VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ECONOMICS

CSOV

44TH INTERNATIONAL CONFERENCE

MME 2026

Mathematical Methods in Economics

FACULTY OF ECONOMICS
VSB – TUO
OSTRAVA

CONFERENCE PROCEEDINGS

ISBN 978-80-248-4880-8 (on-line)
ISSN 2788-3965

SEPTEMBER 9 TO 11, 2026
OSTRAVA

THE MME 2026 CONFERENCE ORGANIZED IN COOPERATION WITH FACULTY OF ECONOMICS, VSB – TECHNICAL UNIVERSITY OF OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ECONOMICS

PROCEEDINGS PUBLISHED BY THE CZECH SOCIETY FOR OPERATIONS RESEARCH

EDITORS Radek Němec, Lucie Chytilová

COVER DESIGN Markéta Šindlerová

PUBLISHER VŠB – Technical University of Ostrava

Faculty of Economics

Department of Systems Engineering and Informatics

PUBLICATION YEAR 2026

NUMBER OF PAGES 401

@ COPYRIGHT the author/authors of each paper

ISBN 978-80-248-4880-8 (on-line)

ISSN 2788-3965

PAPER CITATION EXAMPLE

Author A. (2026). 'Title of the paper'. In: Nemec, R. and Chytilova, L. (eds.) Proceedings of the 44th International Conference on Mathematical Methods in Economics 2026, September 9-11, 2026, Ostrava, Czech Republic, pp x-y.

All papers published in the proceedings have been peer-reviewed by at least 2 independent reviewers. Editors are not responsible for the grammar and quality of language in papers.

CONFERENCE WEBSITE

https://www.ekf.vsb.cz/mme/en

PROGRAM COMMITTEE

CZECH MEMBERS

FRANTIŠEK ZAPLETAL (chair), VSB – Technical University of Ostrava, Czech Republic

HELENA BROŽOVÁ (president of the Czech Society for Operations Research), Czech University of Life Science Prague, Czech Republic

MILOŠ KOPA (vice-president of the Czech Society for Operations Research), Charles University, Czech Republic

ADAM BOROVIČKA, Prague University of Economics and Business, Czech Republic

MARTIN BRANDA, Charles University, Czech Republic

MICHAL ČERNÝ, Prague University of Economics and Business, Czech Republic

PETR FIALA, Prague University of Economics and Business, Czech Republic

MILAN HLADÍK, Charles University, Czech Republic

DUŠAN HRABEC, Tomas Bata University in Zlín, Czech Republic

JOSEF JABLONSKÝ, Prague University of Economics and Business, Czech Republic

FRANTIŠEK KOBLASA, Technical University of Liberec, Czech Republic

MARTINA KUNCOVÁ, College of Polytechnics Jihlava, Czech Republic

MIROSLAV PLEVNÝ, University of West Bohemia, Czech Republic

VÍT PROCHÁZKA, Charles University, Czech Republic

MIROSLAV RADA, Prague University of Economics and Business, Czech Republic

MIROSLAV ŽIŽKA, Technical University of Liberec, Czech Republic

FOREIGN MEMBERS

TOMASZ BRZĘCZEK, Poznań University of Technology, Poland

LORENA CARIDAD LÓPEZ DEL RÍO, University of Córdoba, Spain

MICHAELA CHOCHOLATÁ, Bratislava University of Economics and Business, Slovakia

MARTIN FLÉGL, School of Engineering and Sciences, Tecnológico de Monterrey, Mexico

ARILD HOFF, Molde University College, Norway

LARS M. HVATTUM, Molde University College, Norway

RICHARD F. HARTL, University of Vienna, Austria

JAROSLAV JANÁČEK, University of Žilina, Slovak Republic

FRANCESCA MAGGIONI, University of Bergamo, Italy

VIEROSLAV MOLNÁR, Technical University of Košice, Slovakia

DANIEL P. MERTENS, Jacksonville State University, USA

JORGE L. OYOLA, University of Córdoba, Colombia

EDY PORTMANN, University of Freiburg, Freiburg, Germany

MARIAN REIFF, Bratislava University of Economics and Business, Slovakia

YURY REDUTSKIY, Molde University College, Norway

KENNETH SÖRENSEN, University of Antwerp, Belgium

ERIK ŠOLTÉS, University of Economics, Bratislava, Slovak Republic

MILAN VLACH, Kyoto College of Graduate Studies for Informatics, Japan

GERHARD-WILHELM WEBER, Poznan University of Technology, Poland

ORGANIZING COMMITTEE

LUCIE KLAUBENSCHALK CHYTILOVÁ (chair)

BLANKA BAZSOVÁ, KAMILA FRIEDRICHOVÁ (administrative assistant),

JANA HANČLOVÁ, MIROSLAV HUDEC, JAKUB ILLÍK, DARINA JÍLKOVÁ,

RADEK NĚMEC, VÍTĚZSLAV NOVÁK, PETR ROZEHNAL, MARKÉTA ŠINDLEROVÁ

Department of Systems Engineering and Informatics
Faculty of Economics
VSB - Technical University of Ostrava
Czech Republic

FOREWORD

It is my honour to welcome you to the 44th International Conference on Mathematical Methods in Economics (MME 2026), hosted by the Faculty of Economics, VSB – Technical University of Ostrava.

These Proceedings bring together the scientific contributions accepted for presentation and publication at MME 2026. Following the careful review process, 81 contributions were accepted, including 59 full papers and 22 abstracts. Together, they reflect the broad scope of current research in operations research, econometrics, mathematical modelling, decision sciences, and related quantitative disciplines.

The scientific programme comprised 74 presentations, including 2 plenary lectures, 9 presentations in the PhD Student Competition, and 63 regular presentations. We were also pleased to welcome 83 participants, approximately 40 % of whom came from abroad, confirming the strong international character of the conference. We are especially honoured to host two distinguished plenary speakers, Professor Christa Cuchiero from the University of Vienna and Professor Stefan Trück from Macquarie University in Sydney. Their lectures provide valuable insights into current developments in quantitative and mathematical modelling and complement the wide range of topics represented in the conference sessions.

An important part of MME 2026 is once again the PhD Student Competition, organized with the support of the Czech Society for Operations Research. It provides young researchers with an opportunity to present their work, receive expert feedback, and engage with the broader scientific community.

I would like to sincerely thank all authors, reviewers, plenary speakers, session chairs, and members of the Programme and Organizing Committees for their invaluable contributions. Special thanks go to the Czech Society for Operations Research for its continued support.

On behalf of the Programme and Organizing Committees,

FRANTIŠEK ZAPLETAL

Chair of the Programme Committee of MME 2026

Ostrava, September 2026

TABLE OF CONTENTS

TITLE AND AUTHORS	PP	PAPER #
System Dynamics Model of Carbon Agriculture as a Complex Problem *DANIEL BENEDIKT, JAN RYDVAL, MILAN HOUŠKA, MIRA MAULESHOVA*	13-19	13
Solving the Fixed-Destination Open Vehicle Routing Problem Using Open-Source Tools *MICHAL BOGAR*	20-26	67
Gradual optimizing a sustainable investment portfolio: sequential mean-semivariance-ESG model *ADAM BOROVIČKA*	27-32	71
Seasonality and trend in multiproduct inventory testing and cost modelling *TOMASZ BRZĘCZEK*	33-40	31
Digital development, governance quality, and income inequality: empirical evidence from Vietnamese provinces *HAN BUI, THAO LE, THUY VO, ZDENĚK ZMEŠKAL*	41-47	60
Reinforcement Learning for Perpetual Futures Market Making *MARTIN ČEKAL*	48-53	82
Investment Timing and Strategic Interaction in Cross-Border M&A: A Real Options Perspective *QIYUAN DAI*	54-64	43
Price of Inertia in Emergency Medical System Adjustment *LÝDIA GÁBRIŠOVÁ, JAROSLAV JANACEK, MAREK KVET*	65-71	15
Dynamic portfolio optimization based on volatility-adaptive LSTM and tail-dependence topological network *QIAN GAO*	72-77	19
Early Warning of Sovereign Credit Risk in BRICS: A Comparative GARCH-Based Analysis *SHIKAI GAO*	78-83	26
Dynamic relationship of green economics: renewable energy consumption, CO2 emissions, real GDP and oil prices in European Union countries *JANA HANCLOVA, LUCIE CHYTILOVA, JAKUB ILLÍK*	84-90	50
Structural Factors Influencing AI Copyright Literacy: A Quantitative Analysis from an Academic Survey Study *JANA HECKENBERGEROVÁ, BARBORA KOVÁŘOVÁ*	91-99	44

Mitigating Conflict in Fair Sharing via Resource Design *Robert Hlavatý*	100-106	24
Assessing SDG 11 Efficiency in Czech Regions: An Iterative BoD-DEA Approach with Weight Restrictions *Eva Hrabovská, Markéta Matulová*	107-113	52
Asymmetric evaluation of experts opinions considering their relevancies to the evaluated proposals or candidates *Miroslav Hudec, Miljan Vučetić, David Ulmann, Hanna Kristín Skaftadóttir*	114-119	54
Spatial Patterns of Household Internet Access in EU Regions: An ESDA Approach *Michaela Chocholata*	120-126	78
Minimal Change of Ambulance Deployment to Reach Given Quality of Service *Jaroslav Janacek, Marek Kvet*	127-132	16
Robust Regression of Human Resources and Labor Market Dimensions of TTDI *Jan Kalina, Kristýna Dvořáková*	133-139	6
Predicting Prices and Discovering Structure in Magic: The Gathering Card Data *Nikola Kaspříková*	140-145	69
The economic impact of natural disasters on NUTS2 regions in EU *Daniil Khokhrunov*	146-151	20
Missing Data and Inference in Survey-Weighted Ordered Logistic Models *Ján Kobida and Ľubica Bartová*	152-157	55
Integrated job shop scheduling and bin packing problem *František Koblasa, Miroslav Vavroušek*	158-164	45
Spatial Dynamic Model on Hierarchical Panel Data: Basic Log-likelihood Properties *Martin Konopásek*	165-170	51
The Use of a Two-Step Decomposition Approach in Service Vehicle Routing Planning *Petr Kozel*	171-176	76
The Impact of GDP, Inflation, Industrial Employment, and Trade Openness on Foreign Direct Investment in the Visegrad Group Countries *Radmila Krkošková, Zuzana Neničková, Lucie Waleczek Zotyková*	177-183	4

Econometric Modelling of Tourism Demand in Collective Accommodation Facilities: Evidence from Czech Tourism Data *Radmila Krkošková, Zuzana Neničková*	184-189	10
Performance Dynamics of Czech Breweries: An Updated DEA-Based Assessment *Martina Kuncová, Roman Fiala*	190-196	73
Metrics in Scrum: A quantitative model for evaluating metrics utilization in Scrum teams *Josef Kunhart, Jan Bartoška*	197-203	11
Ambulance Fleet Minimization without Loss of System Quality *Marek Kvet, Jaroslav Janacek*	204-210	17
A Multilevel Analysis of Post-Harvest Food Loss in Sub-Saharan Africa: Commodity, Food Supply Stage, and Regional Interactions *Elvis Kwakye*	211-217	59
Comparative Analysis of Mathematical Programming and Metaheuristic Approaches for the Vehicle Routing Problem with Multiple Soft Time Windows *Farouk Ligali, Dušan Hrabec and Lars Magnus Hvattum*	218-224	57
Firm level evidence on investor sentiment and cross sectional stock market characteristics *Bahate Maidiya*	225-230	8
Counting passengers in public transport buses using door optical sensors *Tomáš Majer, Dalibor Gonda, Ivana Ondrušová*	231-237	58
Structural Modelling of Sustainability Perception: Triangulation of Survey Data and Regional Indicators *Ladislava Míková, Pavlína Forstová Kuráňová, Václav Friedrich*	238-243	40
Comparison of Bootstrap Methods in Robust SSD Portfolio Optimization *Jakub Neugebauer*	244-249	63
Path Dependent Arbitrage Portfolio Strategies *Valentina Piantoni, Denise Mirabella*	250-256	27
Choosing a Value-at-Risk UC Backtest across Standard Tolerance Levels: An Exact-Size Power Analysis *Krzysztof Piontek, Maria Mach-Król*	257-265	68
Roughness and Jump Risk in Cryptocurrency Option Smiles: An Econometric Comparison *Michal Polak*	266-271	61

Regional Apartment Price Disparities in the Czech Republic: Clustering, Supply Elasticity, and Labour Market Analysis with a Focus on the Moravian-Silesian Region 272-278 42
Alena Pozdílková, Martin Pozdílek

Analysis of Productivity in an Automated Production Line and Modeling Its Dynamics 279-285 14
Pavel Pražák, Josef Toman, Andrea Ševčíková

Evaluating Strikers and Forwards in the Bundesliga with a Two-Stage Decision-Making Framework 286-293 37
José Rodea-Zamudio, Martin Flegl, Enrique Delahoz-Domínguez, Eusebio Vargas Estrada

How University Students Learn with Generative AI: Usage Patterns, Study Materials, and Risk Awareness 294-300 86
Petr Rozehnal, Lucie Klaubenschalk Chytilová, Vítězslav Novák, Miroslav Langer

Optimizing the Supply Chain Finance for Increasing the Efficiency 301-309 23
Xiaotong Sang

Explainable Articial Intelligence for Dairy Cow Classication: A Neural Network Approach with SHAP and LIME Interpretability 310-315 1
Jan Saro, Vojtěch Šimůnek

Radial Basis Functions: Analytical Characterization of Polynomial Augmentation and Bias Identification 316-323 84
Vaclav Skala

Fractional Economic Dynamics: Deriving Production Functions via Caputo-Type Memory Invariants 324-329 64
Roman Smirnov

Weekday and Daytime Influence on Work Contours in Projects 330-337 21
Tomas Subrt, Jan Bartoska, Josef Kunhart, Irena Strausova, Simona Huguet

Fuzzy Multi-Attributes decision model for Athletes Position Suitability 338-344 79
Vojtěch Šimůnek, Helena Brozova, Jan Saro

Legitimacy–Cost Gap in Reshoring: Evidence from Central Europe 345-351 35
Markéta Šindlerová, Jaroslav Urminský, Jakub Černoch, Ondřej Melecký

The Impact of Competition and Quality on the Efficiency of Integrated Transportation Systems 352-359 7
Eva Štichhauerová, Miroslav Žižka

Comparison Between Logistic Regression and Machine Learning Methods Using Monte Carlo Simulations 360-365 25
Kryštof Tomsa

Predicting of Realized Volatility of Bitcoin and Ethereum with Sample Entropy 366-371 33
Quang Van Tran, Karel Brůna

Convergence and Differentiation in Slovak Spa Enterprises: A Cluster Analysis of Financial Indicators 372-377 41
Petra Vašaničová, Sylvia Jenčová, Martina Košíková, Marta Miškufová, Kateryna Melnyk

Beyond Predict-then-Optimize: Decision-Focused Learning for Grid-Scale Battery Arbitrage 378-382 65
Vojtěch Vávra

Translating Online Search Behavior into Quantitative Portfolio Views: Evidence from Hong Kong 383-388 28
Jialei Xiong

Modeling number of advertisements and asking prices of flats with von Mises distributions and regression analysis: Moravian-Silesian Region 2018–2025 389-394 39
Jaromir Zahradka, Jaroslav Marek

A DEA-Bootstrap Framework for Assessing Educational Quality in Industrial Engineering Programs 395-401 2
Rohemi Zuluaga-Ortiz, Carlos Garcia-Yerena, Enrique Delahoz-Domínguez

CONTRIBUTIONS

System Dynamics Model of Carbon Agriculture as a Complex Problem

Daniel Benedikt[1], Jan Rydval[2], Milan Houška[3], Mira Mauleshova[4]

Abstract. This paper presents a system dynamics stock-and-flow model implemented in Vensim for simulating carbon credit generation on a mixed crop-livestock farm. The model contains four stocks: Soil Organic Carbon (SOC, tCO2e), Soil Health and Fertility Index (0;1), Livestock Unit (LU), and Carbon Credits Available (tCO2e). Farm GHG emissions are computed from six IPCC-based sources: enteric CH4 with feed quality adjustment, manure CH4 using weighted MCF across four management systems (IPCC, 2006), direct and indirect soil N2O with split emission factors for mineral and organic nitrogen (IPCC, 2019), fuel CO2, upstream fertilizer production emissions, and liming CO2. Credit issuance operates through two channels: verified net sequestration and verified emission reductions against a fixed baseline, each subject to a 5-year MRV (Monitoring, Reporting and Verification) delay. Soil health stock creates a reinforcing loop: improved fertility raises yields and biomass return, enhancing sequestration. Simulations over a 50-year horizon reveal that per-hectare sequestration declines from 1.2 to below 0.3 tCO2e/ha/yr as saturation binds, while the MRV delay creates an initial credit gap posing a financial barrier to adoption. Sensitivity analysis shows that raising practice adoption from 70% to 90% increases cumulative sequestration by 25% over 30 years, moderated by saturation.

Keywords: Carbon Credits & MRV Delay, Carbon Sequestration, GHG Emissions Soil Health & Fertility, Soil Organic Carbon (SOC), System Dynamics Modeling

JEL Classification: C63, Q15

1 Introduction

Carbon farming is a modern sustainable agronomical approach in which agricultural practices are implemented to increase carbon into soils and biomass and reduce on-farm GHG emissions, with the resulting benefit quantified, verified, and converted into tradable carbon credits (Lal, 2020; Smith et al., 2020). It rests on three pillars: sequestration, emission reductions, and credit issuance. Practices such as cover cropping, reduced tillage, diversified rotations, organic fertilization, agroforestry, and grassland restoration raise soil organic carbon (SOC), although accumulation is non-linear and slows as site-specific saturation is approached (IPCC, 2019). Emission-reduction measures, including improved feed, covered manure storage, anaerobic digestion, nitrification inhibitors, precision nutrient management, and fuel or fertilizer substitution, lower CH_4, N_2O, and upstream CO_2 emissions (IPCC, 2006; IPCC, 2019). Credits are issued against a baseline scenario representing business-as-usual farm management, after applying permanence buffers and conservativeness factors. Under the French Label Bas-Carbone CARBON AGRI methodology, verification follows a five-year MRV cycle, creating a delay between practice adoption and revenue realization (Ministère De La Transition Écologique, 2019).

Understanding how these regulatory, biophysical, and economic constraints interact requires an approach that captures accumulation, delays, and feedbacks. System dynamics (SD) provides such a framework by modelling systems through stocks, flows, feedback loops, delays, and non-linearities. In agriculture and natural-resource management, SD has been applied to policy impacts, land-use change, livestock dynamics, and soil-water-nutrient interactions (Liu et al., 2020). Its main strength is linking biophysical processes such as SOC accumulation and mineralization with economic and institutional drivers, including MRV cycles, baseline setting, and farmer adoption. For carbon farming, SD makes it possible to represent sequestration limits, the reinforcing relationship between soil health and productivity, and the financial effects of verification delays. While causal loop diagrams (CLDs) help map system structure, quantitative stock-and-flow diagrams (SFDs) are needed to simulate system behavior as credit issuance and test sensitivity to policy and management choices.

[1] Czech University of Life Sciences, Department of Systems Engineering, Kamýcká 129, 165 00 Praha - Suchdol, benediktd@pef.czu.cz
[2] Czech University of Life Sciences, Department of Systems Engineering, Kamýcká 129, 165 00 Praha - Suchdol, rydval@ pef.czu.cz.
[3] Czech University of Life Sciences, Department of Systems Engineering, Kamýcká 129, 165 00 Praha - Suchdol, houska@ pef.czu.cz.
[4] Czech University of Life Sciences, Department of Systems Engineering, Kamýcká 129, 165 00 Praha - Suchdol, mauleshova@ pef.czu.cz.

The objective of this paper is to develop and analyze a stock-and-flow model of a carbon farming system, where stock dynamics are represented by ordinary differential equations and numerically integrated over time. The model is employed to simulate system behavior and evaluate the sensitivity of the volume of carbon credits generated to varying levels of practice adoption.

2 Materials and Methods

To analyze how practice adoption affects the final volume of carbon credits, a method is needed that captures slow soil carbon accumulation, non-linear saturation, feedbacks, and MRV delays. SD is suitable because it models systems through stocks, flows, feedback loops, and time delays (Sterman, 2000). In agriculture, SD is useful for representing slowly changing biophysical stocks and delayed institutional effects that static or econometric approaches capture poorly. It has already been applied to land-use change, livestock-crop interactions, and green agriculture policies (Liu et al., 2020), and in carbon farming, it links soil carbon, farm management, emissions, and verification cycles under frameworks such as the EU CRCF, Verra, Gold Standard and Label Bas-Carbone (European Commission, 2024).

2.1 Stock-and-Flow Diagram

Following Egerer et al. (2021), a CLD was used to model for climate change adoption in agriculture, identify key feedbacks, and define system boundaries. The CLD serves as a preparatory step for the SFD, which converts the conceptual system structure into a quantitative simulation model (Sterman, 2000). The present paper focuses on the development and analysis of this SFD model.

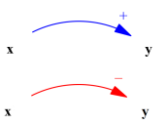
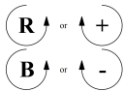
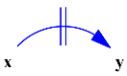
Representation	Interpretation	Mathematical formulation
Stock Variable	Accumulates all inflows and outflows over time; retains its value between time steps. The sole type of "memory" in a SD model	$Y(T) = Y(T_0) + \int_{T_0}^{T} \left(\sum_i In_i(t) - \sum_j Out_j(t) \right) dt,$ (1) where Y is a stock variable and dt is a simulation time step
	Inflow increases the connected stock and outflow decreases the connected stock.	$In(t) = f(x_1, x_2 \ldots, A) \geq 0$ (2) $Out(t) = g(x_1, x_2 \ldots, A) \geq 0$ (3) where x_1, x_2 are auxiliary variables and A is a parameter
Auxiliary variable	Auxiliary variable (other auxiliaries and parameters at each time step; holds no value between steps).	$A(t) = h(x_1(t), x_2(t), \ldots)$ (4)
	Positive and negative causal link. Ceteris paribus, positive is an increase (decrease) in X causes an increase (decrease) in Y above (below) what Y would otherwise be while negative is an inverse relationship between X and Y.	$\frac{\partial Y}{\partial X} > 0..positive,$ $\frac{\partial Y}{\partial X} < 0..negative$ (5) if Y is a stock: $Y = \int (\pm X)\, dt + Y_{T_0}$ (6)
	Reinforcing (positive) loop: generates self-amplifying growth or decline. Balancing (negative) loop: generates goal-seeking or oscillatory behaviour and resists deviation from equilibrium.	
	The effect of X on Y is deferred by delay time t. First-order smooth delay modelled as an auxiliary stock of pipeline material.	

Table 1 System dynamics building blocks (Sterman, 2000, pp 138-139, 193-194)

The SFD operationalizes the system as a set of stock variables (Table 1) whose evolution is governed by inflows and outflows, expressed as ordinary differential equations and integrated numerically over a defined simulation horizon (Forrester, 1994; Sterman, 2000). Stocks represent accumulations, while flows represent the rates at which those accumulations change. This distinction is essential: stocks change only through flows and introduce

continuity and delay into system behavior, where flows respond to the current values of stocks, auxiliaries, and parameters (Schoenberg et al., 2020). Auxiliary variables decompose complex flow expressions into intermediate quantities and make it possible to represent non-linear relationships, threshold effects, and empirically derived response functions through lookup tables of the form $y = g(x)$, which interpolate between tabulated values without requiring a closed analytical form (Sterman, 2000).

The parameterization of the stock-and-flow model combines methodological standards, expert sources, and policy documents. Default emission factors and Tier 2 procedures for enteric and manure CH_4, soil N_2O, and CO_2 from fuel, liming, and upstream fertilizer production follow the 2006 IPCC Guidelines and the 2019 Refinement (IPCC, 2006; IPCC, 2019). Czech-specific data on farm structure were complemented by soil characteristics derived from the BPEJ (evaluated soil-ecological unit) database maintained by the Research Institute for Soil and Water Conservation (VÚMOP, 2024), by crop yields, sown areas, and production volumes for the principal arable rotation crops together with cattle herd numbers from the Situational and Outlook Reports of the Ministry of Agriculture (MZe, 2025), and by management parameters Institute of Agricultural Economics and Information (ÚZEI, 2024).

2.2 Monte Carlo Simulation

Monte Carlo (MC) simulation in SD is a stochastic extension of deterministic stock-and-flow modelling, in which uncertain parameters are sampled repeatedly from predefined probability distributions to generate a distribution of possible model trajectories rather than a single output value (Sterman, 2000). Mathematically, it evaluates how uncertainty in parameters, initial conditions, nonlinear functions, and delay structures propagates through systems of differential or difference equations over time, making it suitable for sensitivity analysis, uncertainty quantification, and robustness testing (Saltelli et al., 2008). In agricultural and environmental models, MC simulation can support scenario evaluation in yields, soil carbon dynamics, policy incentives or adoption rates (Wang et al., 2022).

3 Results

This section presents a farm-level SD model for simulating carbon credit generation on a mixed crop-livestock and only crops farm under the Label Bas-Carbone CARBON AGRI and IPCC. It proceeds in two steps. First, a compact CLD was constructed to determine the system structure. Second, the SFD (Figure 1) is presented together with 50-year simulation outputs for soil organic carbon, on-farm GHG emissions, verified credit issuance, and the sensitivity of cumulative sequestration to practice adoption.

3.1 Carbon Farming Stock and Flow Diagram

The qualitative structure captured in the CLD is operationalized as a quantitative SFD implemented in Vensim (Figure 1). The model consists of 99 variables (4 stock variables and 2 stock delayed variables, 39 auxiliaries, 53 constants and 1 lookup function). The model describes the main reinforcing and balancing loops identified in CLD while adding the algebraic detail required for numerical simulation: emission factors, conversion coefficients, lookup functions, and the explicit material delays governing credit issuance. The simulation horizon is over 50 years with an annual integration step, and the parameterization follows IPCC guidance (IPCC, 2006; IPCC, 2019).

Parameter	Value	Unit	Source	Parameter	Value	Unit	Source
Herd Housing Capacity	50	Livestock units	Model assumption	MCF digestation	0.02	dmnl	IPCC (2006)
EF Manure	0.6	$tCO_2e/LU/yr$	IPCC (2019)	MCF Slurry	0.14	dmnl	IPCC (2006)
EF Enteric CH4	3	$tCO_2e/LU/yr$	IPCC (2019)	EF Mineral Production	0.0065	tCO_2e/kgN	IPCC (2019)
EF N2O Mineral	0.016	kgN_2O-N/kgN	IPCC (2019)	EF Liming	0.000122	tCO_2/kg	IPCC (2006)

Table 2 Description of selected SFD model parameters

The model contains four state variables. *Livestock Herd (LH)* measured in livestock units changes through *restocking (RESR)* and *herd reduction*, *(HR)* with both shaped by *feed availability* and *housing capacity*.

$$LH(T) = Init\ LH\,(T_0) + \int_{T_0}^{T} \big(RESR(t) - HR(t)\big)dt, \tag{7}$$

where:

- $LH\,(T)$ is the number of cattle in the system in time T;
- $Init\ LH\,(T_0)$ is the initial number of cattle in the system in time T_0;
- $RESR(t)$ is the intensity of Restocking rate of cattle over time t;
- $HR(t)$ is the intensity of the Herd Reduction of cattle over time t.

Carbon Credits Available (CCA) (tCO_2e) accumulates *Credits from Sequestration (CIFS)* and *Credits from Emission Reduction (CIFER)* and declines through *credit retirement*. *Enteric* CH_4 depends on *herd size* and *feed quality*; *manure* CH_4 is based on a *weighted methane conversion factor* across *pasture, slurry, solid storage,* and *anaerobic digestion*. *Direct* and *indirect* N_2O *from soils* are driven by *mineral* and *organic nitrogen inputs*, the latter linked to *herd size*. *Credit issuance* operates through two parallel channels, both subject to a permanence *of Penalization Share* and a five-year *MRV delay*. On the emissions side, the difference between *current farm emissions and baseline emissions at project start*, adjusted for *leakage,* gives *Raw Emission Reduction*, which after penalisation and delay becomes *Verified Emission Reduction*. The difference between *average SOC sequestration* and *average SOC loss*, adjusted for leakage, gives *Raw Net Sequestration*, which after becomes *Verified Net Sequestration*.

$$CCA(T) = Init\ Cr(T_0) + \int_{T_0}^{T} \big(\text{CIFER}\ (t) + CIFS(t) - CR(t)\big)dt, \tag{8}$$

where:

- $CCA\ (T)$ is the amount of Carbon Credits Available in the system in time T;
- $Init\ Cr(T_0)$ is the Initial Value of Carbon credits in the system in time T_0;
- $CIFER(t)$ is the intensity of Credits from emission Reduction being created over time t;
- $CIFS(t)$ is the intensity of Credits from Sequestration being created over time t;
- $CR(t)$ is the intensity of the Credits retirement over time t.

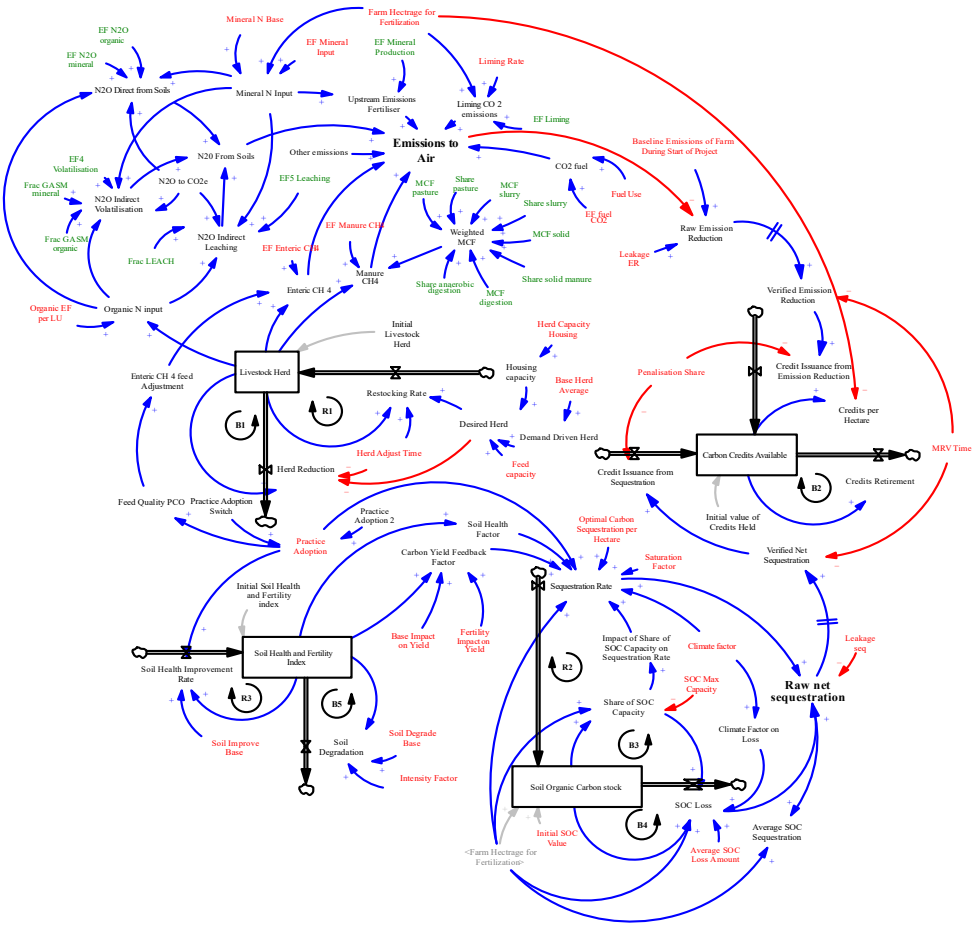


Figure 1 Carbon Farming SFD (Resource: Author's own adaptation in VENSIM)

Soil Organic Carbon Stock (SOCS) (tCO_2e) changes through *Sequestration Rate (SR)* and *SOC loss (SOCL);* sequestration depends on a reference per-hectare rate as a multiplicator, soil health, yield feedbacks, and a non-linear saturation effect linked to the share of maximum *SOC capacity*, while losses are scaled by climate effects.

$$SOCS(T) = Init\ SOCv\,(T_0) + \int_{T_0}^{T} \big(SR\ (t) - SOCL(t)\big)dt, \tag{9}$$

where:

- $SOCS\ (T)$ is the amount of Soil Organic Carbon in the system in time T;
- $Init\ SOCv\,(T_0)$ is the initial Amount of SOC in the system in time T_0;
- $SR(t)$ is the intensity of ekCO2 being sequestrated over time t;
- $SOCL(t)$ is the intensity of the loss of SOC over time t.

The Soil Health and Fertility Index (SOHaFI), which increases the *Soil Health Improvement Rate (SHIR)* and decreases the *Soil Degradation (SDeg)* modulated by farming intensity. The subsystem is feedback linked to the soil organic carbon supply through the Carbon Yield Feedback factor.

$$SOHaFI(T) = Init\ SHaFI\,(T_0) + \int_{T_0}^{T} \big(\mathrm{SHIR}\ (t) - SDeg\ (t)\big)dt, \tag{10}$$

where:

- $SOHaFI\ (T)$ is the amount of Soil Health and Fertility Index in the system in time T;
- $Init\ Soil\ Health\ Improvement\ Rate(T_0)$ is the initial Soil Health and Fertility Index in the system in time T_0;
- $SHIR(t)$ is the intensity of Soil Health Improvement over time t;
- $SDeg(t)$ is the intensity of Soil Degradation over time t.

No relevant real-data time series exist to validate the model's replication of real-world behavior using standard metrics such as MAPE, SMAPE, or R^2. We therefore validate our model's results against the findings of long-term studies. The model's results indicate that the average potential for CO_2 storage and reduction is approximately 1 t eq CO_2 on average, which is consistent with the six-year study by I4CE (2025, p. 56).

3.2 Monte-Carlo Simulation

To verify the model's robustness to uncertainties in the parameters, a MC sensitivity analysis (Figure 2) was performed in accordance with validation procedures for system dynamic models (Sterman, 2000).

Input parameters for MC	MIN	MAX	Average	Standard deviation	Distribution
Storage (kg ekv CO2/ha)	-1 191.00	6 611.00	1 572.60	2 321.61	Normal
Emission GHGs kg ekv CO2/ha	1 415.00	3 047.00	2 266.80	523.18	Normal
Hectrage	237.00	4 334.00	1 188.60	1 137.46	Normal
Number of Livestock Units	0.00	676.20	138.30	233.53	Normal
Fuel Use	25 000.00	446 402.00	128 234.90	116 051.59	Normal
Use of mineral N / ha	0.00	158.00	92.10	54.30	Normal
Use of organic N /LU	0.00	111.00	35.30	37.05	Normal

Table 3 Input parameters for MC (Resource: OK CARBON, 2025)

The model is parameterized against farms from the pilot projects in the Czech Republic from which were data in aggregated form gathered for simulation Monte Carlo (Table 3Table **3**, ranging from purely arable holdings to mixed crop livestock systems with cattle. The absence of a livestock branch on arable-only farms is not a structural break: the herd stock and associated CH_4 and N_2O flows are simply set to zero, while the SOC sub-model continues to draw inputs from crop residues, cover crops and external organic amendments. The pilots can thus be treated as variants of one mixed crop–livestock archetype observed at different stages of integration. Across 10,000 stochastic runs, four structural properties of the model emerge.

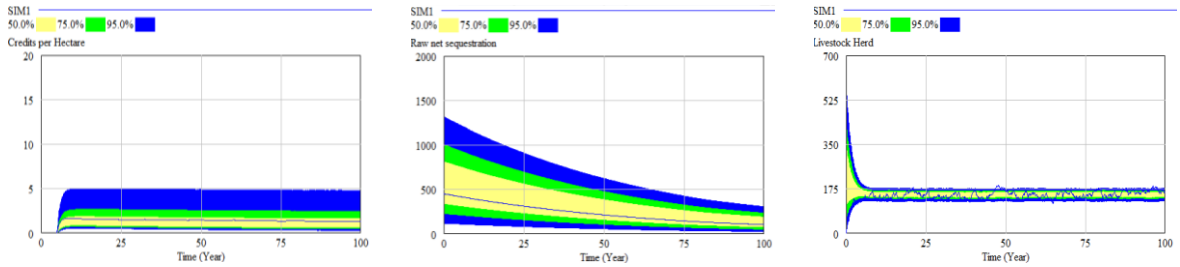


Figure 2 Monte Carlo Simulations (Resource: Author's own adaptation in VENSIM)

Results of MC simulation (Figure 2) shows that First, *Credits per Hectare* never exceed 5 tCO_2e/ha/yr in any simulation, with this upper bound reached only within the 95th percentile band; the median stabilizes between 2–3 tCO_2e/ha which is similar to IC4E (2025, pp 57) for each carbon project variation. Second, the *Livestock Herd* rapidly converges from its initial value (525 LU) to a steady state near around 175 LU within the first decade across all percentiles. Results have been obtained during one year pilot period and partial results need to be anymized. *Raw net sequestration* follows an asymptotically declining trajectory, falling from 500 - 1,500 kgC/ha/yr to 100 - 400 kgC/ha/yr over the 50-year horizon, with the 95 % confidence band progressively narrowing. This quantitatively confirms the nonlinear SOC saturation lookup and reproduces the agronomic principle that sequestration potential decreases as soils approach their carbon storage ceiling.

4 Conclusion

The SD methodology proved to be a suitable framework for representing farm-level carbon credit generation, as it captures the accumulation, delays, and feedbacks that static approaches overlook. The MC simulations confirm that Raw Net Sequestration declines asymptotically over the horizon because rising SOC stock progressively saturates the soil's storage capacity and slows the sequestration rate, while in mixed crop-livestock settings the projects optimize the Livestock Herd toward farm capacity constraints (convergence from ca 525 to ca 175 LU). Average Credits per Hectare stabilise at 2-3 tCO_2e/ha/yr (median), with an upper bound of ca 5 tCO_2e/ha/yr reached only in the 95th percentile. This paper supports further research for the whole carbon agriculture system.

Acknowledgements

This work was supported by the research grant of PEF ČZU under Internal Grant Agency, project no. 2026A003.

References

Egerer S, Valencia Cotera R, Celliers L, Máñez Costa M (2021) A leverage points analysis of a qualitative system dynamics model for climate change adaptation in agriculture. Agricultural Systems 189:103052

European Commission (2024) Communication on Sustainable Carbon Cycles and the Carbon Removal Certification Framework. European Commission, Brussels. https://climate.ec.europa.eu/eu-action/sustainable-carbon-cycles_en

Forrester J W (1994) System dynamics, systems thinking, and soft OR. System Dynamics Review 10(2–3)

I4CE (2025) Six years of carbon certification in France: an assessment of the Label Bas-Carbone. Institute for Climate Economics, Paris. https://www.i4ce.org/wp-content/uploads/2025/06/Six-years-of-carbon-certification-in-France.pdf

IPCC (2006) 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Institute for Global Environmental Strategies (IGES) for the IPCC, Hayama

IPCC (2019) 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Intergovernmental Panel on Climate Change, Geneva

Lal R (2020) Soil organic matter and water retention. Agronomy Journal 112(5):3265–3277

Liu Y et al. (2020) A system dynamics model analysis for policy impacts on green agriculture development. Journal of Cleaner Production

Ministère de la Transition Écologique (2019) Méthode CARBON AGRI: Label Bas-Carbone. Ministère de la Transition Écologique, Paris. https://www.ecologie.gouv.fr/label-bas-carbone

MZe (2025) Situační a výhledová zpráva: Skot – hovězí maso. Ministerstvo zemědělství ČR, Praha. https://mze.gov.cz/public/portal/mze/publikace/situacni-vyhledove-zpravy/zivocisne-komodity-hospodarska-zvirata/skot-hovezi-maso

Saltelli A et al. (2008) Global Sensitivity Analysis: The Primer. Wiley, Chichester

Schoenberg W, Davidsen P, Eberlein R (2020) Understanding model behavior using the loops that matter method. System Dynamics Review 36(2):158–190

Smith P et al. (2020) How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration. Global Change Biology 26(1):219–241

Sterman J D (2000) Business Dynamics: Systems Thinking and Modeling for a Complex World. Irwin/McGraw-Hill, Boston

ÚZEI (2024) Nákladovost zemědělských výrobků – výběrové šetření nákladů a výnosů zemědělských komodit. Ústav zemědělské ekonomiky a informací, Praha. https://www.uzei.cz/publikacni-cinnost/databaze/nakladovost-zemedelskych-vyrobku

VÚMOP (2024) Bonitace zemědělského půdního fondu – databáze BPEJ. Výzkumný ústav meliorací a ochrany půdy, Praha. https://bpej.vumop.cz/

Wang W, Li K, Liu Y, Lian J, Hong S (2022) A system dynamics model analysis for policy impacts on green agriculture development: a case of the Sichuan Tibetan Area. Journal of Cleaner Production 371:133562

Cavana R. Y et al. (2021) Feedback Economics: Economic Modeling with System Dynamics. Cham: Springer

Solving the Fixed-Destination Open Vehicle Routing Problem Using Open-Source Tools

Michal Bogár[1]

Abstract. Repeated delivery is a practical routing problem for nonprofit organizations, where reliability must be maintained with limited financial and planning resources. This paper addresses the reorganization of daily delivery routes for an urban nonprofit operation and formulates the case as a Fixed-Destination Open Vehicle Routing Problem (FDOVRP). Unlike the classical Vehicle Routing Problem or Open Vehicle Routing Problem, vehicles do not return to the depot after serving customers and the route endpoints are not freely selected by the optimizer. In this case, each vehicle starts at a common depot and must finish at the same predetermined headquarters node. The study combines mathematical modeling with a reproducible optimization workflow. A FDOVRP formulation defines the fixed start and terminal, customer assignment, flow conservation, and capacity constraints. Address data are geocoded with Nominatim and transformed into a directed travel-time cost matrix using the Open Source Routing Machine (OSRM). The full empirical instance is then solved with cuOpt and compared with the original routes used by the organization. The optimized solution reduced total travel time from 14,228.4 seconds to 12,266.5 seconds, saving 1,961.9 seconds, approximately 32.7 minutes, or 13.79%. The results show that meaningful process optimization can be achieved without increasing fleet size or relaxing the operational requirement of a fixed terminal. The paper contributes an FDOVRP formulation and demonstrates a practical, repeatable workflow based on open-source tools.

Keywords: Vehicle Routing Problem, Fixed-Destination Open Vehicle Routing Problem, route optimization, open-source

JEL Classification: C61, C44, R41, L31

1 Introduction

Recurring logistics decisions are a practical constraint for nonprofit and charitable organizations, where service reliability must be maintained with limited money, staff time, and planning capacity. This paper is motivated by a Bratislava nonprofit organization that provides daily meal delivery to socially vulnerable seniors. The operation involves repeated daily routing, so even modest improvements can accumulate into meaningful savings. At the same time, limited budgets make commercial optimization services difficult to access, especially when routes have to be recomputed frequently. The case therefore calls for a routing approach that is precise enough for daily use but suitable enough for a nonprofit setting.

The described task belongs to the Vehicle Routing Problem (VRP) family, in which customer stops are assigned to vehicles and the order of visits is selected to minimize a cost measure such as distance, travel time, or monetary cost. Toth and Vigo (2002) define routing as a cost minimization problem over vehicle tours, normally with vehicles leaving from and returning to a depot. Braekers, Ramaekers, and Van Nieuwenhuyse (2016) show how practical routing applications motivate variants with different fleet structures, service rules, and route-end conditions. The closest standard variant to the studied problem is the open version of VRP, which removes the requirement that each vehicle must return to the depot after serving all the customers. This variant is commonly referred to as the Open Vehicle Routing Problem (OVRP) and is a well-established extension of the classical VRP (Li, Golden, and Wasil, 2007).

Other related routing families also show that nonstandard route endpoints are practically relevant. Open routing with driver nodes allows routes to terminate at specified driver-related locations (Aksen, Özyurt, and Aras, 2007). School bus routing can likewise involve known starts and a fixed school destination, as discussed by Li and

[1] Bratislava University of Economics and Business, Faculty of Economic Informatics, Department of Operations Research and Econometrics, Dolnozemská cesta 1, 852 35 Bratislava, Slovak Republic, michal.bogar@euba.sk

Fu (2002) and by Bögl, Doerner, and Parragh (2015). Half-open and open-route variants provide additional background for endpoint flexibility. These variants are useful reference points, but they differ from our case in their operational context, objectives and constraints.

The studied nonprofit case has a more specific structure, where each route starts at the delivery depot, but each vehicle must finish at the same predetermined headquarters node, where the vehicles are parked after delivery. This is not a classic closed VRP tour, because the vehicles do not return to the depot, and it is not a fully open route, because the final node is fixed in advance. We refer to this structure as a Fixed-Destination Open Vehicle Routing Problem (FDOVRP). FDOVRP is less established than classic VRP and OVRP, but it captures a practical routing pattern that appears when daily delivery routes have a shared operational endpoint. The main contribution of this paper is the formulation and application of the FDOVRP in a real-life nonprofit setting, supported by an open-source workflow suitable for daily use.

2 Problem Formulation and Mathematical Model

This section further describes the fixed-destination open vehicle routing problem introduced above. In the described operation, meals are loaded at an initial depot, delivered to customers across the city, and the vehicles then finish at a separate headquarters node. Since the routing takes place in an urban setting and is repeated daily, travel time is used as the cost measure instead of physical distance. The detailed address data and node counts are described in the next section.

The routing structure differs from both the classical Vehicle Routing Problem and the standard Open Vehicle Routing Problem. In the classical case, vehicle routes normally return to the depot. In the open version, the return-to-depot requirement is removed and the final endpoint is chosen by the optimization. Here, each vehicle starts at the same depot but must end at the same predetermined headquarters node. This structure is referred to in this paper as the fixed-destination open vehicle routing problem (FDOVRP). All available vehicles are used, each performs one route, and equal capacities help keep volunteer workloads bounded without forcing identical route durations.

2.1 Mathematical Model

The model is written as a mixed-integer formulation on the complete directed graph induced by the depot, customer nodes, and terminal headquarters. Let n be the number of delivery customers and K the number of available vehicles. Node 1 is the depot, nodes $2, \ldots, n+1$ are customers, and node $n+2$ is the terminal headquarters. The travel time from node i to node j is denoted by c_{ij} and is measured in seconds. Demand at node i is q_i, with $q_1 = q_{n+2} = 0$, vehicle capacity is V_k, and $\bar{V}$ is the Big-M constant used in the Miller-Tucker-Zemlin (MTZ) constraint. The binary variable x_{ij}^k equals 1 if vehicle k travels directly from i to j and 0 otherwise. The variable u_i^k denotes the cumulative load assigned to vehicle k after reaching node i. The objective is to minimize total travel time over all vehicles while assigning every customer to exactly one route. The main modeling adaptation is the fixed-endpoint structure: constraint (4) forces each vehicle to depart from the depot, while constraints (5) and (6), together with customer flow conservation in (3), make the headquarters the only feasible route endpoint.

$$\min z = \sum_{k=1}^{K} \sum_{i=1}^{n+2} \sum_{\substack{j=1 \\ j \neq i}}^{n+2} c_{ij} x_{ij}^k \tag{1}$$

Subject to:

$$\sum_{k=1}^{K} \sum_{i=1}^{n+2} x_{ij}^k = 1, \qquad j = 2, 3, \ldots, n+1 \tag{2}$$

$$\sum_{i=1}^{n+2} x_{ij}^k = \sum_{\ell=1}^{n+2} x_{j\ell}^k, \qquad j = 2, 3, \ldots, n+1, \quad k = 1, 2, \ldots, K \tag{3}$$

$$\sum_{j=2}^{n+1} x_{1j}^k = 1, \qquad k = 1, 2, \ldots, K \tag{4}$$

$$x_{i1}^k = 0, \qquad i = 1, 2, \ldots, n+2, \quad k = 1, 2, \ldots, K \tag{5}$$

$$x_{(n+2)j}^k = 0, \qquad j = 1, 2, \ldots, n+2, \quad k = 1, 2, \ldots, K \tag{6}$$

$$u_i^k + q_j - \bar{V}(1 - x_{ij}^k) \leq u_j^k, \qquad i, j = 1, 2, \ldots, n+2, \quad i \neq j, \quad k = 1, 2, \ldots, K \tag{7}$$

$$u_i^k \leq V_k, \qquad i = 2, 3, \ldots, n+1, \quad k = 1, 2, \ldots, K \tag{8}$$

$$u_1^k = 0, \qquad k = 1, 2, \ldots, K \tag{9}$$

$$x_{ij}^k \in \{0, 1\}, \quad u_i^k \in \mathbb{R}_0^+, \qquad i, j = 1, 2, \ldots, n+2, \quad k = 1, 2, \ldots, K \tag{10}$$

The objective function (1) minimizes total travel time for all vehicles in the fleet. Constraint (2) ensures that each customer node is visited exactly once, while (3) preserves vehicle flow by requiring the same vehicle that enters a customer node to leave it. Constraint (4) sends every vehicle from the depot exactly once, and only to a customer node. Constraint (5) prevents vehicles from returning to the depot, and (6) prevents vehicles from leaving the terminal headquarters after arrival. The requirement that each route ends at the headquarters is therefore not represented by constraint (6) alone; it follows from the combination of the mandatory depot departure in (4), customer-node flow conservation in (3), the prohibition on returning to the depot in (5), and the rule in (6) that no arc may leave the terminal once it is reached. Constraint (7) is an MTZ-style load propagation constraint that tracks cumulative vehicle load and prevents disconnected customer subtours. For larger instances, future work could compare the MTZ-style formulation used here with alternative formulations. Constraint (8) enforces vehicle capacity; equal capacities and full fleet use support bounded workloads, but they do not impose identical route durations or stop counts. Finally, (9) initializes the depot load at zero, and (10) defines the binary route variables and nonnegative load variables. The next section uses this formulation as the conceptual basis for constructing the travel-time matrix and computing feasible routes.

3 Data and Solution Workflow

The empirical part of the study is based on an internal address dataset provided by the nonprofit organization. The file represents one particular daily delivery order from the studied operation and contains 100 address nodes in total: 98 customer stops, one initial depot, and one terminal headquarters. The number of customer stops therefore reflects the orders included in that daily delivery day, in practice, the exact number of stops may fluctuate slightly from day to day according to received orders. The customer nodes represent actual delivery addresses, while the depot and headquarters define the fixed start and terminal nodes used in the FDOVRP model. The same internal file also preserves the original assignment of customer stops to delivery routes, which is used as the baseline for evaluating the reorganized solution. The address file is not publicly available because it contains sensitive data.

3.1 Travel-Time Cost Matrix

The solver requires a numerical cost matrix rather than pure addresses. Because the available data contained only street addresses, the first processing step converted each address to latitude and longitude coordinates. This geocoding step used Nominatim, an OpenStreetMap based service for translating address descriptions into geographic coordinates (OpenStreetMap Wiki, n.d.). The coordinate list was then submitted to the Open Source Routing Machine (OSRM) Table service, which computes pairwise routing durations on an OpenStreetMap road network (Luxen and Vetter, 2011). The resulting input is a directed 100 x 100 travel-time matrix measured in seconds, with one row and one column for each address. The matrix is directed because travel times in an urban street network can differ by direction due to one-way streets, turn restrictions, and local traffic layout. Travel time is used as the optimization cost instead of physical distance because the delivery procedure depends on actual driving in a city environment, not just distance.

3.2 Optimization Workflow

Before applying the full workflow, a small GAMSPy/CPLEX instance was used to verify that the fixed start and fixed terminal route structure behaved as intended. However, solving the full 100 node, three-vehicle instance with exact MIP solvers such as CPLEX was not practical within the available time and cost constraints, so the final workflow used cuOpt as a routing solver for the real dataset. cuOpt was selected because it supports vehicle-routing style inputs and can be accessed in a workflow suitable for repeated daily planning. cuOpt was released as open-source software in 2025, and its algorithmic approach combines GPU acceleration with metaheuristic methods for large-scale vehicle-routing optimization (Neskovic, 2025; Çördük, Sielski, and Chung, 2024). In this implementation, the Managed Service API access model described by NVIDIA (n.d.) allowed the prepared data to be passed to the solver without maintaining a local optimization instance. The cuOpt solver uses the travel-time matrix, three vehicles, where each vehicle has a capacity of 40 units, the shared depot as the start node, and the shared headquarters as the terminal node. The three vehicles correspond to the actual number of vehicles available for the daily delivery operation, so they represent the real fleet size. The capacity limit of 40 delivery nodes per vehicle was selected according to the observed original route structure, where no route exceeded 40 stops. cuOpt returns the ordered sequence of stops for each vehicle, and those sequences shall then be evaluated against the retained baseline used before optimization. The next section uses this workflow to compare the current delivery arrangement with the reorganized solution.

4 Results

The latest run evaluates the reorganized three-vehicle solution against the original delivery arrangement. Costs are measured as travel time in seconds. The comparison uses the same empirical dataset as described in the previous chapter, where the baseline represents the current assignment used by the organization, while the reorganized solution allows the nodes to be reassigned across the three routes. The section therefore reports only the aggregate effect of the whole joint reorganization run. The main comparison is summarized in Table 1.

Indicator	Original arrangement	Reorganized solution
Total travel-time (s)	14,228.4	12,266.5
Vehicles used	3	3
Absolute saving (s)		1,961.9
Relative saving		13.79%

Table 1 Aggregate comparison of the original and reorganized delivery arrangements

At the aggregate level, Table 1 shows that the selected reorganization improves the objective value without changing the number of vehicles. The optimized solution reduced the total travel time from 14,228.4 s to 12,266.5 s, an absolute saving of 1,961.9 s and a relative saving of 13.79%. Because the delivery operation is repeated daily, the travel time reduction accumulates across service days rather than remaining a one-time saving.

The selected solution assigns all 98 delivery customers across the three available vehicles while respecting the capacity limit of 40 deliveries per vehicle. The capacity of vehicles was chosen because no route in the original assignment exceeded 40 delivery nodes. Table 2 reports these route allocation counts, where depot and terminal endpoint are excluded and it can be seen that the assigned number of nodes did not change drastically when comparing the original and reorganized routes.

	Number of delivery nodes	
Vehicle route	**Original arrangement**	**Reorganized solution**
VEHICLE-1	34	39
VEHICLE-2	34	32
VEHICLE-3	30	27

Table 2 Route-level allocation of delivery-customer nodes used for capacity verification

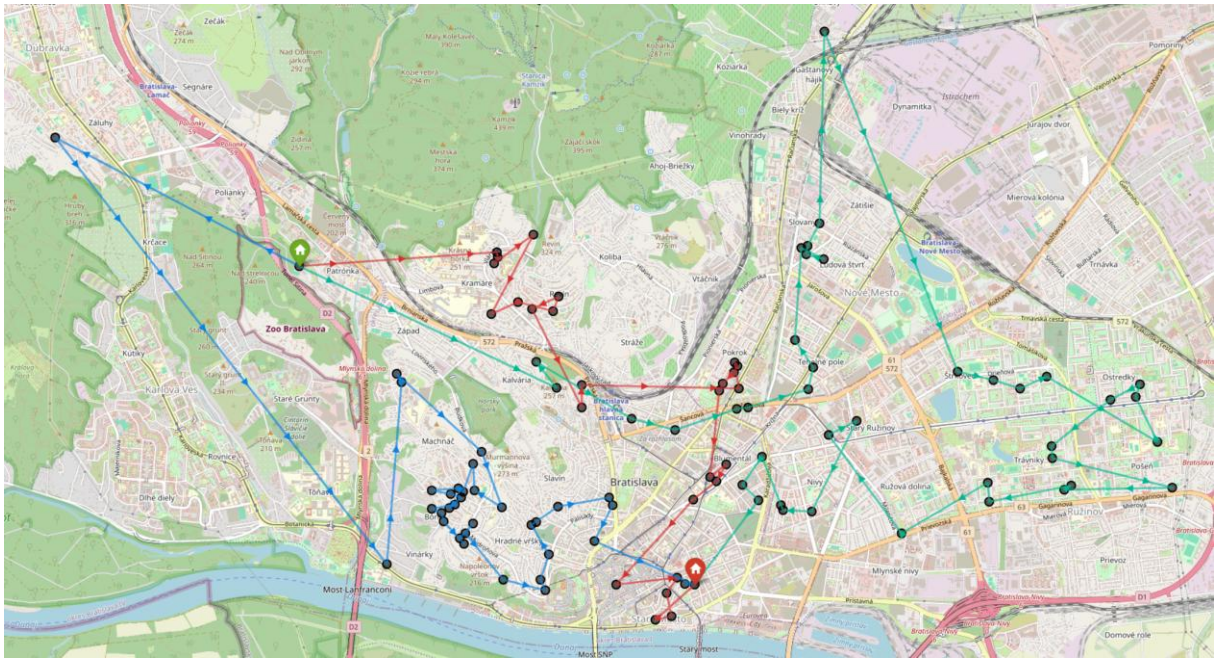


Figure 1 Anonymized map of the reorganized delivery routes

Figure 1 is used as a qualitative check on the spatial structure of the selected solution. It shows that the three vehicle routes cover the customers across the western, central, and eastern parts of the service area while retaining the common start and terminal node logic described in the model. The visualization is intentionally anonymized as it distinguishes route colors and delivery points, but it does not disclose street addresses or complete route order. Because cuOpt can return slightly different feasible solutions across runs, the reported result should be interpreted as the best run available. The map should therefore be read together with the aggregate table, which is the main numerical evidence in this section. For the studied organization, the result indicates that joint route reorganization can reduce daily driving time while preserving the fixed depo and terminal, and capacity structure required for the operation.

4.1 Discussion

The saving of 1,961.9 seconds is generated each day the delivery process is repeated. Expressed in operational terms, the result corresponds to approximately 32.7 minutes of reduced driving time per service day and over a month (21 working days), this equals about 11.4 hours of reduced travel time for the three vehicle fleet. The exact monetary value cannot be calculated from the present data alone, because it depends on wage rates, fuel prices, vehicle operating costs, and the number of actual service days, but the time saving gives a good basis for such an estimate. However, the benefit is not limited to direct cost savings. The released time can also be used to absorb additional customers, reduce overtime pressure, improve schedule reliability, or maintain the same service level without increasing the fleet size.

Although the presented case concerns meal delivery, the same model structure is relevant for other routing tasks in which vehicles do not return to their starting depot. Examples include delivery service performed by contract drivers who pick up goods at a depot and finish at a predetermined point or at drivers' homes, school bus or employee shuttle routing where passengers are collected and the route ends at a school or workplace, and

healthcare logistics in which samples are collected from multiple sites and transported to a laboratory or a hospital. In such settings, modeling the operation as a closed tour would not represent the real endpoint requirement. The FDOVRP structure therefore provides a closer representation of operations where the end of the route is operationally meaningful, not just a return to the start.

5 Conclusion

This paper addressed the daily delivery-route reorganization problem of a nonprofit organization as a Fixed-Destination Open Vehicle Routing Problem. In this problem, all vehicles start at the depot, serve assigned delivery customers, and finish at a predetermined headquarters node. The empirical case used 98 customer nodes, one depot, one fixed terminal headquarters, and three available vehicles, so the analysis focused on joint route reorganization rather than separate route-by-route changes. The empirical reorganization of three routes reduced total travel time from 14,228.4 s to 12,266.5 s, which corresponds to an absolute saving of 1,961.9 s and a relative saving of 13.79%. The daily saving of 32.7 minutes equals about 11.4 hours over 21 working days, which can free capacity without increasing fleet size. The same structure may also apply to contract delivery, shuttle routing, or healthcare sample transport where routes end somewhere other than the starting depot. Methodologically, the study contributes an FDOVRP formulation for operations in which vehicles start from a common depot but must finish at a predetermined headquarters node, rather than returning to the depot or ending at freely chosen locations. This result closes the main question of the paper by showing that the modeled problem structure can be preserved while improving the travel time objective.

The result shows that meaningful efficiency gains can be achieved without changing the number of vehicles or relaxing the fixed depot and terminal node structure required by the operation. Because the delivery process is repeated daily, the time savings are substantial. The workflow is also suitable for nonprofit logistics with constrained resources because it relies on address-level data, a Nominatim geocoding step, an OSRM travel-time cost matrix, and cuOpt routing runs that can be repeated when the address set or requirements change. In practical terms, the section demonstrates an optimization procedure that can compare the retained baseline with a reorganized three-route plan while keeping the operational assumptions transparent. A limitation of the workflow is that the OSRM travel-time matrix is derived from OpenStreetMap road data and a routing profile. It does not capture live congestion, temporary disruptions, or daily variation in travel speeds. Future work could incorporate historical or real-time traffic data to evaluate whether traffic adjusted travel time estimates would change the selected routes.

Acknowledgements

This work was supported by VEGA 1/0435/26 Investment Effectiveness and Spatial Spillover Effects in the EU Green Transformation: A Spatial Econometric and Machine Learning Approach and as part of the grant tasks for young teachers, researchers, and doctoral students within project A-25-103/3020-08 Examining Interdisciplinary Strategies for Effective Environmental Management and Sustainable Development.

References

Aksen D, Özyurt Z, Aras N (2007) Open vehicle routing problem with driver nodes and time deadlines. Journal of the Operational Research Society 58(9):1223-1234. https://doi.org/10.1057/palgrave.jors.2602249

Bögl M, Doerner K F, Parragh S N (2015) The school bus routing and scheduling problem with transfers. Networks 65(2):180-203. https://doi.org/10.1002/net.21589

Braekers K, Ramaekers K, Van Nieuwenhuyse I (2016) The vehicle routing problem: State of the art classification and review. Computers & Industrial Engineering 99:300-313. https://doi.org/10.1016/j.cie.2015.12.007

Çördük A, Sielski P, Chung M (2024) Record-breaking NVIDIA cuOpt algorithms deliver route optimization solutions 100x faster. NVIDIA Technical Blog. https://developer.nvidia.com/blog/record-breaking-nvidia-cuopt-algorithms-deliver-route-optimization-solutions-100x-faster/ Accessed 20.01.2026

Li F, Golden B, Wasil E (2007) The open vehicle routing problem: Algorithms, large-scale test problems, and computational results. Computers & Operations Research 34(10):2918-2930. https://doi.org/10.1016/j.cor.2005.11.018

Li L Y, Fu Z (2002) The school bus routing problem: A case study. Journal of the Operational Research Society 53(5):552-558. https://doi.org/10.1057/palgrave.jors.2601341

Luxen D, Vetter C (2011) Real-time routing with OpenStreetMap data. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, pp 513-516. https://doi.org/10.1145/2093973.2094062

Neskovic G (2025) NVIDIA open-sources cuOpt, ushering in new era of decision optimization. NVIDIA Blog. https://blogs.nvidia.com/blog/cuopt-open-source/ Accessed 20.01.2026

NVIDIA (n.d.) NVIDIA Managed Service Overview. NVIDIA cuOpt documentation. https://docs.nvidia.com/cuopt/user-guide/25.02/mg-server-overview.html Accessed 05.02.2026

OpenStreetMap Wiki (n.d.) Nominatim. https://wiki.openstreetmap.org/wiki/Nominatim Accessed 05.02.2026

Toth P, Vigo D (eds) (2002) The vehicle routing problem. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9780898718515

Gradual optimizing a sustainable investment portfolio: sequential mean-semivariance-ESG model

Adam Borovička[1]

Abstract. The two-factor view of investment, as a cornerstone of modern portfolio theory, has recently been challenged by the concept of sustainability. It is therefore advisable to expand an investor's risk-return profile to include an ESG (Environmental, Social, and Governance) perspective. Mean-risk models are then expanded into a three-objective form. This modification also applies to a well-known mean-semivariance model, which does not penalize positive deviations and is not limited by potential non-normality in the return distribution. The mean-semivariance-ESG model, however, reflects all three characteristics simultaneously. The final investment decision is thus largely dependent on the investor's additional preferences, typically in the form of threshold values or the weights of the objectives, which may be uncomfortable for some, particularly less experienced individuals.
Not only in the Czech environment, the growing popularity of sustainability, or social impact, has not yet reached a similar level of importance as return and risk characteristics. Simultaneous consideration of these characteristics may therefore not be a necessary optimization path. This assumption allows for the development of a sequential mean-semivariance-ESG model. First, the set of efficient portfolios is defined using a "classical" mean-semivariance model. Subsequently, an investment with the maximum level of the ESG metric is identified within this space. This approach therefore does not require any additional investor preference in the three-objective space. The user-friendliness of the proposed approach of gradual optimization is demonstrated in a case study on the selection of a portfolio of highly popular mutual funds. The empirical results of non-traditional sequential and conventional simultaneous optimization procedures are critically compared.

Keywords: ESG, mean-semivariance, mutual fund, portfolio, sequential

JEL Classification: C44, C61, G11

1 Introduction

The concept of sustainability is undoubtedly a defining feature of our times, influencing various aspects of human life. Investment decisions are no exception. According to a survey conducted annually by Perfect Crowd for the Czech Capital Market Association (CCMA), as investment appetite grows, the aspect of sustainability is increasingly becoming a key consideration for investors (CCMA, 2026). This perspective is often reflected through the concept of ESG (Environmental, Social, and Governance).

The methodology and tools used to support investment decisions are naturally adapting to the growing importance of sustainability. This is also true of the model pillar of modern portfolio theory, which has been transformed into a mean-variance-ESG version. Mean-variance portfolio problem was thus augmented through a linear ESG utility by (e.g) Blitz et al. (2024), or Garcia-Bernabeu et al. (2024). A nonlinear ESG adjustment of mean-variance model has been proposed by Müller and Joubrel (2025). Tis concept was further critically analyzed by Müller (2026). Since there can be dozens of ESG indicators, mean-variance-ESG models are developed with an associated procedure for selecting "sustainable" characteristics (Goodarzi et al., 2022, or Fereydooni et al., 2024). In light of relatively recent research on the incorporation of ESG factors into the cornerstone of modern portfolio theory, other mean-risk models have not yet undergone a more thorough review. The incorporation of sustainable behavior has been partially implemented in the case of the mean-semivariance model, which removes the limitations on penalizing positive deviations from the average return and, by extension, on the assumption of normality of returns. A sort of mean-semivariance-ESG model is being developed by Brătianu (2024) or Abate et al. (2023).

[1] Prague University of Economics and Business, Department of Econometrics, W. Churchill Sq. 4, Prague, Czech Republic, adam.borovicka@vse.cz.

What do all the models mentioned above have in common? They aim to identify the most suitable combination of return, risk, and ESG factors, typically based on additional information provided by the investor. These are thus reference (threshold) values for selected characteristics or their respective weights. Moreover, not only in the Czech context, while the ESG factor is certainly taken into account, the risk-return profile holds an unshakable position. Under such conditions, the provision of additional information is potentially problematic, even misleading, especially for less experienced/informed investors.

I propose a solution in the form of a sequential investment decision-making process. In the first phase, a set of efficient portfolios would be determined using a mean-semivariance model, i.e., without the influence of the sustainability factor. In the second phase, a portfolio with the highest ESG factor value would then be identified. This results in a proposal of a sequential mean-semivariance-ESG model that enables the maximization of investment sustainability within the space of mean-risk efficient investment opportunities. This process thus reflects the significant role of ESG criteria within the context of the core mean-risk perspective, thereby naturally reflecting the typical (inexperienced) prioritization of investment objectives without requiring further intervention.

In addition to the primary aim of designing a user-friendly methodology that reflects the return-risk-ESG perspective, a secondary, equally important objective is to demonstrate the proposed approach in a typical investment scenario. The empirical study thus focuses on the selection of an investment portfolio of mutual funds, which are attracting an increasingly broad (lay) investment audience. The Czech collective investment market provides a representative illustration of the empirical advantages of the sequential approach over standard simultaneous three-objective optimization process. It also critically examines the still-debated (conflicting) relationship between returns and ESG factors.

Following the introduction, Section 2 describes the sequential model. Section 3 presents an empirical investment study involving mutual funds. Section 4 then summarizes the findings and outlines further possible ideas and directions for research.

2 Sequential mean-semivariance-ESG model

The following section provides a necessary description of the mean-semivariance model. An innovative concept of sequential optimization using this model may then be introduced.

2.1 Mean-semivariance model

H. Markowitz himself was critical of his original (diversification) concept as incorporated into the mean-variance model (Markowitz, 1952). In the context of investment reality, the assumption of non-normal returns seems very strong. Penalizing positive deviations from the average return may appear irrational from an investment perspective. To limit the shortcomings of the original approach, H. Markowitz introduced the concept of semivariance (Markowitz, 1959). Bollestero (2005) effectively incorporated this idea into the following model

$$\begin{aligned} &\min \quad \sum_{i=1}^{n}\sum_{j=1}^{n} sem\sigma_{ij} x_i x_j \\ &\sum_{i=1}^{n} r_i x_i \geq r^* \\ &\sum_{i=1}^{n} x_i = 1 \\ &x_i \geq 0 \qquad i = 1,2,...,n, \end{aligned} \tag{1}$$

where x_i, $i = 1,2,...,n,$ or x_j, $j = 1,2,...,n,$ denotes the share of the *i*-th, or *j*-th asset in the portfolio, r_i, $i = 1,2,...,n,$ is a return of the *i*-th asset and r^* defines a reference level for the portfolio return. $sem\sigma_{ij}$, $i, j = 1,2,...,n,$ denotes a semicovariance between return of the *i*-th and *j*-th asset. The following holds for each couple of assets

$$sem\sigma_{ij} = \sigma_{ij} - b_{ij} \qquad i, j = 1,2,...,n, \tag{2}$$

where σ_{ij}, $i, j = 1,2,...,n,$ is a covariance between return of the *i*-th and *j*-th asset. b_{ij}, $i, j = 1,2,...,n,$ reflecting the positive return cases (above mean), is calculated as

$$b_{ij} = \beta_i \beta_j v\left(\mathbf{r_m} > r_m^o\right) \qquad i, j = 1, 2, ..., n, \tag{3}$$

where the following holds

$$\begin{aligned} \beta_i &= \frac{\sigma_{im}}{\sigma_m^2} \qquad i = 1, 2, ..., n, \\ \beta_j &= \frac{\sigma_{jm}}{\sigma_m^2} \qquad j = 1, 2, ..., n \\ v\left(\mathbf{r_m} > r_m^o\right) &= \frac{1}{p} \sum_{k=1}^{p} \left[\max\left(r_{m_k} - r_m^o, 0\right) \right]^2 . \end{aligned} \tag{4}$$

The vector $\mathbf{r_m} = (r_{m_1}, r_{m_2}, ..., r_{m_p})^T$ contains the returns of the market portfolio (m) over p time periods. The expected return (mean) of the market portfolio is $r_m^o = \frac{1}{p} \sum_{k=1}^{p} r_{m_k}$. The covariance between the return of the i-th, or j-th asset and the return of the market portfolio is reflected by σ_{im}, $i = 1, 2, ..., n,$, or σ_{jm} $j = 1, 2, ..., n$. Finally, σ_m^2 represents the variance of the portfolio return and r_{m_k}, $k = 1, 2, ..., p$, reflects the return of the market portfolio in the k-th time period.

Although this model is widely used, its implementation is not without obstacles. I see the main challenge in determining the market portfolio, as the results can understandably be sensitive to its composition. Ballestero (2005) points out the naive strategy, representing all considered assets with equal weight. However, the market portfolio can replicate the current composition of a benchmark (stock) index through proportional representation. The genesis of the market portfolio is determined by the specific investment situation, specifically the investment strategy, the portfolio of the preselected assets, the investment time horizon, or the investor's experience.

2.2 Mean-semivariance model in a sequence with an ESG factor

The output of the mean-semivariance optimization process is the identification of efficient portfolios—that is, investments with the best possible simultaneous combination of two key characteristics reflecting investment performance and the associated risk. On this set, we are finding a solution with the highest possible value of the function expressing the level of sustainability as measured by the relevant ESG factor. Formally, the following mathematical model is thus solved

$$\begin{aligned} &\max \quad \sum_{i=1}^{n} s_i x_i \\ &\mathbf{x} \in E \end{aligned} \tag{5}$$

where s_i, $i = 1, 2, ..., n,$ denotes the ESG score of the i-th asset and E is the set of efficient portfolios according to model (1) represented by vector of shares $\mathbf{x} = (x_1, x_2, ..., x_n)^T$. The solution thus represents a portfolio offering the best possible combination of return, risk, and ESG factors, in line with sequential three-factor optimization.

3 Gradual optimizing an investment portfolio with ESG mutual funds

The empirical section demonstrates the potential of the proposed methodological concept for selecting an investment portfolio of mutual funds within the Czech capital market.

3.1 Investment strategy and data

According to CCMA (2026) and other surveys, this has emerged as one of the most common investment strategies aimed at providing at least partial financial security in retirement. In an era of a deficit-ridden, permanently unsustainable pension system, this largely long-term investment strategy is not surprising. While the purpose of the investment may, on the one hand, lead to a more cautious attitude toward risk, on the other hand, it appears that this constructive mindset is becoming increasingly prevalent among the younger generation, which exhibits greater signs of boldness. It should be noted that the role of environmental, social, and sustainability responsibilities is also becoming increasingly entrenched.

According to the survey mentioned above, mutual funds are the most common form of investment, primarily due to their accessibility and professional management, despite higher fees. Virtually every bank in the Czech market offers its own funds or acts as an intermediary for affiliated institutions. Let's focus on clients of Česká spořitelna, which, with approximately 4.6 million clients, holds a dominant position in the Czech banking sector (Česká spořitelna, 2026). This institution is not the only one to categorize mutual funds according to ESG perspective.

Table 1 lists all selected ESG mutual funds along with their key characteristics. Due to the long-term nature of the investment strategy, a five-year historical period from May 2021 to April 2026 has been selected for their valuation. While this period largely represents good times in the financial markets, it is not without its share of (negative) shocks or periods lacking strong trends. Return is measured as the average monthly return. Individual risk is represented by the (standard) semideviation, which characterizes negative volatility. Of course, the semicovariances of returns are calculated in the background. Sustainability (labeled ESG) is measured using the internally developed ESGenius concept (Erste Asset Management, 2026). The necessary data on share prices, as well as links to the funds' annual reports containing ESG data, are available on the website of the Investment Center of Česká spořitelna (2026).

Mutual fund	**Return** [%]	**Risk** [%]	**ESG [score]**
Erste Bond Euro Corporate	-0.13	1.52	63
Erste Bond Europe High Yield	0.04	1.44	58
Erste Responsible Bond	0.11	1.25	69
Erste Future Invest	-0.02	3.60	67
Erste Responsible Stock America	0.73	3.60	63
Erste Responsible Stock Europe	0.39	2.96	69
Erste Responsible Stock Global	0.63	3.22	63
Erste Stock Biotec	0.33	4.56	57
Erste WWF Stock Environment	-0.39	6.02	85
ESG Mix 10	0.21	0.77	74
ESG Mix 30	0.24	1.58	72
ESG Mix 50	0.27	1.97	69

Table 1 Investment characteristics of the selected mutual funds

3.2 Sustainable portfolio making

Using the mean-semivariance model (1), an efficient frontier was determined (see Figure 1), which represents all efficient investment decisions within the return-risk profile.

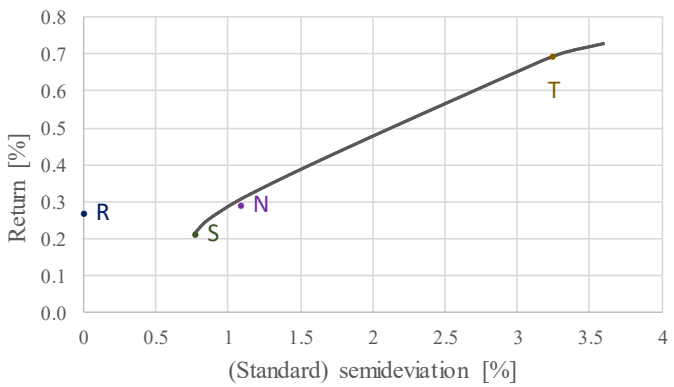


Figure 1 Efficient frontier and selected key points

It should be noted that negative covariance relationships were not particularly expected. Although these are mutual funds with different focuses (bond, equity, and balanced), they have performed well over the long term. Short-term opposing behavior thus tends to (merely) weaken the positive correlation. This fact potentially indicates that mutual funds have rather limited diversification capabilities.

The funds capable of forming effective portfolios in terms of risk-return profile are the equity funds *Erste Responsible Stock America* and *Erste Responsible Stock Global*, and the mixed fund *ESG Mix 10* with a smaller equity component. While the return of ESG Mix 10 is not among the highest, it has by far the lowest volatility. Furthermore, together with Erste Stock Biotec, it has the lowest semicovariance among returns. Unfortunately, this apparent diversification strength is undermined by the fund's second-highest return variability, which prevents it from playing a more significant role in the investment. As the demand for higher returns grows, the share of the two aforementioned equity funds naturally increases at the expense of the mixed fund.

3.3 Sequential ESG perspective

Figure 1 highlights several points worth noting, the characteristics of which are summarized in Table 2.

Point	**Return** [%]	**Risk** [%]	**ESG [score]**
Risk-free (R)	0.29	0	x
Sortino (T)	0.7	3.28	63
Sequential (S)	0.21	0.77	73.87
Naive mean-semv-ESG (N)	0.3	1.06	72.16
ESG prior	-0.39	6.02	85

Table 2 Key investment points

Using model (5), a portfolio (see point S) with the following composition is obtained: 0.74 % *Erste Stock Biotec* and 99.26 % *ESG Mix 10*. The low share of the equity fund is due to the lowest correlation of returns. Dominance of the balanced fund is due to the extremely low semivariance of its returns. This fact makes it the portfolio with the lowest risk. Without additional information (reference levels or characteristic weights), we obtain a return-risk efficient portfolio (even with a minimal level of risk) with the highest possible ESG factor value.

Regardless of the ESG factor and additional preferences, we can traditionally obtain a portfolio with the best "return-risk" ratio by maximizing the Sortino ratio. It consists of *Erste Responsible Stock America* (69.10 %) and *Erste Responsible Stock Global* (30.90 %). The investment then clearly reaches the near-opposite end of the effective return-risk spectrum (see point T)[2] which is due primarily to a flatter efficient frontier and, consequently, a relatively higher risk-free rate of return. An interesting comparison is the preference-based decision-making in the three-dimensional space of the "standard" mean-semivariance-ESG model, which simultaneously optimizes all three observed characteristics. Point N represents a naive strategy with equal weights for the characteristics. A portfolio, composed of funds *Erste Responsible Stock America* (17.04 %) and *ESG Mix 10* (82.96 %), lies just below the efficient frontier, which demonstrates a weak, risk-negative impact of ESG factors. As the demand for higher ESG increases, the result is ultimately a one-asset portfolio (*Erste WWF Stock Environment*) with the lowest (negative) return and the highest risk (see ESG prior in Table 2).

This empirical analysis supports the idea that sacrificing part of the return is necessary to ensure the sustainability of the investment. More importantly, however, it demonstrates the considerable utility of a designed sequential approach in cases where an investor wishes to simultaneously consider return, risk, and, secondarily, sustainability, but is unwilling or hesitant to prioritize these factors quantitatively. If, after all, an additional preference does emerge, the investment decision, naturally shifts to a different value spectrum.

4 Conclusion

The purpose of this article is to assist in decision-making, particularly for less experienced/informed investors who are unable or reluctant to express more specific preferences. In addition to two key investment characteristics, such investors are also concerned with the sustainability of their investment activities. For these purposes, the "standard" three-factor mean-variance-ESG model is not suitable, as it requires additional preference information to make a final decision. Furthermore, the variance concept exhibits well-known shortcomings. The goal is therefore to integrate the ESG factor into a mean-semivariance model, a topic not yet widely discussed in the current literature. A sequential mean-semivariance-ESG model is thus proposed, and its user-friendly implementability was demonstrated using the case of investment decision-making with mutual funds. The empirical analysis is supplemented by a comparable "unweighted" application of the mean-semivariance model

[2] The risk-free rate of return is derived from the current interest rate on Czech treasury bills of 3.41% p.a., which corresponds to a monthly rate of approximately 0.29 % (see point R).

(without ESG factor). Since this methodology provides a portfolio with a higher ESG rating and lower return, the aforementioned idea of a trade-off between returns and sustainability is confirmed. Understandably, the ESG Mix 10 fund stands out among the three most frequently included funds (along with the Erste Responsible Stock America/Global) in risk-return efficient portfolios, as it exhibits the highest ESG score but also the lowest return among them. Although the substitution relationship is underscored by the extreme case of the Erste WWW Stock Environment fund, it cannot be generalized. After all, some empirical studies show the opposite.

In future research, the model could, if necessary, be expanded to include preferences regarding the number of assets, their sector classification, and similar factors, thereby effectively moving beyond the "preference-free" framework. The proposed concept could also be extended to other risk measures (VaR, etc.).

Acknowledgements

The research project was supported by Grant No. F4/18/2024 of the Internal Grant Agency, Faculty of Informatics and Statistics, Prague University of Economics and Business.

References

Abate G, Basile I, Ferrari P (2024) The integration of environmental, social and governance criteria in portfolio optimization: An empirical analysis. Corporate Social Responsibility and Environmental Management 31: 2054–2065. https://doi.org/10.1002/csr.2682

Ballestero E (2005) Mean-Semivariance Efficient Frontier: A Downside Risk Model for Portfolio Selection. Applied Mathematical Finance 12:1–15. https://doi.org/10.1080/1350486042000254015

Blitz D, Chen M, Howard C, Lohre H (2024) 3D Investing: Jointly Optimizing Return, Risk and Sustainability. Financial Analysts Journal 80: 22–40. https://doi.org/10.1080/0015198X.2024.2335142

Brătian V (2024) Portfolio Optimization Under the Mean-semivariance Behavioral Hypothesis. Empirical Evidence of the Dependence Between Optimal Portfolio Structure and ESG Risk Scores. Management of Sustainable Development 16: 1–13. https://doi.org/10.54989/msd-2024-0011

CCMA (2026) Czechs are showing growing appetite for investing and interest in modern investment methods. (in Czech) https://www.akatcr.cz/Dokumenty/Aktuality/tiskov225-zpr225va-akat-u-cechu-roste-investorsk221-apetit-i-z193jem-o-modern205-zpusoby-investov193n205 [cited 2026-05-15]

Česká spořitelna (2026) Who we are. https://www.csas.cz/en/about-us/who-we-are [cited 2026-05-15]

Erste Asset Management (2026) Erste AM ESGenius® Score. https://www.erste-am.at/en/private-investors/sustainability/esgenius-score [cited 2026-05-15]

Fereydooni A, Barak S, Sajadi S M A (2024) A novel online portfolio selection approach based on pattern matching and ESG factors. Omega 123: 102975. https://doi.org/10.1016/j.omega.2023.102975

Garcia-Bernabeu A, Hilario-Caballero A, Tardella F, Pla-Santamaria D (2024) ESG integration in portfolio selection: A robust preference-based multicriteria approach. Operations Research Perspectives 12: 100305. https://doi.org/10.1016/j.orp.2024.100305

Goodarzi F, Abdollahzadeh V, Zeinalnezhad M (2022) An integrated multi-criteria decision-making and multi-objective optimization framework for green supplier evaluation and optimal order allocation under uncertainty. Decision Analytics Journal 4: 100087. https://doi.org/10.1016/j.dajour.2022.100087

Investment Center of Česká spořitelna (2026) Responsible investment. https://www.investicnicentrum.cz/en/products/esg-products [cited 2026-05-15]

Markowitz H M (1952) Portfolio selection. Journal of Science 7: 77–91

Markowitz H M (1959) Portfolio Selection: Efficient Diversification of Investments. John Wiley & Sons, Inc.

Müller L (2026) Diversification effects of ESG penalties in sustainable mean–variance portfolios. Journal of Asset Management 27: 12. https://doi.org/10.1057/s41260-026-00446-2

Müller L, Joubrel M (2025) A novel approach to sustainable mean-variance portfolio optimization: Accounting for ESG-related uncertainty. Finance Research Letters 85(C): 108056. https://doi.org/10.1016/j.frl.2025.108056

Seasonality and trend in multiproduct inventory volumes data testing and cost modelling

Tomasz Brzęczek[1]

Abstract. We examine multi-product sales seasonality to identify its volatility/dynamics cost and sum with classical stationary single-period inventory expected cost function. Two-way ANOVA is applied to trend and seasonality testing in product quarterly volume sales that can face both effects, one of them or none. Empirical case study of tyres and wheels equipment category for agriculture is analyzed. The minimal expected cost model is formulated. Instead solving it, we evaluate its upper bound as the expected cost of lost sales, overstock and seasonality for forecasted sales feasible solution. We show that seasonality cost per product equals 0-7% and increasing in a product price, demand/order and the seasonality deviation. For all product the cost accounts for 3%. Product pooling that compensates interseason variability leads to lower cost.

Keywords: Seasonality, trend, two-way ANOVA, sales forecasting, single-period inventory, expected cost of inventory

JEL Classification: C53, C8, L25, M21, M31

1 Introduction

This research aims at identification of demand caused stock adjustment and its financing in addition to classical single-period inventory model. Hence, including not only lost sales expected cost and overstock keeping expected cost but also the financial cost of mean demand seasonal cyclicality. Literature review in section 2 shows the research gap. In section 3 we apply econometric model of sales time series that is crucial to identification of seasonality and its cost. In Empirical Results section we provide data description and the analysis of a wholesaler company's product category including 11 products. This empirical analysis is also to show structure of the total expected cost including purchase value. Seasonality cost is assumed to be financial cost of product category purchases and accounts for 6-14% on the aggregate total expected cost and purchase cost yearly. Expected cost of stockkeeping and lost sales accounts to 3-4%. The results are achieved for the service level aiming at the forecasted sales. For symmetrically distributed demand as the normal one that is used here the service level equals 50%.

2 Literature review

Demand unpredictability is pointed out as the main problem for warehouse operations by (Davarzani & Norrman, 2015). This research aims at identification of demand fluctuation and connected stock financing cost and expected overstock cost balancing with lost sales cost. We analyze agricultural equipment category stock items. Hence, a quarter seasons are found as appropriate period comparing to fast moving goods demand analysis and stock setting based on weekly forecasting period (Namwad et al., 2024) or a daily sales weekly seasonal cycle modelled by (Ehrenthal et al., 2014). Unlike others, we allow for positive and negative seasonality effects to compensate for a whole product category financing cost. Category net seasonality generates interest cost of credit line amount that is ensured by the bank and is to be used by a company at the amount that follows seasonal demand of product category. A hike in seasonal demand and the inventory leads to higher level of financing interests or own capital cost. We do not analyze interest rate progression or additional fees to be charged by suppliers due to order quantity change like in multi-period setting in (Urban, 2000). Furthermore, we estimate the lost sales cost as well as overstock keeping cost and their expected values with the newsvendor model application.

To capture for seasonal hike or low in demand one can also set EOQ with a budget seasonal constraint and the season length cap for the replenishment period. Sales linear trend with seasonality versus Brown's motion depletion rate process were used to solve dynamically economic order quantity (EOQ) and lead (replenishment) period in order to minimize sum of costs: order, keeping (depletion), shortage by (Pervin et al., 2018). (Mukherjee

[1] Poznan University of Technology/Logistics Institute, Engineering Management Department, 2nd Rychlewskiego street, 61-934 Poznań, Poland, tomasz.brzeczek@put.poznan.pl. ORCID: 0000-0002-2505-5089

et al., 2025) solved similar model replacing the depletion with ecological degradation cost. Seasonality compensation due product assortment width is discussed in (Brzeczek, 2016), (Brzeczek, 2020) but only between product categories.

The research elaborates jointly on the seasonal demand effect and stochastic fluctuation impact on expected total cost of inventory. The newsboy problem considers minimization of the expected inventory mismatch cost (identical to the maximization of the expected profit of a retailer) stationary for each season in the cycle. It is called also the classical single-period inventory model called newsvendor that was worked out analytically into numerous extensions in the literature (Zhou et al., 2015). The model is single-period so captures for inter seasonal effects of demand only in terms of various means of demand that we estimate with time-series forecasts. So, we capture for four quarters seasonal demand financing cost if product is seasonal. In the category we jointly analyze seasonal and non-seasonal products with the defined expected cost objective and apply it to the sales forecasted values as the order solution. Review of other cost objective extensions can be found in (Khouja et al., 2020), (Luo et al., 2015). Let's note that for budget constraint multi-product setting a lot of optimization models and numerical methods are proposed by (Abdel-Malek et al., 2004), (Abdel-Malek & Montanari, 2005), (Abdel-Malek et al., 2020) and (Zhang & Hua, 2010) although seasonality is not captured.

3 Methods

3.1 Sales volume model and a forecast

A product purchase should be done at the latest before the beginning of a quarter to receive an order. We use a time-series data from previous calendar years denoted with:
y_{it} : a sales volume observed value,
$i = 1, 2, \dots, I$: product index,
$t = 1, 2, \dots, T$: period index of a series that gives trend and seasonality calibration, we analyse $T = 28$ quarters,
$l \in \{1, 2, 3, 4\}$: quarter number,
$\hat{y}_{it}$: a sales volume from theoretical model or a forecast extrapolation of each product sales volume into a linear trend model of quarter t sales:

$$g(t) = b_1 t + b_0 \tag{1}$$

that under insignificant trend becomes an average $g(t) = \frac{\sum_{t=1}^{t=28} y_t}{28}$, that either include an additive season l effect:

$$\hat{y}_{it} = g(t) + q_l, \tag{2}$$

or multiplicative one:

$$\hat{y}_{it} = g(t) * q_l, \tag{3}$$

from which random realisation standard error is given with:

$$s(e_i) = \sqrt{\frac{\sum_{t=1}^{28} (y_{it} - \hat{y}_{it})^2}{T - K}} \tag{4}$$

and is used to set a vertical section at the graph to present sales volume confidence interval:

$$\hat{y}_{it} - s(e_i) \leq y_t \leq \hat{y}_{it} + s(e_i) \ . \tag{5}$$

We use this sales volume forecast to estimate the seasonality stock adjustment cost and connected stochastic demand fluctuation cost. To handle stochastic fluctuations the newsboy model is suitable as it concerns for expected cost of stock-keeping if overstock occurs as well as captures for lost sales expected opportunity cost.

3.2 Seasonality financing cost

In each quarter of the year the company orders and pertains stock level z_{it} based on the forecast so $z_{it} = \hat{y}_{it}$. Hence, quarterly purchase cost equals:

$$C_t = \sum_{i=1}^{I} c_i z_{it} = \sum_{i=1}^{I} c_i \hat{y}_{it}\,, \qquad t = 29, 30, 31, 32. \tag{6}$$

c_i: unit cost of product i is fixed during all seasons. Hence, denoting quarterly average purchase cost with $\bar{C} = (C_{29} + C_{30} + C_{31} + C_{32})/4$, we find seasonal variability of purchase cost in this basic scenario $s(C)$ as follows:

$$s(C) = \sqrt{\frac{1}{L}\sum_{t=29}^{t=32}(C_t - \bar{C})^2} \tag{7}$$

and using company financing yearly non-capitalized interest rate r we find linked turnover capital seasonality financing cost that accounts for:

$$\boldsymbol{\varphi} = s(C)r\,. \tag{8}$$

The rate of return r is fixed inter quarterly in our model. This cost is being attributable to stock running and financing variability during the demand cycle. This cost is annual phenomenon so its decomposition into stocking and selling periods like in $\boldsymbol{\varphi} = \varphi_{29} + \varphi_{30} + \varphi_{31} + \varphi_{32}$ is problematic. It could be valued equally as in the equation $\varphi_t \cong s(C)r/L$. Actually, deviations from average seasonality equal to 0 could be asymmetrical and need proportional inter seasons de-aggregation. Other approach could use unit seasonal cost c'_{it} per order unit as following:

$$c'_{it} = \frac{\boldsymbol{\varphi} c_i \hat{y}_{it}}{4\sum_{i=1}^{I} c_i \hat{y}_{it}}\ . \tag{9}$$

This would ease a lot the tractability of the total expected cost minimization that may not follow the forecasted order values $C_t = \sum_{i=1}^{I} c_i z_{it} \neq \sum_{i=1}^{I} c_i \hat{y}_{it}$.

3.3 Expected cost of inventory/sales mismatch to stochastic demand

Off course, in business practice random fluctuations of demand are usual and therefore apart from seasonal costs we should also assess expected cost of inventory stock mismatch to demand fluctuations. Hence, total category inventory expected cost $\boldsymbol{C}_t$ in quarterly outbreak accounts for:

$$\boldsymbol{C}_t = \varphi_t + \theta_t + C_t\ , \tag{10}$$

and consists of mentioned costs. Following the newsvendor-like model for a product in (Abdel-Malek et al., 2004) we define $\theta_t + C_t$ as the aggregate for all products:

$$C_t + \theta_t = \sum_{i=1}^{I}[c_i z_{it} + \theta(z_{it}, x_{it})] =$$

$$= \sum_{i=1}^{I}\left[c_i z_{it} + h_i \int_0^{z_{it}}(z_{it} - x_{it})f_i(x_{it})dx_{it} + k_i \int_{x_{it}}^{\infty}(x_{it} - z_{it})f_i(x_{it})dx_{it}\right], \tag{11}$$

- $\theta(z_{it}, x_{it})$: the expected mismatch cost of item i order quantity stock to stochastic demand x_{it},
- $z_{1t}, z_{2t}, \dots, z_{It}$: order quantities that are identical to stock quantities (no inter-period stock),
- f_i : probability density function, for example normal distribution density function,
- x_{1t}, x_{2t}, x_{It}: random variable representing stochastic demand with expected value $\hat{y}_{it}$ and with volatility standard deviation SE_i, so $x_{it} \sim (\hat{y}_{it}, SE_i)$.
- h_i : unit cost of stock keeping, Initially (11) was formulated in the literature in such a way that defines h_i as the utilization or loss fee that is a part or whole purchase cost c_i; we define its value smaller than a cost and interpret it as the sales out of quarterly unsold stock with a loss.
- k_i : unit cost of unsupplied demand (opportunity cost that equals lost profit, but we do not capture for a unit cost of unsatisfied demand).

3.4 The optimization model

Arisen stochastic multi-product and multi-period non-linear novel formulation as following:

$$\boldsymbol{C} = \boldsymbol{C}_{29} + \boldsymbol{C}_{30} + \boldsymbol{C}_{31} + \boldsymbol{C}_{32} =$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{I} c'_{it} z_{it} + c_i z_{it} + h_i \int_0^{z_{it}}(z_{it} - x_{it})f_i(x_{it})dx_{it} + k_i \int_{z_{it}}^{\infty}(x_{it} - z_{it})f_i(x_{it})dx_{it}\ . \tag{12}$$

Hereafter, in empirical results we just evaluate (7-8, 12) for the feasible solution of $z_{it} = \hat{y}_{it}$ instead minimization of (12) with the optimal solution z^*_{it}. This is because of two reasons. One is that scope of this work concerns methodology integration of demand and seasonality forecasting into the optimization model including seasonality cost minimization. Second reason is that solving procedure is to be worked out although Lagrangian first order conditions equation set size equals to IT. Showing that (12) is convex is done for classical NV in the literature. The sum of convex and linear seasonal function component is also covex then. The optimal solution and its minimal expected cost (12) is not solved just its upper bound is calculated in results as the expected cost of lost sales, overstock and seasonality cost value for feasible solution equal to forecast of sales $z_{it} = \hat{y}_{it}$.

4 Empirical results

4.1 Data analysis

The category includes $I = 11$ items. Item names are provided in Table 1.

Item number	Product
1	Tractor front tyre
2	Tractor front tyre
3	Trailer tyre
4	Truck tyre
5	Other tyre
6	Inner tube
7	Vents
8	Agriculture machine wheels
9	Single wheel trolley & equipment
10	Rims
11	Accumulators

Table 1 Items in product category

Time scope is $T = 28$ quarterly periods in years 2013-2019.

Each item sales time series follows steps of testing and demand estimation that are following:

- Perform two-way ANOVA for year number and for quarter number explanatory variables.
- If both vars significantly (for achieved F-statistics $p - \text{value empirical} \leq 0.05$) affect sales volume then detrend series by computing first order differences.
- Repeat Two-Way ANOVA for transformed data.
- Repeat steps 2 and 3 until the trend is not significant and seasonality significance test is viable.

Finally, ANOVA significance test of the seasonality averages variance result is not biased when a tendency exists in data $\hat{y}_{it}$

For product 1 and 2 results of testing and estimation of data are presented at Figure 1.

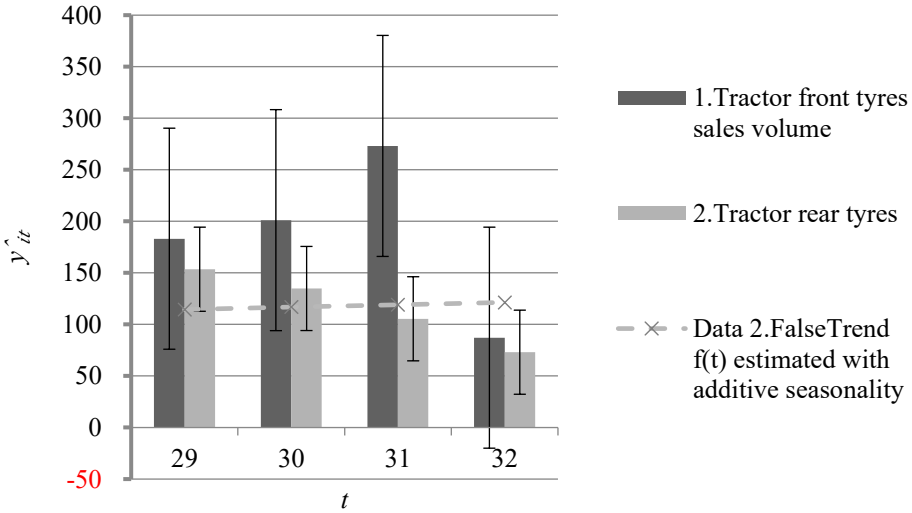


Figure 1 Sales forecasts: item 1 (product): tractor front tyre; item2: tractor rear tyre

Results are presented as the following:

- a bar that shows the forecasted expected sales $\hat{y}_{it}$ including additive seasonal component q_l;

- for $i = 1$ is regression of forecasted expected value is: $\hat{y}_{i=1;t} = b_{0;i=1} + q_{l;i=1}$, $t = 29, 30, 31, 32,$
- product two appropriate data model is of the same type $\hat{y}_{i=2;t} = b_{0;i=2} + q_{l;i=2}$, $t = 29, 30, 31, 32.$

However, for product 2 thanks to ANOVA we avoid false trend. This product raw sales observations show a time slope of regression line to be significant at *p-value* 0.1. Note that first quarter and fourth quarter false trend values are beyond no-trend forecast confidence intervals at Fig.1. While seasonal effects are smaller than SE.

Using two-way ANOVA deals also with second data issue that is seasonality identification only when series shows no trend. Otherwise, series is differentiated and examined again for seasonality. It was identified for item 3, item 8 and item 10 that testing across yearly and quarterly grouped averages with F statistics showed the trend and seasonality both are significant even if correlated. Results shown at Figure 2. Figure 2 presents trend and seasonality identification for products number 3-11 in the category. Let's also notice that all sales volume trends were matched with seasonality multiplicative while other series with additive by fitness to data measured with R-square ratio.

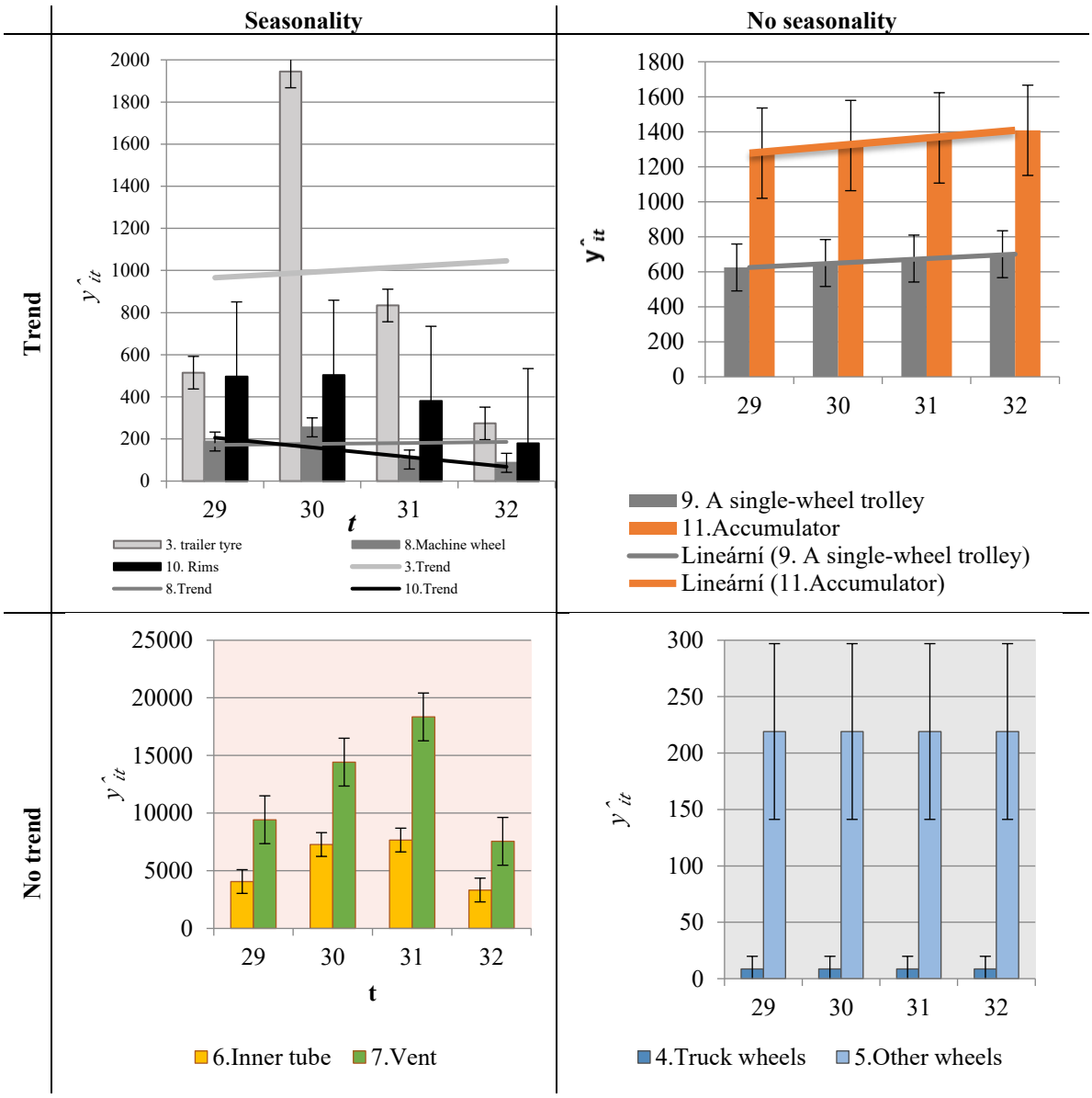


Figure 2 Classification of items according to seasonality and trend identified with Two-way ANOVA

4.2 Cost estimates

The results were achieved for the following unit cost values:

- $r = 10\%$ nominal per cent rate,

- Stock-keeping cost is declared by the company were about 10% of item unit cost: $\frac{h_i}{c_i} = 0.1$,
- Opportunity cost of lost sales accounts varies between items and on average $\frac{k_i}{c_i} = 0.2$;

The values of costs are prices are empirically collected and are presented in Table 2.

c_i	151.6	927.4	61.3	2264.0	40.0	21	1.71	289.92	16.35	75.50	210.70
p_i	173.8	1060.2	72.6	2655.1	48.5	26	2.45	344.10	22.24	88.47	249.14

Table 2 Cost c_i, price p_i (PLN) values for products

Item i =	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**Category**
$x_{i;t=29}$ [1]	183	154	515	8.73	219	4068	9431	188	625	496	1278	N/A [2]
$x_{i,30}$	201	135	1945	8.73	219	7281	14425	255	650	504	1321	N/A
$x_{i,31}$	273	106	834	8.73	219	7667	18348	102	675	380	1365	N/A
$x_{i,32}$	87	73	274	8.73	219	3330	7555	87	700	180	1408	N/A
$\theta(x_{i,29}, Y_{i,29})$ [3]	1445	3687	536	1984	149	2917	753	1493	403	2502	6125	21994
$\theta(x_{i,30}, Y_{i,30})$	1484	3682	536	1984	149	2917	753	1493	403	2514	6125	22040
$\theta(x_{i,31}, Y_{i,31})$	1569	3633	536	1984	149	2917	753	1453	403	2302	6125	21824
$\theta(x_{i,32}, Y_{i,32})$	1130	3379	536	1984	149	2912	753	1412	403	1964	6125	20747
$c_{i,29}x_{i,29}$ [4]	30502	142449	31551	19765	8763	85434	16127	54520	10219	37448	269275	706053
$c_{i,30}x_{i,30}$	41417	125106	119210	19765	8763	152904	24667	73950	10628	38052	278335	892797
$c_{i,31}x_{i,31}$	45289	97841	51106	19765	8763	161000	31375	29580	11036	28690	287606	772051
$c_{i,32}x_{i,32}$	13222	67793	16790	19765	8763	69929	13919	25230	11445	13590	296666	557112
φ_{29} [5]	N/A [5]	-	-	-	-	-	-	-	-	-	-	25191
φ_{30}	-	-	-	-	-	-	-	-	-	-	-	25191
φ_{31}	-	-	-	-	-	-	-	-	-	-	-	25191
φ_{32}	-	-	-	-	-	-	-	-	-	-	-	25191

[1] number of item units at stock is feasible solution $z_{it} = \hat{y}_{it}$,
[2] N/A - Not Applicable,
[3] expected mismatch cost (PLN) $\theta(z_{it}, x_{it})$ and the category mismatch expected cost as in (11).
[4] item purchase cost (PLN) and the category purchase cost C_t in the last cell of the row,
[5] seasonal supply variability cost φ_t (PLN) – hardly attributable to items as shown in (7, 8).

Table 3 Cost c_i, price p_i (PLN) values for products

In Figure 3 total inventory cost of product category is presented at left hand side vs. item 10 cost structure (right hand side).

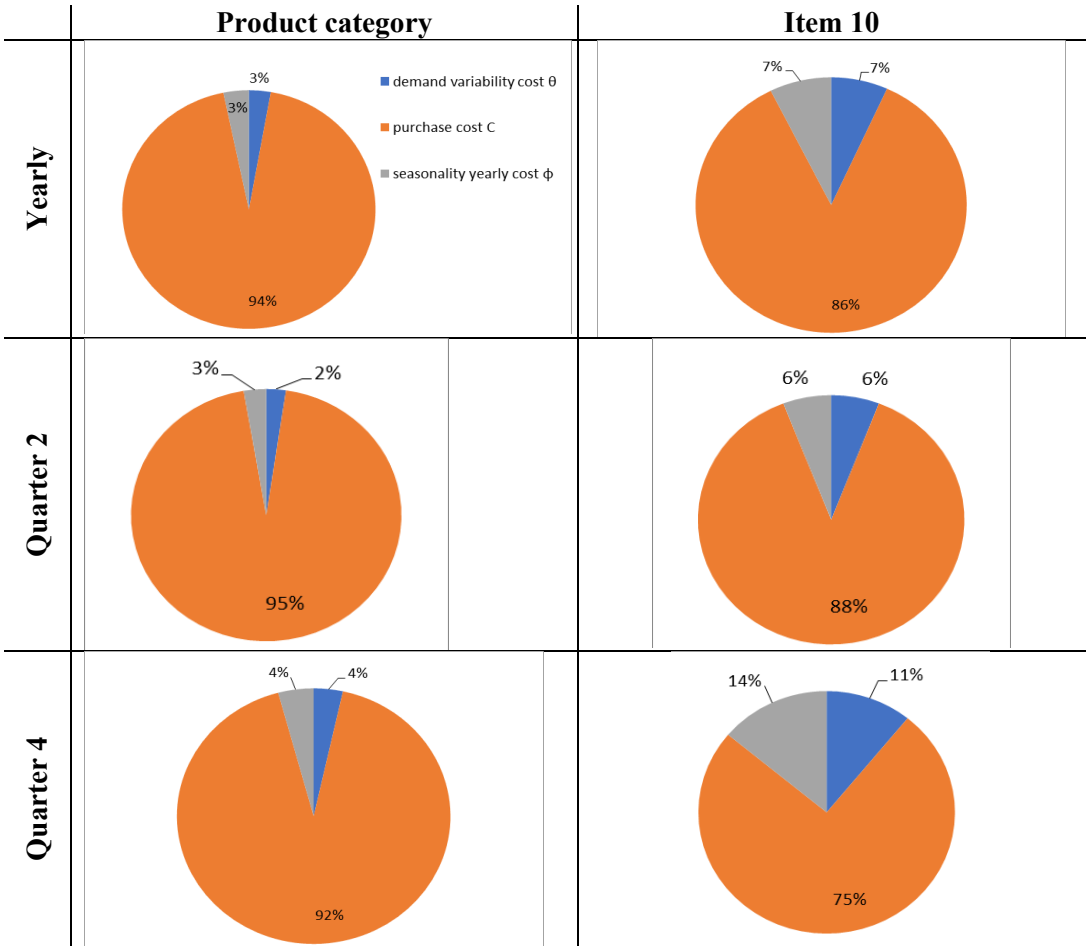


Figure 3 Cost C_t structure estimation for predicted quarterly mean demand to be stocked

5 Conclusion and discussion

Owing to this analysis method and results we come-up with findings:

- Thorough testing of quarterly sales for seasonality and trend with two-way ANOVA gives confidence in forecast performance until game-changer occurs in the future. Especially a quarter season.
- Demand seasonal variability cost as well as the cost of inventory mismatch to random demand contribute to the modelled cost function in a single-quarter as well as all quarters aggregated cost.
- Seasonal inventory management cost is product category attributable because of seasonal purchases and sales pooling compensation (Brzeczek, 2016). For example a few seasonal items low season takes place in quarter Q4 and for others in Q1 in the examined data. Some considerations for its decomposition into product unit cost give inventory optimization future work direction.
- Seasonal adjustment inventory cost is distributed uniformly into seasons because of whole year selling of a seasonal product considered here. Empirically checked for the company it accounts for 3% of total inventory yearly cost including purchase. These costs share increases in low season due to lower purchase. This cost itself doesn't have to change monotonically in a stock quantity,
- Seasonality cost depends on: interest rate, a number of seasonal products, seasonal effects,
- To compensate the seasonal sales variability a company can use pension plans quarterly contributions (Szczepański & Brzęczek, 2016) or financial assets risk (Uğurlu & Brzeczek, 2023).
- The random variability causes stock-keeping cash-cost. Is determined with h_i/k_i ratios.
- Mismatch cost is proportional to overstock and lost sales cost parameters.
- Optimization model (12) needs further operationalization and solving procedure especially if contains seasonal cost formulation (7) not (8).

Acknowledgements

Contribution to conference MME'26 Ostrava proceedings from Poznan University of Technology 0812/SBAD/4245.

References

Abdel-Malek L, Montanari R (2005) An analysis of the multi-product newsboy problem with a budget constraint. International Journal of Production Economics 97(3): 296–307. https://doi.org/10.1016/j.ijpe.2004.08.008

Abdel-Malek L, Montanari R, Morales L C (2004) Exact, approximate, and generic iterative models for the multi-product Newsboy problem with budget constraint. International Journal of Production Economics 91(2): 189–198. https://doi.org/10.1016/j.ijpe.2003.09.004

Abdel-Malek L, Shan P, Montanari R (2020) A Constructive Methodology to Solving the Capacitated Newsvendor Problem: an Approximate Approach. Operations Research Forum 1(2). https://doi.org/10.1007/s43069-020-0008-7

Brzeczek T (2016) Using portfolio theory to predict the impact of reduction in product with on sales. Journal of Business Economics and Management 17(6): 1222–1236. https://doi.org/10.3846/16111699.2015.1120769

Davarzani H, Norrman A (2015) Toward a relevant agenda for warehousing research: literature review and practitioners' input. Logistics Research 8(1). https://doi.org/10.1007/s12159-014-0120-1

Ehrenthal J C F, Honhon D, Van Woensel T (2014) Demand seasonality in retail inventory management. European Journal of Operational Research 238(2): 527–539. https://doi.org/10.1016/J.EJOR.2014.03.030

Khouja M, Christou E, Stylianou A (2020) A heuristic approach to in-season capacity allocation in a multi-product newsvendor model. Omega 95: 102252. https://doi.org/10.1016/J.OMEGA.2020.102252

Luo Z, Wang J, Chen W (2015) A risk-averse newsvendor model with limited capacity and outsourcing under the CVaR criterion. Journal of Systems Science and Systems Engineering 24(1): 49–67. https://doi.org/10.1007/s11518-015-5263-3

Mukherjee A K, Maity G, Jablonsky J, Roy S K, Weber G W (2025) Multi-objective sustainable inventory model with shortages and radio frequency identification under uncertain environment. International Journal of Systems Science. Operations & Logistics 12(1). https://doi.org/10.1080/23302674.2025.2470917

Namwad R S, Mishra N K, Ranu P, Jain P (2024) Optimizing Inventory Management with Seasonal Demand Forecasting in a Fuzzy Environment. Journal Europeen Des Systemes Automatises 57(4): 1091–1102. https://doi.org/10.18280/jesa.570416

Pervin M, Roy S K, Weber G-W (2018) Analysis of inventory control model with shortage under time-dependent demand and time-varying holding cost including stochastic deterioration. Annals of Operations Research 260(1–2): 437–460. https://doi.org/10.1007/s10479-016-2355-5

Szczepański M, Brzęczek T (2016) Zarządzanie ryzykiem w pracowniczych programach emerytalnych. Uwarunkowania instytucjonalne, ekonomiczno-fiskalne i demograficzne. PTE Poznań

Uğurlu K, Brzeczek T (2023) Distorted probability operator for dynamic portfolio optimization in times of socio-economic crisis. Central European Journal of Operations Research 31(4): 1043–1060. https://doi.org/10.1007/s10100-022-00834-0

Urban T L (2000) Supply contracts with periodic, stationary commitment. Production and Operations Management 9(4): 400–413. https://doi.org/10.1111/j.1937-5956.2000.tb00466.x

Zhang B, Hua Z (2010) A portfolio approach to multi-product newsboy problem with budget constraint. Computers & Industrial Engineering 58(4): 759–765. https://doi.org/https://doi.org/10.1016/j.cie.2010.02.007

Zhou Y, Chen X, Xu X, Yu C (2015) A Multi-Product Newsvendor Problem with Budget and Loss Constraints. International Journal of Information Technology and Decision Making 14(5): 1093–1110. https://doi.org/10.1142/S0219622014500448

Digital Development, Governance Quality, and Income Inequality: Empirical Evidence from Vietnamese Provinces

Han Bui Thi Ngoc[1], Thao Le Thi Kim[2], Thuy Thanh Tam Vo[3], Zdeněk Zmeškal[4]

Abstract. This study examines the impact of Information and Communication Technology (ICT) and institutional quality on income inequality across 63 Vietnamese provinces over the period 2018–2022. The System Generalised Method of Moments (System GMM) is employed as the main estimator to address endogeneity and dynamic effects. The results indicate that ICT development significantly reduces income inequality, as measured by the Gini coefficient. Institutional quality, proxied by the Provincial Governance and Public Administration Performance Index (PAPI), also contributes to lowering inequality. Furthermore, institutional quality strengthens the inequality-reducing effect of ICT through a significant moderating role. Key governance dimensions such as transparency, accountability, and corruption control play a crucial role, whereas citizen participation and public service delivery are not statistically significant. These findings emphasize the importance of combining digital development with institutional improvement to promote inclusive growth.

Keywords: ICT, income inequality, institutional quality

JEL Classification: C33, C44, O15, O33, O53

1 Introduction

Over decades, income inequality has remained a pressing concern for policymakers (Ho et al., 2025). With the overwhelm of digital technologies, technological breakthroughs, including AI, Big Data or Internet of Things, have reshaped the structure of labor markets and economic chances as they have been key drivers for the digital economy (Lee et al., 2026). The broad use of ICT has prompted numerous scholars to examine its impact on addressing income inequality. Findings reveal mixed empirical evidence from existing studies. Digitalisation was found to mitigate income inequality in G20 countries as its interactive effects with trade openness and foreign direct investment reinforce the narrowing of the income gap (Yin & Choi, 2023). However, some found that ICT widens inequality as those who have a higher education level are better equipped to use ICT, whereas disadvantaged groups may struggle to capture opportunities and are excluded from the game (Ho et al., 2025). Besides, previous studies also found the relationship between institutional quality and ICT. The growth of the Internet fosters transparency and reduces information asymmetries, thereby helping to curb corruption (Fan et al., 2021), and institutional quality plays a role as a mediator of e-government effects on corruption (Adam, 2020).

Despite its one-party system, Vietnam offers a unique context for this study, as significant subnational governance variations captured by the PAPI index provide a valuable setting to examine institutional roles. Concurrently, since the 1986 Doi Moi reforms, Vietnam has achieved rapid ICT expansion and ambitious digital economy goals. However, these gains remain unevenly distributed. Income inequality persists across regions, urban-rural divides, and ethnic minority groups (Nguyen, 2025). This paradox of digital advancement coexisting with persistent inequality makes Vietnam a compelling case for investigating the mechanisms of inclusive development. Upon these insights, this paper aims to contribute to the research gap by investigating the effect of ICT on income inequality and the moderating role of institutional quality in the context of Vietnam, with 3 research questions: Does ICT adoption affect income inequality? Does institutional quality moderate the relationship between ICT and income inequality? Does the structural evolution of institutional quality reduce income inequality in Vietnam? To answer these questions, we employed a panel dataset of 63 cities and provinces in Vietnam, prior to the

[1]Technical University of Ostrava, Economic, 17. listopadu 2172/15 ·597 321 111, Czech Republic, han.bui.thi.ngoc.st@vsb.cz

[2] Nong Lam University Ho Chi Minh City, Department of Business Administration, Quarter 33, Linh Xuan Ward. Ho Chi Minh City, Vietnam, thao.lethikim@hcmuaf.edu.vn

[3]Technical University of Ostrava, Economic, 17. listopadu 2172/15 ·597 321 111, Czech Republic, thuy.thanh.tam.vo.st@vsb.cz

[4]Technical University of Ostrava, Economic, 17. listopadu 2172/15 ·597 321 111, Czech Republic, zdenek.zmeskal@vsb.cz

administrative merger, from 2018 to 2022, and used system GMM methods. This study contributes to existing findings in 3 ways. Firstly, it is given evidence on the link between ICT, institutional quality and income inequality. Then it is tested the moderating role of institutional quality to comprehend the conditional channels through which ICT affects distributional outcomes. Eventually, we provide managerial implications for relevant parties.

2 Literature Review and Hypothesis Development

2.1 Theoretical Framework

This study is grounded in Skill-Biased Technological Change (SBTC) and New Institutional Economics. At its core is Income Inequality, measured by the Gini coefficient, where higher values reflect a greater income concentration (Sultanova et al., 2024). The SBTC framework posits that ICT advancements may disproportionately raise wages for high-skilled labor, potentially widening the Gini index (Richmond & Triplett, 2018). Conversely, New Institutional Economics (North, 1990) argues that institutions dictate resource distribution. In transitional economies like Vietnam, institutional evolution focuses on transparency and accountability, following a monotonic inequality-reducing path (Asamoah, 2021). As governance matures, it establishes mechanisms to mitigate wealth gaps and ensure equitable sharing of structural gains.

2.2 ICT and Income Inequality

Information and Communication Technology refers to the digital infrastructure and tools, primarily mobile telecommunications and internet access, that facilitate the flow of information and economic transactions (Asongu & Odhiambo, 2019). The impact of ICT on income distribution is increasingly viewed through the lens of digital inclusion. In a transitional economy like Vietnam, widespread mobile and internet access can act as a foundational driver for socioeconomic change among marginalized groups by lowering transaction costs and providing direct market access (Richmond & Triplett, 2017). This technology helps bridge the wealth gap by empowering those in remote areas with real-time information and financial resources that were previously inaccessible (Asongu & Odhiambo, 2019). Furthermore, the capacity of ICT to foster a fairer distribution of wealth is particularly potent in regions with existing disparities, as it enables an information-flattening effect (Sultanova et al., 2024). Based on the premise that middle-income nations possess the necessary infrastructure to leverage technology for broader participation in the digital economy (Bello et al., 2022), the first hypothesis is proposed.:

Hypothesis 1 (H1): Information and Communication Technology (ICT) has a negative impact on income inequality (GINI) in Vietnam, leading to a reduction in the wealth gap.

2.3 Institutional Quality and Income Inequality

The role of institutional quality, specifically measured by provincial governance (PAPI), is a cornerstone in the effort to achieve equitable growth. Unlike the extractive frameworks seen in some transitional regions, the structural evolution of institutional quality in Vietnam is hypothesized to follow a monotonic inequality-reducing path (Asamoah, 2021). Within this framework, improvements in local governance enhance the transparency of public administration and the accountability of officials, ensuring that politically connected elites do not monopolize economic opportunities. The high-quality institutions, characterized by transparent contract enforcement and secure property rights, ensure equitable access to economic resources across all social strata rather than just the elite (Tran et al., 2021). By lowering entry barriers and facilitating a fairer allocation of public resources, superior governance creates a redistributive effect that narrows the wealth gap. Furthermore, as institutions mature, they enhance the capacity to regulate market forces, ensuring that economic gains are shared more broadly (Batuo et al., 2022). Consequently, in the context of Vietnam's commitment to social equity, the following hypothesis is proposed.

Hypothesis 2 (H2): The structural evolution of institutional quality (PAPI) reduces income inequality (GINI) in Vietnam.

2.4 The Moderating Role of Institutional Quality

Institutional quality is hypothesized to moderate the ICT-inequality nexus through a counterbalancing mechanism. In environments with weak governance, ICT serves as an autonomous alternative, allowing citizens to bypass bureaucratic hurdles and lack of transparency independently (Richmond & Triplett, 2017). However, as PAPI scores rise, formal public services and modern infrastructure become more efficient, while robust accountability mechanisms, such as anti-corruption measures, begin to significantly narrow the wealth gap (Fan et al., 2021). In

such high-governance contexts, the reliance on ICT as a self-managed instrument for economic empowerment becomes less urgent, leading to a diminished marginal impact on income redistribution. This suggests that superior governance acts as a stabilizing force that reduces the equality-generating pressure of new technologies (Adam, 2020). Therefore, the third hypothesis is formulated.

Hypothesis 3 (H3): Institutional quality (PAPI) moderates the relationship between ICT and income inequality by strengthening the inequality-reducing effect of ICT.

3 Methodology

3.1 Data Sources and Variable Definitions

This study employs a balanced panel dataset covering 63 Vietnamese provinces from 2018 to 2022. The data are synthesised from official sources: the General Statistics Office of Vietnam (GSO) for income inequality (GINI), business activity (ACT - number of active enterprises), and Gender (gender ratio); the Ministry of Information and Communications for the provincial ICT Index; and the Ministry of Planning and Investment for realised Foreign Direct Investment (FDI). Institutional quality is proxied by the Provincial Governance and Public Administration Performance Index (PAPI). Additionally, Trade Openness (OT) is calculated as the total value of provincial exports and imports relative to local GDP.

3.2 Estimation methodology applied

The dynamic panel data models allow for modelling the lagged dependence structures and autoregressive dynamics, making them suitable for analysing persistent socioeconomic outcomes such as income inequality. The basic methodological contributions are of Arellano and Bond (1991) and Blundell and Bond (1998). This study employs the two-step System Generalized Method of Moments (System GMM) to address potential endogeneity, reverse causality, unobserved provincial heterogeneity, and the persistence of income inequality. The method is appropriate for the short panel dataset of 63 Vietnamese provinces over 2018–2022 because it combines equations in first differences and levels to improve estimation efficiency. The lagged dependent variables are included to capture dynamic effects, while GINI, ICT, PAPI, and the ICT × PAPI interaction term are treated as endogenous or predetermined variables and instrumented using their lagged values. The estimations are implemented in Stata using the xtabond2 software module, with restricted instrument lags to reduce instrument proliferation and ensure that the number of instruments remains below the number of provincial groups. Model validity is assessed using the Arellano–Bond serial correlation tests and the Hansen test of overidentifying restrictions; particularly, a non-significant AR (2) statistic and a Hansen p-value above 0.05 indicate no second-order serial correlation and support the validity of the instrument set. The dynamic panel model in first differences can be expressed as follows:

$$\Delta Y_{i,t} = \beta\, \Delta X_{i,t} + \delta \Delta Y_{i,t-1} + \Delta \mu_i + \varepsilon_{i,t}, \quad (1)$$

based on the theoretical framework of the research by (Huynh, 2025). This research develops an empirical framework to examine the impact of information and communication technology (ICT) on income inequality across provinces in Vietnam. To estimate the direct effect of ICT on income inequality, Model 1 is proposed as follows:

$$\Delta GINI_{i,t} = \alpha \Delta GINI_{i,t-1} + \beta_1 \Delta ICT_{i,t} + \beta_2 \Delta PAPI_{i,t} + \sum \beta_q \Delta \mathbf{control}_{i,t} + \Delta \mu_{i,} + \varepsilon_{i,t}\,. \quad (2)$$

To further assess whether institutional quality moderates the relationship between ICT development and income inequality, Model 2 introduces an interaction term between ICT and institutional quality as follows:

$$\Delta GINI_{i,t} = \alpha \Delta GINI_{i,t-1} + \beta_1 \Delta ICT_{i,t} + \beta_2 \Delta PAPI_component_{i,t} + \beta_3 \left(\Delta ICT_{i,t}\, x \Delta\, PAPI_{i,t}\right) + \sum \beta_q \Delta \mathbf{control}_{i,t} + \Delta \mu_{i,t} + \varepsilon_{i,t}\,, \quad (3)$$

where $GINI_{i,t}$ is income inequlity, $ICT_{i,t}$ is the information and communication technology index; $PAPI_{i,t}$ represents Institutional quality index components, including $AC_{i,t}$ (Accountability), $Trans_{i,t}$ (Transparent), $CORUP_{i,t}$ (Corruption control), Public-$AD_{i,t}$ (Public administration), $CITI_{i,t}$ (Citizen participation), and Public-$SER_{i,t}$ (Public services).Control variables include: $OT_{i,t}$ is open trade, $FDI_{i,t}$ is Foreign direct investment, $Gender_{i,t}$ is Gender ratio, $ACT_{i,t}$ is Number of active enterprises, $\mu_{i,}$ represents unobserved province-specific effects; and $\varepsilon_{i,t}$ is the idiosyncratic error term.

4 Results and Discussion

This section provides the results of descriptive statistics, baseline regressions, and interaction model estimations.

4.1 Descriptive statistics

The descriptive statistics in Table 1 shows that the variables vary across observations, with most variables having reasonable dispersion. PAPI index and Gender are relatively stable, while GINI, ICT index, and LnFDI show greater variation. Overall, the data are sufficiently distributed and suitable for further analysis.

Variable	Obs	Mean	Std. Dev.	Min	Max
GINI	315	0.361	0.058	0.203	0.525
ICT index	315	0.427	0.138	0.086	0.941
PAPI index	310	42.5	4.191	34.96	48.059
Gender	315	99.97	2.444	94.360	106.5
OT	315	1.297	1.867	0.002	15.514
LnFDI	272	10.787	2.631	0.693	15.936
lnACT	315	8.561	1.048	6.400	12.521

Table 1 Summary statistics

Table 2 presents the correlation matrix among the variables. The results show that most variables are significantly correlated with each other. ICT is positively associated with OT, LnFDI, and LnACT, while PAPI has a weak negative correlation with most variables. In addition, the correlation coefficients are all below 0.8, indicating that multicollinearity is not a serious concern in the model.

	ICT	PAPI	Gender	OT	LnFDI	lnACT
ICT	1					
PAPI	-0.041	1				
	(0.474)					
Gender	-0.274***	0.063	1			
	(0.000)	(0.266)				
OT	0.201***	-0.213***	-0.295***	1		
	(0.000)	(0.000)	(0.000)			
LnFDI	0.399***	-0.108*	-0.264***	0.427***	1	
	(0.000)	(0.079)	(0.000)	(0.000)		
lnACT	0.534***	-0.129**	-0.390***	0.269***	0.605***	1
	(0.000)	(0.028)	(0.000)	(0.000)	(0.000)	

*** at the 1% level, ** at the 5% level, and * at the 10% level

Table 2 Correlation matrix

4.2 Result from regression

The results in Table 3 indicate that the impact of ICT on income inequality depends significantly on institutional quality. First, the ICT variable has a negative coefficient and is statistically significant in most models (1), (3), (4), (5) and (6), indicating that the development of information and communication technology contributes to reducing income inequality. However, in model (2), ICT has a positive coefficient and is significant, which may reflect the moderating role of institutional quality in the relationship between ICT and inequality. In models (7) and (8), ICT is no longer statistically significant, implying that when more detailed components of PAPI are included in the model, the individual impact of ICT may be diminished.

For the PAPI, Model (1) shows a negative and statistically significant coefficient, indicating that better institutional quality reduces income inequality. However, in the interaction model, the interpretation of PAPI and ICT changes because their effects are conditional on each other. Although the coefficient of PAPI becomes positive in Model

(2), this coefficient should not be interpreted as the overall direct effect of institutional quality. More importantly, the negative and significant ICT × PAPI interaction suggests that the effect of ICT on income inequality depends on the level of public governance quality. Specifically, better governance strengthens the inequality-reducing effect of ICT development, implying that provinces with higher institutional quality are better able to use ICT development as a tool to reduce income inequality.

When PAPI is decomposed into specific components (model 3–8), factors such as accountability (AC), transparency (TRAS), and corruption control (CORUP) all have negative and statistically significant impacts, indicating that these aspects play an important role in reducing income inequality. Meanwhile, components such as public administration, citizen participation, and public services lacked clear statistical significance, implying that their impact was not strong enough or was dependent on other factors. Regarding the control variables, the Gender variable consistently had positive coefficients and was statistically significant in most models, suggesting that gender may contribute to increased income inequality. Conversely, trade openness (OT) and foreign direct investment (LnFDI) primarily had negative impacts, implying that economic integration and attracting foreign investment contributed to reducing inequality. The lnACT variable was not statistically significant in most models.

Variable	(1) GINI	(2) GINI	(3) GINI	(4) GINI	(5) GINI	(6) GINI	(7) GINI	(8) GINI
ICT	-0.103***	3.773**	-0.142***	-0.253**	-0.323***	-0.313**	-0.142	-0.086
PAPI	-0.003**	0.044**						
Gender	0.004***	0.003	0.003***	0.007***	0.007***	0.007**	0.003***	0.002***
OT	-0.004***	-0.007**	-0.004***	-0.007*	-0.012**	-0.007	-0.002	-0.001
LnFDI	-0.005**	0.000	-0.001**	-0.002	-0.001	-0.001	-0.001*	-0.001*
lnACT	-0.004	-0.019**	-0.001	-0.000	0.008	0.007	0.000	-0.002
ICT x PAPI		-0.084**						
AC			-0.018***					
TRAS				-0.043**				
Corup					-0.050**			
Public-AD						-0.0443		
CITI							-0.004	
Public-SER								-0.003
N	197	197	215	215	215	215	212	215
AR1	0.008	0.075	0.002	0.025	0.014	0.004	0.001	0.002
AR2	0.826	0.614	0.578	0.266	0.172	0.363	0.412	0.370
Hansen	0.243	0.864	0.081	0.492	0.748	0.262	0.187	0.116

**** at the 1% level, ** at the 5% level, and * at the 10% level*
Components PAPI: AC is Accountability, Trans is transparency, Corup is corruption control, Public-AD is public administration, CITI is citizen participation, Public-SER is public services

Table 3 Estimation Results for Income Inequality Model

5 Conclusion

This study examines the roles of Information and Communication Technology (ICT) and institutional quality in shaping income inequality across 63 Vietnamese provinces over the period from 2018 to 2022. Based on a balanced panel dataset and employing the System Generalised Generalised Method of Moments (GMM), the main contribution is the empirical examination of ICT, PAPI-based institutional quality, and their interaction in explaining provincial income inequality in Vietnam.

The results strongly support Hypothesis 1 (H1), indicating that ICT development has a statistically significant negative effect on income inequality. This finding confirms that digitalisation contributes to narrowing the income gap by enhancing access to information and expanding economic opportunities for disadvantaged groups.

Additionally, the results are consistent with Hypothesis 2 (H2), as the institutional quality, measured by the Provincial Governance and Public Administration Performance Index (PAPI). It is found significantly reduce income inequality. This highlights the critical role of effective governance, transparency, corruption, and accountability in ensuring a more equitable distribution of economic resources. Furthermore, the interaction analysis provides strong evidence in support of Hypothesis 3 (H3). The negative and significant coefficient of the interaction term between ICT and institutional quality suggests that better governance strengthens the inequality-reducing effect of ICT. In other words, institutional quality acts is a complementary factor that enhances the inclusiveness of digital transformation. In addition, the control variables provide further insights into the determinants of income inequality. Gender is found to have a positive and statistically significant effect across most specifications, suggesting that gender disparities may contribute to widening income gaps. In contrast, trade openness (OT) generally exhibits a negative effect, indicating its role in reducing inequality through enhanced economic integration. Similarly, foreign direct investment (FDI) tends to lower income inequality, although its significance varies across models, implying that its distributive impact may depend on local economic conditions.

Overall, the study demonstrates that both ICT development and institutional improvement are mutually reinforcing drivers of inclusive growth in Vietnam. These findings imply that policies aimed at reducing income inequality should not only focus on expanding digital infrastructure but also on improving governance quality to maximize the distributive benefits of technological progress. Despite these contributions, several limitations remain open, including the relatively short time frame and constraints in provincial-level data. Future research could extend the study period, incorporate additional institutional indicators, and explore nonlinear dynamics to provide deeper insights into the ICT-inequality relationship in developing economies.

References

Adam I O (2020) Examining E-Government development effects on corruption in Africa: The mediating effects of ICT development and institutional quality. Technol. Soc. 61: 101245. https://doi.org/10.1016/j.techsoc.2020.101245

Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The Review of Economic Studies 58(2): 277–297

Asamoah L A (2021) Institutional Quality and Income Inequality in Developing Countries: A Dynamic Panel Threshold Analysis. Prog. Dev. Stud. 21: 123–143. https://doi.org/10.1177/14649934211016715

Asongu S A, Odhiambo N M (2019) How enhancing information and communication technology has affected inequality in Africa for sustainable development: An empirical investigation. Sustain. Dev. 27: 647–656. https://doi.org/10.1002/sd.1929

Batuo M E, Kararach G, Malki I (2022) The dynamics of income inequality in Africa: An empirical investigation on the role of macroeconomic and institutional forces. World Dev. 157: 105939. https://doi.org/10.1016/j.worlddev.2022.105939

Bello A A, Renai J, Hassan A, Akadiri S S, Itari A R (2022) Synergy effects of ICT diffusion and foreign direct investment on inclusive growth in Sub-Saharan Africa. Environ. Sci. Pollut. Res. 30: 9428–9444. https://doi.org/10.1007/s11356-022-22689-3

Blundell R, Bond S (1998). Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics 87(1): 115–143

Fan Q, Kuper P, Choi Y H, Choi S (2021a) Does ICT development curb firms' perceived corruption pressure? The contingent impact of institutional qualities and competitive conditions. J. Bus. Res. 135: 496–507. https://doi.org/10.1016/j.jbusres.2021.06.062

Ho T L, Ngoc L H, Ho T H (2025) Digital inclusion or exclusion? Exploring the moderating role of governance in digitalization's impact on income inequality in developing countries. Res. Glob. 10: 100283. https://doi.org/10.1016/j.resglo.2025.100283

Huynh Quoc V (2026) Institutional quality and infrastructure as catalyst to the effects of ICT on provincial total factor productivity. Compet. Rev. Int. Bus. J. 36: 748–768. https://doi.org/10.1108/CR-03-2025-0105

Kim Thi Tran O, Dinh Le H, Hong Viet Nguyen A (2021) Role of institutional quality in economic development: A case study of Asian countries. Probl. Perspect. Manag. 19: 357–369. https://doi.org/10.21511/ppm.19(2).2021.29

Lee M-H, Yu C, Kao C-H (2026) ICT development and income Inequality: The moderating role of the entrepreneurial environment. Telecommun. Policy 50: 103084. https://doi.org/10.1016/j.telpol.2025.103084

North D C (1990) Institutions, institutional change and economic performance. Cambridge university press

Richmond K, Triplett R E (2018) ICT and income inequality: a cross-national perspective. Int. Rev. Appl. Econ. 32: 195–214. https://doi.org/10.1080/02692171.2017.1338677

Sultanova G, Djuraeva R, Usmanova A (2024) The Impact of the ICT and Export Diversification on Income Inequality: Panel Quantile Regression Analysis, in: Koucheryavy, Y., Aziz, A. (Eds.), Internet of Things, Smart Spaces, and Next Generation Networks and Systems, Lecture Notes in Computer Science. Springer Nature Switzerland, Cham, pp 328–338. https://doi.org/10.1007/978-3-031-60997-8_30

Sy Tinh N (2025) Digital Economy Development in Vietnam: Challenges and Policy Directions. Int. J. Adv. Multidiscip. Res. Stud. 5: 1267–1271. https://doi.org/10.62225/2583049X.2025.5.5.5105

Tewathia N, Kamath A, Ilavarasan P V (2020) Social inequalities, fundamental inequities, and recurring of the digital divide: Insights from India. Technol. Soc. 61: 101251. https://doi.org/10.1016/j.techsoc.2020.101251

Yin Z H, Choi C H (2023) Does digitalization contribute to lesser income inequality? Evidence from G20 countries. Inf. Technol. Dev. 29: 61–82. https://doi.org/10.1080/02681102.2022.2123443

Deep Reinforcement Learning for Perpetual Futures Market Making

Martin Cekal[1]

Abstract. This paper studies market making in BTC and ETH perpetual futures with a simple event-driven reinforcement-learning setup. PPO chooses quote skew and quote width, while the environment tracks execution gains and inventory losses separately. A temporary reduction of the adverse-fill penalty is used as an intended stabilizing device during early training. Across simulated and historical data, PPO quotes more selectively, carries smaller inventory, and usually ends with better portfolio value than a zero-action baseline. BTC is tighter than ETH, so the same control logic requires market-specific calibration, yet the qualitative policy behavior remains consistent across both markets. The results are therefore encouraging in a narrow sense: a compact and interpretable RL policy can improve inventory control and over-all portfolio outcomes in two different crypto perpetual environments, even though robust profitability claims remain out of scope.

Keywords: Deep Reinforcement Learning, Market Making, Perpetual Futures, Proximal Policy Optimization

JEL Classification: C63, G11, G12

1 Introduction

Perpetual futures market making balances spread capture, adverse selection, and inventory risk on an event-driven stream of book and trade updates. Classical inventory-based dealer models (Ho, Stoll, 1981, Avellaneda, Stoikov, 2008) and their algorithmic-trading extensions (Cartea, Jaimungal, Penalva, 2015) frame this trade-off, while reinforcement learning offers a data-driven alternative (Spooner et al, 2018). In this paper PPO (Schulman et al, 2017) adjusts quote skew and quote width in BTC and ETH perpetuals, while the environment keeps trade edge and inventory revaluation separate. The main empirical pattern is simple: compared with both a zero-action policy and an Avellaneda–Stoikov benchmark, PPO behaves mainly as a conservative inventory-control rule, quoting wider and trading less often, but usually ending with smaller exposure and better terminal portfolio value.

Five ingredients matter most. First, the control is written compactly through reservation adjustment and reservation spread in log-price space. Second, step-wise portfolio value decomposes into trade edge and inventory mark-to-market fluctuation, while maker trade edge splits into immediate quote edge and fill-time fluctuation. Third, scheduled damping of adverse fill fluctuation through ϕ_{fluc} is used as a motivated early-training device rather than as an experimentally isolated variance-reduction result. Fourth, the same accounting decomposition is examined on both simulated and historical perpetual data. Fifth, the BTC–ETH comparison shows why the same market-making engine has to be interpreted with market-specific spread and order-flow calibration.

2 Perpetual Market-Making Environment

At decision step t, the raw observation is

$$o_t = (q_t, ols_t, rv_t), \tag{1}$$

where q_t is signed inventory, $ols_t = \log(a_{mkt,t}) - \log(b_{mkt,t})$ is the observed log spread, and rv_t is a fast annualized realized-volatility estimate. Writing $o_t = (o_{1,t}, o_{2,t}, o_{3,t})$, each component is clipped and normalized before entering PPO:

$$\tilde{o}_{i,t} = \frac{2\,\mathrm{clip}(o_{i,t}, \ell_i, u_i) - (u_i + \ell_i)}{u_i - \ell_i}. \tag{2}$$

The notation distinguishes raw observation o_t from normalized policy input $\tilde{o}_t$. The action is $a_t = (ra_t, rs_t)$, where ra_t shifts the quote center and rs_t sets quote width. PPO outputs normalized action values in $[-1, 1]$; they are then mapped to admissible quote controls before prices are submitted.

[1] Department of Banking and Insurance, University of Economics, Nám. W. Churchilla 1938/4, 130 67 Praha 3, martin.cekal@vse.cz

Symbol	Meaning
o_t	Raw observation vector.
$\tilde{o}_t$	Normalized observation given to PPO.
q_t	Signed inventory.
ols_t	Log observed spread.
rv_t	Fast realized-volatility signal.
ra_t	Reservation adjustment.
rs_t	Reservation spread.
rs_{min}, rs_{max}	Lower and upper admissible reservation-spread bounds.
p_t	Execution price.
$t', \mathcal{M}_t, \mathcal{T}_t$	Execution event inside step t; maker and taker event sets in that step.
t^-	Time at the beginning of step t.
s_t, d_t	Executed size and buy/sell direction; $s_t d_t$ is signed executed size.
$m_{quote,t}$	Mid price at quote placement time.
$m_{exec,t}$	Mid price at fill-processing (execution) time.
PV_t	Portfolio value.
ψ_{quote}	Immediate maker quote edge.
ψ_{fluc}, ϕ_{fluc}	Favorable/adverse fill-time fluctuation.
ϕ_{take}	Taker loss relative to the mid.

Table 1 Main notation

For perpetuals, quotes are placed as

$$\log a_{quote,t} = \log m_t + ra_t + \frac{rs_t}{2}, \qquad \log b_{quote,t} = \log m_t + ra_t - \frac{rs_t}{2}. \tag{3}$$

In implementation the spread action is centered on the observed market spread, so zero spread action targets the current market spread. The constants rs_{min} and rs_{max} are the lower and upper admissible values of the reservation spread in log-price units; in the reported perpetual experiments they are 0 and 5×10^{-4}, respectively. Writing the normalized spread action as $u_t^{(rs)} \in [-1, 1]$, the mapping is

$$rs_t = \begin{cases} rs_{min} + (u_t^{(rs)} + 1)(s_{mkt,t} - rs_{min}), & u_t^{(rs)} \le 0, \\ s_{mkt,t} + u_t^{(rs)}(rs_{max} - s_{mkt,t}), & u_t^{(rs)} > 0. \end{cases} \tag{4}$$

Hence $u_t^{(rs)} = 0 \Rightarrow rs_t = s_{mkt,t}$ because the normalized action value 0 is the pivot of the piecewise map, while $rs_{min} = 0$ is the lower bound of the denormalized reservation spread. Final prices are tick-aligned to the exchange tick size.

2.1 Fill Conditions and Model Scope

The model considers only the best ask and best bid. Maker fills are triggered by quote touches or trade prints crossing the standing quote. Taker fills cross the observed best quote. If the strategy crosses the spread, the taker term is

$$\phi_{t'}^{take} = s_{t'} d_{t'} (m_{exec,t'} - p_{take,t'}), \tag{5}$$

where p_{take} is the current best ask for taker buys and the current best bid for taker sells. Maker fees are zero and taker fees are five basis points, so the execution price already includes fee adjustment. No additional slippage term is assumed beyond top-of-book crossing.

This is a top-of-book execution model, not a full exchange-matching model. It preserves what is needed for the accounting problem: quote placement, possible execution against the observed market, and the later mark-to-market effect of resulting inventory. Queue position, hidden liquidity, latency races, and market impact are not modeled, so absolute PnL magnitudes are model-based diagnostics rather than deployment-ready execution estimates.

3 PnL Decomposition and Reward

The central accounting point is that maker execution quality and inventory mark-to-market movement are distinct objects. Let t' denote an individual execution event inside decision step t, and let $\mathcal{M}_t$ and $\mathcal{T}_t$ collect the maker and taker executions in that step. Let $m_{quote,t'}$ denote the prevailing mid when a quote is placed, and let $m_{exec,t'}$ denote the mid when that execution is processed.

For maker fills, the immediate quote edge is

$$\psi_{t'}^{quote} = s_{t'} d_{t'} (m_{quote,t'} - p_{t'}), \tag{6}$$

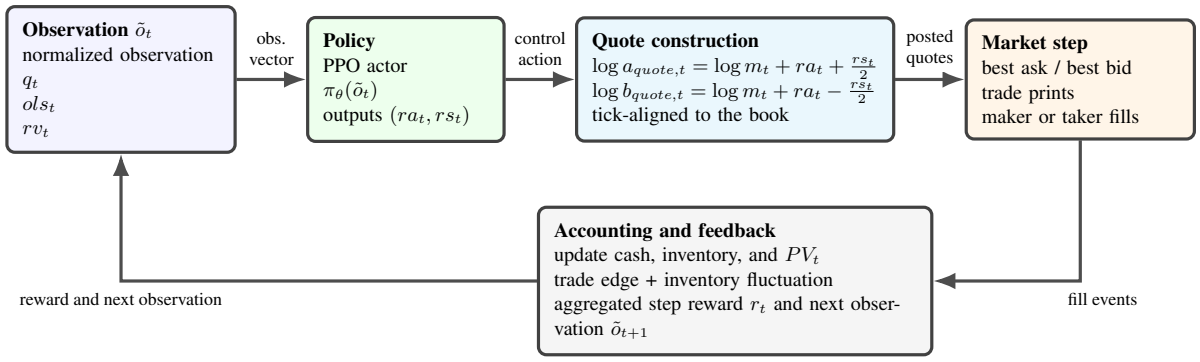


Figure 1 Perpetual control loop. The policy receives the normalized observation, outputs (ra_t, rs_t), quotes both sides, and then receives fills, inventory fluctuation, and the aggregated step reward.

and the fill-time fluctuation term is

$$fluc_{t'} = s_{t'} d_{t'} (m_{exec,t'} - m_{quote,t'}). \tag{7}$$

Its positive and negative parts are

$$\psi_{t'}^{fluc} = \max(fluc_{t'}, 0), \qquad \phi_{t'}^{fluc} = \min(fluc_{t'}, 0). \tag{8}$$

Aggregating executions inside one decision step,

$$\psi_t^{quote} = \sum_{t' \in \mathcal{M}_t} \psi_{t'}^{quote}, \quad \psi_t^{fluc} = \sum_{t' \in \mathcal{M}_t} \psi_{t'}^{fluc}, \quad \phi_t^{fluc} = \sum_{t' \in \mathcal{M}_t} \phi_{t'}^{fluc}, \tag{9}$$

so the total maker contribution at step t is

$$\psi_t^{quote} + \psi_t^{fluc} + \phi_t^{fluc} = \sum_{t' \in \mathcal{M}_t} s_{t'} d_{t'} (m_{exec,t'} - p_{t'}). \tag{10}$$

Inventory revaluation is tracked separately. Here t^- means the time at the beginning of step t, so q_{t^-} and m_{t^-} are the inventory and mid price carried into the step:

$$step_fluctuations_t = q_{t^-}(m_t - m_{t^-}). \tag{11}$$

Hence

$$\Delta PV_t = trade_edge_net_t + step_fluctuations_t, \tag{12}$$

where $trade_edge_net_t = \psi_t^{quote} + \psi_t^{fluc} + \phi_t^{fluc} + \phi_t^{take}$ and $\phi_t^{take} = \sum_{t' \in \mathcal{T}_t} \phi_{t'}^{take}$. This split is central for reward design and training schedules, because the components can be weighted gradually to support stable learning.

4 Reinforcement Learning Formulation

The environment is an event-driven MDP (Sutton, Barto, 2018), solved with PPO (Schulman et al, 2017). The agent input is the normalized observation vector $\tilde{o}_t$ and the action is the continuous pair (ra_t, rs_t). In the reported setup one action is selected every 1000 ms and then held fixed while all order-book and trade events inside that interval are processed sequentially. The transition kernel is therefore event-driven inside each decision interval. Reward is accumulated over the interval by differencing cumulative counters between interval start and interval end. The shaped reward uses interpretable components with scheduled weights; in particular, the adverse fill term ϕ_{fluc} is damped early in training and ramped up later as a design choice intended to avoid making early updates dominated by adverse fill-time noise.

The scalar reward is a weighted sum of components,

$$r_t = \sum_k w_k R_{k,t}, \tag{13}$$

with the linear training schedule

$$w_k(u) = w_k^{start} + u\big(w_k^{end} - w_k^{start}\big), \qquad u \in [0, 1]. \tag{14}$$

This matters for interpretation because the schedule changes exploration pressure early in training without changing the evaluation accounting itself. No ablation experiment is reported for this schedule, so the schedule is interpreted as a motivated training design rather than as evidence of a measured variance reduction.

Term	Start	End
ψ_t^{quote}	1.0	1.0
ψ_t^{fluc}	1.0	1.0
ϕ_t^{fluc}	0.1	0.6
ϕ_t^{take}	1.0	1.0
Risk aversion	10.0	0.0
Delta aversion	0.1	0.1
Spread aversion	0.0002	0.0

Table 2 Key reward weights used in training

Market	Snapshots [M]	Tick [USD]	q50 [bps]	q90 [bps]	1-tick states
BTC	44.38	0.50	0.31	2.08	45.02%
ETH	61.63	0.05	0.73	2.34	35.88%

Table 3 Half-year spread summary. Columns q50 and q90 are spread quantiles in basis points relative to repre-sentative prices of 48,147 USD for BTC and 3,426 USD for ETH

5 Data and Experimental Design

The historical data consist of tick-by-tick trades and tick-by-tick top-of-book market quotes for BTC and ETH perpetual futures from 2021-10-01 to 2022-03-31. The sample contains 44.38 million BTC top-of-book snapshots and 61.63 million ETH snapshots. Raw tick counts are not directly comparable because one BTC tick is 0.50 USD while one ETH tick is 0.05 USD. Relative to representative prices of about 48,147 USD for BTC and 3,426 USD for ETH, the median spread is about 0.31 bps for BTC and 0.73 bps for ETH. At the 90th percentile the relative spreads are closer, about 2.08 bps and 2.34 bps. Thus BTC is still narrower in relative terms, but less dramatically than raw tick counts suggest.

The half-year period is first used to calibrate model parameters for the simulation engine. Such engine is then used for preliminary PPO model training. In the second phase, historical data is then used to retrain PPO on historical data itself. There is no separate validation period at this point. The reported PPO policy is the final saved model from the retraining historical run - no intermediate checkpoint is saved. The reported results are therefore in-sample. It should be read as a stress test of inventory control rather than an out-of-sample profitability backtest.

The simulator has three parts. The mid price follows a Heston-like process, the spread follows a six-state Markov chain, and buy and sell trade executions are generated as Bernoulli events with market-specific logistic probabilities driven by the standardized relative spot move. The Heston-style parameters $(\kappa, \theta, \xi, \rho, v_0)$ are fitted separately for BTC and ETH from five-minute historical perpetual-mid returns. A running annualized realized-variance estimate is used as the observable proxy for the latent variance path; v_0 is the filtered variance at the simulator start. After fitting, these parameters are kept fixed in the simulator. Spreads are calibrated from 250 ms last-quote samples over the whole dataset. Raw spread ticks are grouped into six states: $\{1, 4, 8, 11, 16, 29\}$ for BTC and $\{1, 4, 7, 11, 16, 31\}$ for ETH. Transition probabilities are then estimated from adjacent spread states. This uses 11.53 million BTC calibration snapshots and 16.35 million ETH snapshots. The tightest spread state is more persistent for BTC, with persistence 0.756 versus 0.670 for ETH, which again shows that ETH is usually wider.

The simulator is used mainly as a warm-up device. It gives PPO plausible BTC or ETH-specific price, spread, and order-flow conditions before historical fine-tuning. The main evidence in the paper comes from the fine-tuned policy on historical dataset and from the inventory-control decomposition, not from the simulator fit itself.

Data cleaning is based on a simple assumption that trades executions are more reliable than orderbook quotes. First, positive prices and amounts are ensured, then trades at the same millisecond are aggregated by side and price; if several prices occur in the same side and millisecond, the low and high prices are split across adjacent milliseconds to preserve their ordering. Fourth, trades are joined with the cleaned best bid and ask, and when a trade price is misaligned with the recorded orderbook, the corresponding ask or bid is adjusted to the buy or sell execution price. Fifth, missing quote sides are forward-filled in order to ensure non-missing values on both sides.

6 Results

Table 5 shows a stable qualitative pattern on the full half-year historical screen in both markets. PPO has the best average terminal portfolio value among the three policies for both BTC and ETH. On BTC, PPO improves mean terminal PV from -65.55 (zero-action) and -20.60 (Avellaneda–Stoikov) to -11.52 USD. On ETH, PPO

Market	Spread states [ticks]	Tight persistence	$(\kappa, \theta, \xi, \rho, v_0)$	s_0	Per-side print prob.	Tick [USD]
BTC	$1, 4, 8, 11, 16, 29$	0.756	$(5.4, 3.8, 5.3, 0.1, 0.24)$	48147	0.6515	0.50
ETH	$1, 4, 7, 11, 16, 31$	0.670	$(6.5, 2.6, 4.1, 0.4, 0.3)$	3426	0.3832	0.05

Setting	Value	Setting	Value
Episode length	300 steps	Decision interval	1000 ms
PPO rollout length	60	Episodes	1000
Discount factor γ	0.999	GAE λ	0.95
Clip range	0.1	Action bounds	$ra_t \in [-5, 5] \times 10^{-5}, rs_t \in [0, 5] \times 10^{-4}$
Actor network	$128, 128$	Critic network	$256, 256$
Activation	Tanh	Optimizer	Adam
Learning rate	3×10^{-4}	PPO update epochs	3
Minibatch size	full rollout buffer	PPO seed	52
Training phases	simulated warm-up, historical fine-tuning	Reported PPO policy	final saved model
Validation sample	none	Historical screen	in-sample

Table 4 Simulator calibration, PPO settings, and how the reported policy is chosen. The reported PPO policy is the final saved model after the pre-specified training run; the historical screen is not used to select an intermediate checkpoint

Market	Policy	Intervals (ok/total)	Mean terminal PV [USD]	Std terminal PV [USD]
BTC	PPO	182/182	-11.52	33.46
BTC	Zero-action	182/182	-65.55	175.45
BTC	Avellaneda–Stoikov	182/182	-20.60	54.41
ETH	PPO	182/182	-8.16	11.49
ETH	Zero-action	182/182	-59.15	133.42
ETH	Avellaneda–Stoikov	182/182	-13.89	22.04

Table 5 Full half-year historical multi-day evaluation from 2021-10-01 to 2022-03-31 using 182 matched evaluation intervals per market (all intervals valid). Numbers are from the historical multi-day summaries for BTC and ETH. The screen is in-sample with respect to calibration and is not used for checkpoint selection

improves mean terminal PV from -59.15 and -13.89 to -8.16 USD.

The average rankings in Table 5 should not be interpreted as strict dominance on every day or as statistically robust evidence of profitability. The large standard deviations, especially for zero-action, indicate occasional extreme inventory-driven downside intervals. This is precisely why the decomposition in (12) is useful: it distinguishes local execution quality from delayed exposure costs.

The half-year historical plots in Figure 2 point in the same direction. Zero-action has visibly larger negative tail events in both markets, while PPO remains much closer to zero across most evaluation intervals. This is an illustrative inventory-control diagnostic for (12): the main policy difference is not only immediate execution edge but also the mark-to-market cost of inventory carried through adverse intervals.

7 Discussion

The simulator and historical-evaluation limitations should be interpreted strictly. Queue position, hidden liquidity, latency races, and market impact are not modeled, so absolute PnL levels are model-based diagnostics rather than deployable execution estimates. The historical screen is also in-sample with respect to calibration and has no separate validation-based checkpoint selection. Any apparent edge may depend materially on omitted matching mechanics, on the calibrated simulator priors, or on the particular historical intervals used for illustration. The article therefore supports a narrower claim: under a compact observation design and a transparent accounting split, PPO can learn an interpretable inventory-sensitive quoting rule that transfers qualitatively across BTC and ETH once the simulator is recalibrated by market. It does not establish statistically robust profitability.

8 Conclusion

The paper can be summarized in one sentence: PPO improves the perpetual market-making policy mainly by controlling inventory rather than by maximizing fill count or immediate spread capture. That pattern appears in both BTC and ETH and in both simulated and historical pipelines, although the numerical scale depends strongly on market-specific spread and order-flow calibration. The historical screen should be read as illustrative evidence of inventory control under an in-sample calibrated setup, not as statistically robust evidence of profitable execution.

The main methodological value is interpretability. The compact observation, explicit quote controls, and decomposition into trade edge and inventory fluctuation make it clear why the policy wins or loses. The main limitation

Aspect	BTC	ETH	Implication for control
Central spread	Lower relative median spread, about 0.31 bps, with 45.02% one-tick states.	Higher relative median spread, about 0.73 bps, with 35.88% one-tick states.	The same widening action changes trading intensity differently across markets.
Spread dynamics	Tight states are more persistent in the calibrated spread chain.	Spread states are more diffuse and less concentrated in the tightest regime.	ETH requires a more cautious calibration of quoting aggressiveness.
Observed print arrivals	Higher one-second per-side simulator rate.	Lower one-second per-side simulator rate.	Passive narrow quoting is less uniformly rewarded in ETH.
Historical baseline behavior	Zero-action accumulates meaningful but still moderate inventory.	Zero-action accumulates very large inventory on the historical screen.	Inventory-sensitive skew is especially important in ETH.
PPO effect	Lowers inventory and improves terminal portfolio value.	Same qualitative effect, with stronger historical inventory-control contrast.	The control logic transfers even though numerical scales differ.

Table 6 Qualitative BTC–ETH comparison and control implications

Figure 2 Full half-year historical multi-day terminal portfolio value by evaluation interval (2021-10-01 to 2022- 03-31). Top: BTC. Bottom: ETH. Each market contains 182 matched intervals; PPO, zero-action, and Avellaneda–Stoikov are evaluated on the same realized intervals

is equally clear: this is still a top-of-book simulator. Broader historical testing, richer exchange mechanics, and finer modeling of wide-spread states are natural next steps.

References

Avellaneda M, Stoikov S (2008) High-frequency trading in a limit order book. Quantitative Finance 8(3):217–224

Cartea Á, Jaimungal S, Penalva J (2015) Algorithmic and High-Frequency Trading. Cambridge University Press, Cambridge

Ho T, Stoll H R (1981) Optimal dealer pricing under transactions and return uncertainty. Journal of Financial Economics 9(1):47–73

Schulman J, Wolski F, Dhariwal P, et al (2017) Proximal policy optimization algorithms. arXiv:1707.06347

Spooner T, Fearnley J, Savani R, et al (2018) Market making via reinforcement learning. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), pp 434–442

Sutton R S, Barto A G (2018) Reinforcement Learning: An Introduction, 2nd ed. MIT Press, Cambridge, MA

Investment Timing and Strategic Interaction in Cross-Border M&A: A Real Options Perspective

Qiyuan Dai[1]

Abstract. Cross-border mergers and acquisitions (M&A) are exposed to firm-value fluctuations, exchange-rate risk, and uncertainty over technological synergies, making acquisition valuation and timing inherently dynamic. This paper develops an integrated real-options framework that jointly analyzes optimal investment timing and strategic interaction in cross-border M&A. The acquisition opportunity is modeled as a pre-commitment option to defer and solved using the Least-Squares Monte Carlo (LSMC) method on a discrete monthly grid. A bargaining mechanism is further introduced to capture surplus sharing between the acquiring and target firms. Using the Siemens–Altair acquisition as an illustrative high-technology cross-border case, the baseline acquisition option value is estimated at approximately EUR 1.22 billion, with an average exercise time of 1.5416 years and an exercise probability of 39.53%. Sensitivity analysis shows that firm-value dynamics, volatility, and technological synergy are the principal drivers of acquisition value, whereas exchange-rate movements and integration costs have smaller effects. Bargaining reduces the value retained by the acquirer and the likelihood of commitment while having only a limited effect on timing. Robustness tests confirm that the main conclusions are stable across alternative exercise frequencies, regression specifications, and technological-transition assumptions.
The paper contributes to the M&A literature by integrating real-options timing, exchange-rate exposure, technological uncertainty, and strategic bargaining within a unified framework.

Keywords: real options, mergers and acquisitions, game theory

JEL Classification: G32, G34

1 Introduction

Cross-border mergers and acquisitions (M&A) provide firms with an important means of accessing strategic assets and technological capabilities, particularly in high-technology industries. However, such transactions are exposed to multiple sources of uncertainty, including fluctuations in target firm value, exchange-rate movements, and uncertainty regarding technological synergies. These factors make acquisition decisions inherently dynamic, requiring firms to determine not only whether to invest but also when to make an irreversible acquisition commitment.

Real options theory provides a natural framework for analyzing such timing decisions because it emphasizes the value of waiting under uncertainty and irreversibility (Dixit and Pindyck, 1994). However, most real-options models adopt a single-agent perspective and treat strategic interaction as exogenous. By contrast, game-theoretic approaches emphasize bargaining and competition but are often abstract from dynamic investment timing under uncertainty. Consequently, optimal timing and strategic interaction are often analyzed separately, although both are central to cross-border M&A decisions.

This paper develops an integrated framework to analyze optimal acquisition timing and strategic interaction under multiple sources of uncertainty. It addresses three questions: (RQ1) how firm-value, exchange-rate, and technological uncertainty affect acquisition value and exercise behavior; (RQ2) what determines optimal acquisition timing; (RQ3) how bargaining power influences acquisition value and the likelihood of investment.
The acquisition opportunity is modeled as a pre-commitment option to defer and solved using the Least-Squares Monte Carlo (LSMC) method on a discrete monthly grid. The model incorporates firm-value dynamics, exchange-rate movements, and technological uncertainty and is extended with a bargaining mechanism that determines how technological surplus is shared between the acquiring and target firms. Using the Siemens–Altair transaction as an illustrative high-technology cross-border case, the results show substantial option value and delayed, state-dependent acquisition commitment. Firm-value dynamics, volatility, and technological synergy are the main drivers of acquisition value, whereas exchange-rate effects are comparatively smaller. Bargaining reduces the value retained by the acquirer and the likelihood of commitment but has only a limited effect on optimal timing. The paper therefore contributes by

[1] VSB - Technical University of Ostrava, Department of Finance,17. listopadu 2172/15
708 00 Ostrava-Poruba, Czech Republic, Qiyuan.dai@vsb.cz

integrating multidimensional uncertainty, optimal stopping, and strategic bargaining within a unified framework for cross-border M&A.

2 Literature Review

Cross-border mergers and acquisitions (M&A) are exposed to multiple sources of uncertainty that can affect both acquisition valuation and investment timing. Policy and economic uncertainty may alter investment incentives, regulatory conditions, and deal completion, while exchange-rate movements affect acquisition costs and expected returns in cross-border transactions (Paudyal et al., 2021; Bonaime et al., 2018; Si et al., 2024; Li and Zhang, 2024). In high-technology M&A, technological uncertainty is particularly important because acquisition value depends on technological compatibility, integration success, and the realization of expected synergies (Li et al., 2021; Wei et al., 2025). Existing studies, however, often examine these sources of uncertainty separately, with more limited attention to their joint effects on acquisition timing and strategic decision-making.

Real options theory provides a natural framework for analyzing investment decisions under uncertainty and irreversibility. McDonald and Siegel (1986) show that firms may optimally delay irreversible investment until sufficiently favorable conditions emerge, highlighting the value of waiting. Dixit and Pindyck (1994) further formalize this insight within a stochastic investment framework in which optimal investment timing depends on the trade-off between immediate exercise and the continuation value of preserving flexibility. This perspective is particularly relevant to cross-border M&A because target value, exchange rates, and technological outcomes can evolve before the acquiring firm makes an irreversible commitment.

When several sources of uncertainty interact, closed-form solutions are generally difficult to obtain, making numerical methods necessary. The Least-Squares Monte Carlo (LSMC) approach provides a flexible solution by combining simulated state paths with backward estimation of continuation values. Previous research emphasizes that continuation-value approximation and regression specification are important determinants of numerical accuracy in simulation-based real-options models (Smith, 2005; Nadarajah et al., 2017). LSMC and related simulation-based approaches have also been applied to investment problems characterized by market and technological uncertainty (Ahmadi and Bratvold, 2023; Najafi and Aghaei, 2024). Its ability to accommodate several interacting state variables makes LSMC particularly suitable for the multidimensional acquisition-timing problem considered in this paper.

Real options theory has also been applied directly to mergers and acquisitions, where acquisition opportunities can be interpreted as options to invest. Firms therefore face a trade-off between delaying acquisition to obtain additional information and committing earlier to capture strategic benefits (Dunis and Klein, 2005). Existing studies use real-options reasoning to analyze takeover premiums, uncertain synergies, and strategic acquisition decisions (Barbopoulos et al., 2019; Čirjevskis, 2020; Ebina et al., 2022). These studies demonstrate that managerial flexibility can materially affect acquisition value. However, much of the real-options literature retains a predominantly single-agent perspective and gives more limited attention to the strategic interaction between acquirers and target firms.

M&A decisions also involve strategic interaction because transaction prices and the distribution of merger surplus depend on negotiation between participating firms. Game-theoretic studies therefore examine merger decisions, bargaining, acquisition pricing, and contractual design. Early work models merger decisions as strategic games in which firms anticipate the responses and incentives of other participants (van den Honert and Stewart, 1992). Subsequent studies integrate uncertainty with acquisition pricing and contractual mechanisms. Yu and Xu (2011), for example, combine real-options and game-theoretic reasoning to analyze acquisition pricing under uncertainty, while Lukas et al. (2012) examine how contractual arrangements reflect strategic interaction under uncertain acquisition outcomes.
Despite these developments, the real-options and game-theoretic approaches remain only partially integrated. Real-options models primarily emphasize uncertainty, managerial flexibility, and optimal investment timing, whereas game-theoretic models focus more strongly on strategic interaction and surplus allocation. Moreover, relatively limited attention has been paid to how firm-value uncertainty, exchange-rate exposure, and technological uncertainty jointly affect acquisition timing and the distribution of acquisition gains. This paper addresses this gap by modeling cross-border M&A as a multidimensional optimal-stopping problem and incorporating a bargaining mechanism through which the target captures part of the incremental technological surplus. The resulting framework allows acquisition value, optimal timing, and the likelihood of acquisition commitment to be analyzed jointly.

3 Methodology Description

This paper develops a single-stage real-options model of cross-border acquisition timing, in which the acquirer retains flexibility before irreversible commitment. The baseline optimal-stopping model is extended with a bargaining mechanism to capture surplus sharing between the acquirer and target.

3.1 Uncertainty and Stochastic Processes

The acquisition opportunity is driven by target-firm value, exchange-rate movements, and technological uncertainty. All state variables are evaluated on a common monthly grid ($\Delta t = 1/12$); continuous-time diffusion equations motivate the stochastic environment, while valuation uses their discrete-time simulation.

Firm Value Process

Let V_t denote the target firm's stand-alone value in U.S. dollars. Under the benchmark valuation measure, the traded market-based component follows a GBM with risk-neutral drift:

$$\frac{dV_t}{V_t} = \mu_V^Q\, dt + \sigma_V\, dW_t^V, \qquad \mu_V^Q = r_{\text{USD}} - \delta_V \tag{1}$$

where μ_V^Q is the risk-neutral drift, r_{USD} the U.S. risk-free rate, δ_V a payout or convenience-yield adjustment, σ_V annualized firm-value volatility, and W_t^V a standard Brownian motion. The monthly simulation is:

$$V_{n+1} = V_n \exp\left[\left(\mu_V^Q - \frac{1}{2}\sigma_V^2\right)\Delta t + \sigma_V\sqrt{\Delta t}\,\varepsilon_{n+1}^V\right] \tag{2}$$

The benchmark uses the risk-neutral drift; historical and managerial-growth estimates are retained only as sensitivity scenarios.

Exchange Rate Dynamics

Let S_t denote the EUR/USD exchange rate in U.S. dollars per euro. Because the acquisition is priced in U.S. dollars but evaluated in euros, currency movements affect the euro-denominated payoff:

$$\frac{dS_t}{S_t} = \mu_S^Q\, dt + \sigma_S\, dW_t^S, \qquad \mu_S^Q = r_{\text{USD}} - r_{\text{EUR}} \tag{3}$$

where r_{USD} and r_{EUR} are the U.S. and euro-area risk-free rates. The monthly process is:

$$S_{n+1} = S_n \exp\left[\left(\mu_S^Q - \frac{1}{2}\sigma_S^2\right)\Delta t + \sigma_S\sqrt{\Delta t}\,\varepsilon_{n+1}^S\right] \tag{4}$$

The firm-value and exchange-rate shocks may be correlated with coefficientρ_{VS}, estimated from target-stock and EUR/USD returns over the same calibration window.

Technological Uncertainty

Technological uncertainty is represented by a two-state Markov process $Xn \in \{x_L, x_H\}$, corresponding to low- and high-synergy states. Because this state is not directly traded, its transition probabilities are scenario-based but anchored to transaction and industry information.

$$P(\Delta t) = \begin{pmatrix} p_{LL}(\Delta t) & p_{LH}(\Delta t) \\ p_{HL}(\Delta t) & p_{HH}(\Delta t) \end{pmatrix} \tag{5}$$

Because decisions are evaluated monthly, annual transition assumptions are converted to the corresponding monthly probabilities:

$$p_{LH}(\Delta t) = 1 - \left(1 - p_{LH}^{\text{annual}}\right)^{\Delta t}, \qquad p_{HH}(\Delta t) = \left(p_{HH}^{\text{annual}}\right)^{\Delta t} \tag{6}$$

with $p_{LL}(\Delta t) = 1 - p_{LH}(\Delta t)$ and $p_{HL}(\Delta t) = 1 - p_{HH}(\Delta t)$. This ensures consistency between technological transitions and the monthly simulation grid.

The state vector is therefore $Y_n = (V_n, S_n, X_n)$, yielding a Markov decision problem on a common discrete-time state space.

3.2 Construction of Real Options Framework

The acquisition opportunity is modeled as a pre-commitment option to defer. Before a binding offer or definitive agreement, the acquirer may continue due diligence, assess technological fit, monitor market and exchange-rate conditions, negotiate preliminary terms, or abandon the opportunity. Post-announcement execution, where contractual flexibility is substantially reduced, lies outside the option window.

Acquisition Payoff

The payoff is evaluated from the acquirer's perspective in euros. Technological synergy is additive and depends on the technological state:

$$\Omega_n = \Theta X_n \tag{7}$$

where Θ is the scale of potential technological and strategic synergy and X_n is the technological state. The immediate acquisition payoff is:

$$\Pi_n = \frac{V_n + \Theta X_n - P_n}{S_n} - I \tag{8}$$

where V_n is target stand-alone value, P_n the acquisition price (both in USD), S_n the EUR/USD rate, and I integration cost in euros. The benchmark price is anchored to the observed transaction value and becomes partly endogenous in the bargaining extension.

Optimal Stopping Problem

Let T be the finite investment horizon and τ a stopping time on {0, Δt, $2\Delta t$, ..., T}. The value of the acquisition opportunity is:

$$F(0, Y_0) = \max_{\tau} E^Q[e^{-r_{\text{EUR}}\tau} \max(\Pi_\tau, 0) \mid Y_0] \tag{9}$$

where r_{EUR} is the euro-denominated risk-free discount rate and E^Q denotes expectation under the benchmark valuation measure. The Bellman recursion is:

$$F(n, Y_n) = \max\{\Pi(n, Y_n), e^{-r_{\text{EUR}}\Delta t} E^Q[F(n+1, Y_{n+1}) \mid Y_n]\} \tag{10}$$

Acquisition is optimal when immediate payoff is at least as large as the continuation value; otherwise, the acquirer preserves the option to wait.

Least-Squares Monte Carlo Valuation

Because the model contains multiple continuous state variables and a discrete technological state, the stopping problem has no closed-form solution. It is therefore solved using Least-Squares Monte Carlo (LSMC) following Longstaff and Schwartz (2001).

The model simulates M paths of Y_n on the monthly grid: V_n and S_n follow Equations (2) and (4), while X_n evolves according to $P(\Delta t)$.

At terminal date N, the path value is:

$$F_N^{(m)} = \max\left\{\Pi_N^{(m)}, 0\right\} \tag{11}$$

Backward induction then discounts the future path value at each decision date:

$$Z_{n+1}^{(m)} = e^{-r_{\text{EUR}}\Delta t} F_{n+1}^{(m)} \tag{12}$$

The continuation value is estimated by least-squares regression:

$$C(n, Y_n) = \mathbb{E}^Q[Z_{n+1} \mid Y_n] \approx \sum_{k=1}^{K} a_{k,n} \phi_k(Y_n) \tag{13}$$

The basis includes a constant, linear and quadratic terms, and interactions in V_n S_n, and X_n, allowing nonlinear continuation values.

Exercise occurs when:

$$\text{Exercise at } n \text{ if } \Pi_n^{(m)} \geq C\left(n, Y_n^{(m)}\right) \; and \; \Pi_n^{(m)} > 0 \tag{14}$$

The option value is the average optimally stopped and discounted payoff:

$$\hat{F}(0) = \frac{1}{M} \sum_{m=1}^{M} e^{-r_{\text{EUR}}\tau_m} \max\left(\Pi_{\tau_m}^{(m)}, 0\right) \tag{15}$$

All simulated paths are retained in the continuation regression. This preserves information on the value of waiting when immediate exercise is unattractive and improves numerical stability when exercise is rare.

3.3 Construction of Game-Theoretic Framework

Bargaining Structure

The benchmark model treats the acquisition price as externally anchored. To capture negotiation, the target is allowed to appropriate part of the expected technological surplus.

Let β denote target bargaining power. Although β is theoretically defined over [0,1], the empirical analysis restricts the scenario range to [0,0.5]to focus on economically plausible bargaining outcomes in which the target captures no more than half of the incremental technological surplus, while avoiding extreme surplus-allocation cases given that the observed acquisition price already incorporates a transaction premium.

Endogenous Acquisition Payoff

Let P_{base} denote the price component reflecting the target's stand-alone value and baseline premium, and $\Omega_n = \Theta X_n$ the technological surplus. The negotiated price is:

$$P(n,\beta) = P_{\text{base}} + \beta\Omega_n \tag{16}$$

The acquirer's immediate payoff becomes:

$$\Pi_n^B = \frac{V_n + \Omega_n - P(n,\beta)}{S_n} - I \tag{17}$$

Substitution gives:

$$\Pi_n^B = \frac{V_n + (1-\beta)\Theta X_n - P_{\text{base}}}{S_n} - I \tag{18}$$

Thus, bargaining determines the share of technological surplus retained by the acquirer rather than scaling the entire acquisition payoff.

Calibration of Bargaining Power

Because β is not directly observable, the announced premium provides an empirical anchor, while alternative bargaining conditions are evaluated using a scenario grid. A surplus-based proxy is:

$$\beta_{\text{obs}} \approx \frac{P_{\text{obs}} - V_0 - \text{Premium}_{\text{base}}}{E[\Theta X]} \tag{19}$$

where P_{obs} is the observed transaction value and $\text{Premium}_{\text{base}}$ captures the part of the premium attributable to general control benefits rather than technological surplus. Because this decomposition is imperfect, the analysis does not rely on a single estimated β.

Interaction with Optimal Stopping

Bargaining affects optimal stopping through the immediate payoff. A higher β reduces retained synergy and can lower option value and exercise probability, but it does not alter the stochastic processes for $Vn, Sn, or\ Xn$. Its effect on timing may therefore be weaker than its effect on value, particularly when waiting is driven mainly by uncertainty resolution.

4 Empirical Results

This section applies the proposed framework to the Siemens–Altair acquisition as an illustrative high-technology cross-border M&A case. The empirical analysis covers baseline real-options valuation and optimal timing, sensitivity analysis, bargaining effects, and robustness checks.

4.1 Case Description

The empirical application focuses on Siemens AG's acquisition of Altair Engineering Inc., announced in October 2024 and completed in March 2025. Siemens offered USD 113 per share in an all-cash transaction, implying an enterprise value of approximately USD 10 billion and a premium of about 19% over Altair's unaffected closing price.

Siemens is a global industrial technology company, while Altair specializes in simulation software, high-performance computing, and data analytics. The transaction therefore provides a suitable setting for examining cross-border acquisition timing under firm-value, exchange rate, and technological uncertainty.

The model captures managerial flexibility during the pre-commitment evaluation stage rather than the post-announcement execution period. Accordingly, the announced transaction is used as an empirical calibration benchmark, while the option itself represents the flexibility available before irreversible acquisition commitment.

4.2 Parameter Specification and Calibration

The Siemens–Altair calibration distinguishes three types of inputs: market-based parameters, risk-neutral drift and discounting parameters, and observable-anchored scenario parameters.

Market-based inputs are estimated from pre-announcement Altair and EUR/USD data over October 2022–October 2024. The initial firm value is USD 8,192.528 million, firm-value volatility is 33.56%, the initial EUR/USD rate is 1.0815, exchange-rate volatility is 7.05%, and the estimated correlation between firm-value and exchange-rate shock is 0. 083014.

Risk-neutral valuation uses a euro discount rate of 2.31% and a U.S. risk-free rate of 4.17%. Accordingly, the benchmark firm-value drift is 4.17%, while the exchange-rate drift is 1.86%, reflecting the interest-rate differential. The historical Altair drift is retained only as a sensitivity scenario rather than as the valuation benchmark.

Parameters that cannot be directly estimated from traded market data are calibrated from transaction information and treated as scenario-based inputs. The synergy scale Θ is set to USD 1.0 billion, informed by Siemens' disclosed expectations of more than USD 500 million in medium-term annual revenue impact and more than USD 150 million in annual cost synergies. Technological synergy is represented by $X \in \{0.75,1.25\}$ with annual transition probabilities $p_{LH} = 0.3$ and $p_{HH} = 0.8$, converted to the monthly grid as described in Section 3.1.3. The acquisition price is anchored to the announced enterprise value of approximately USD 10 billion, while integration cost is set to 3% of transaction value, or EUR 277.35 million. Bargaining power is varied over $\beta \in \{0,0.5\}$, with $\beta = 0$ corresponding to the no-bargaining benchmark.
All reported baseline results use a two-year horizon, monthly decision dates, 100,000 simulated paths, and the full polynomial LSMC basis containing linear, squared, and interaction terms. Table 4.1 summarizes the calibration.

Parameter	Baseline treatment	Calibration basis
Initial firm value V_0	USD 8,192,528 thousand	Unaffected Altair enterprise value before announcement; market-based anchor.
Firm-value volatility σ_V	33.56%	Annualized from Altair daily stock returns over Oct 2022-Oct 2024.
Initial exchange rate S_0	1.0815 USD/EUR	EUR/USD rate at valuation date
Exchange-rate volatility σ_S	7.05%	Annualized from daily EUR/USD log returns over the same estimation window.
Correlation ρ_{VS}	0.083014	Correlation between Altair daily stock returns and EUR/USD daily log returns.
Euro discount rate r_{EUR}	2.31%	Two-year euro area AAA / German Bund risk-free rate around 30 Oct 2024
U.S. risk-free rate r_{USD}	4.17%	Two-year U.S. Treasury yield
Firm-value drift μ_V^Q	4.17%	Risk-neutral drift, assuming no payout adjustment
Exchange-rate drift μ_S^Q	1.86%	Interest-rate differential between U.S. and euro rates
Synergy scale Θ	USD 1,000,000 thousand	Anchored to Siemens disclosed revenue and cost synergies
Technological state X	0.75,1.25	Low- and high-synergy realization states
Transition probabilities	Annual $p_{LH} = 0.3$;$p_{HH} = 0.8$, converted monthly	Moderate improvement and high persistence assumptions
Acquisition price P	USD 10,000,000 thousand	Announced Siemens–Altair enterprise-value benchmark
Integration cost I	EUR 277,350 thousand	3% of transaction value converted into euros
Bargaining power β	0-0.5, $\beta = 0$ benchmark	Scenario range; target share of incremental technological surplus
Time horizon T	2 years	Pre-commitment evaluation horizon.
Time step Δt	1/12	Monthly decision interval
Simulation paths M	100,000	simulation setting

Table 4.1 Parameter Specification and Calibration

4.3 Real Options Valuation

This section presents the baseline results for the Siemens–Altair acquisition opportunity under the risk-neutral calibration in Section 4.2. The main outcomes are summarized by the acquisition option value, optimal exercise timing, exercise probability, and state-dependent exercise behavior.

Option Value

The baseline LSMC valuation gives an acquisition option value of EUR 1,220.64 million, with a standard error of EUR 6.16 million. The positive option value reflects the economic benefit of retaining flexibility before making an irreversible acquisition commitment.

For Siemens, this flexibility is particularly relevant because the attractiveness of acquiring Altair depends jointly on the target's future stand-alone value, exchange-rate conditions, and the realization of technological synergies. The result is consistent with the strategic rationale of the transaction, in which Altair's simulation, data analytics, and AI capabilities represented uncertain future strategic benefits rather than purely current financial value.

Optimal Timing

The baseline average exercise time is 1.5416 years, while the exercise probability is 39.53%. Acquisition commitment is therefore delayed in most simulated paths and occurs only when the immediate payoff becomes sufficiently attractive relative to the continuation value of waiting.

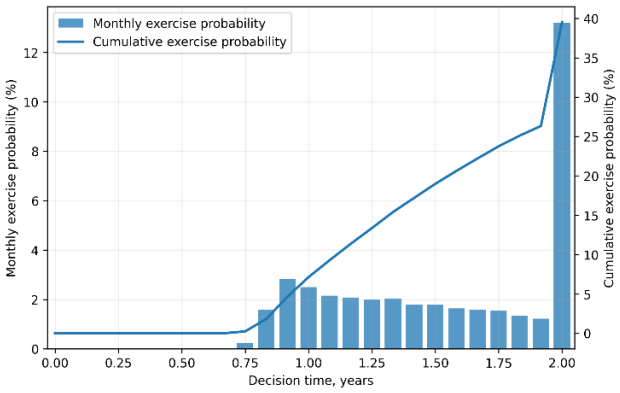


Figure 4.1 Exercise Intensity

Figure 4.1 shows limited exercise activity during the early part of the two-year horizon, followed by increasing commitment as the end of the decision period approaches. This pattern reflects the value of preserving flexibility while information about market conditions and technological integration becomes available. In the Siemens–Altair setting, the result is consistent with the economic role of screening, due diligence, technological assessment, financing preparation, and negotiation before irreversible commitment.

Exercise Region and Optimal Stopping Boundary

Figure 4.2 illustrates state-dependent exercise and continuation regions. Exercise is concentrated primarily in higher target-value states, whereas continuation dominates when target value remains lower. Exchange-rate conditions also affect the decision, but their influence is weaker than the target-value dimension.

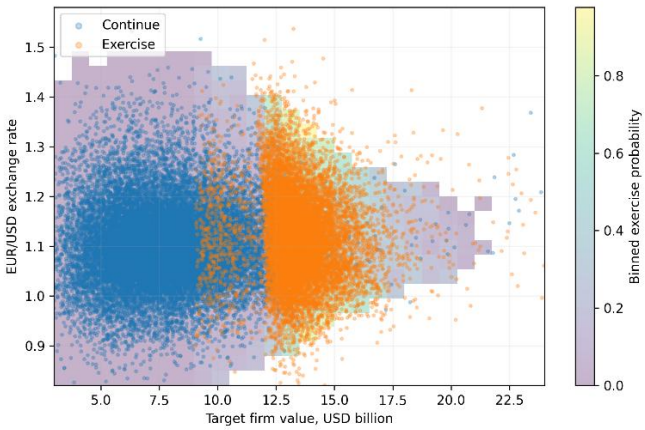


Figure 4.2 Exercise Region

The overlap between exercise and continuation state shows that acquisition cannot be characterized by a single threshold in firm value or exchange rates alone. Instead, the stopping decision depends jointly on firm value, currency conditions, and technological synergy. This finding supports the use of a multidimensional LSMC framework and indicates that Siemens should commit only when the immediate strategic and financial payoff exceeds the value of continued evaluation.

4.4 Sensitivity Analysis

Sensitivity analysis examines the effects of firm-value drift, exchange-rate drift, firm-value volatility, technological synergy, and integration cost on acquisition value and exercise timing. Figure 4.3 shows that acquisition value is most responsive to firm-value dynamics, particularly expected growth and volatility, while exchange-rate movements and integration costs have comparatively smaller effects. Technological synergy also has a clear positive effect on acquisition value, whereas its influence on exercise timing is limited.

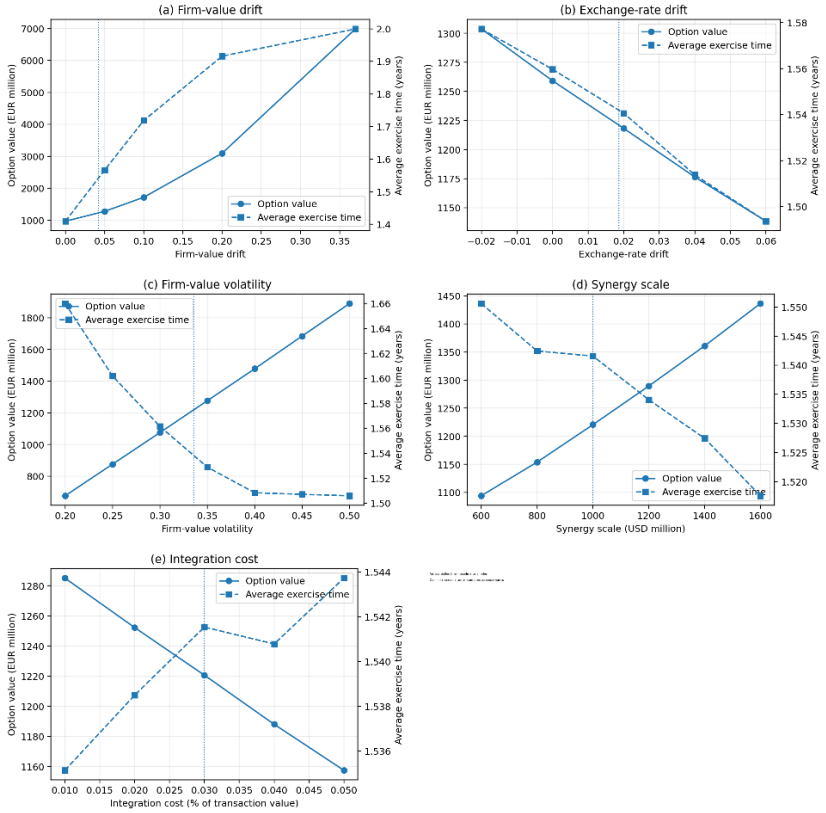


Figure 4.3 Sensitivity of Acquisition Option Value and Exercise Timing to Key Parameters

These results are consistent with the strategic logic of the Siemens–Altair transaction. Siemens acquired Altair primarily to strengthen its capabilities in industrial software, simulation, data analytics, and AI rather than to exploit short-term financial or currency conditions. Accordingly, stronger expected growth and greater technological synergy substantially increase the value of preserving the acquisition opportunity. Higher firm-value volatility also increases option value because Siemens can retain flexibility while uncertainty is resolved and committed when favorable states emerge. This highlights the economic role of the pre-commitment option: uncertainty is not only a source of risk but can also create value when the acquirer retains discretion over timing. The original case analysis likewise identifies growth, volatility, and technological synergy as the principal valuation drivers.

By contrast, exchange-rate movements and integration costs mainly affect the financial attractiveness of the transaction rather than its underlying strategic rationale. Currency conditions change the euro value of the dollar-denominated acquisition surplus, while higher integration costs reduce the net value available from combining the two firms. However, neither factor materially changes the strategic complementarity between Siemens and Altair. Similarly, technological synergy mainly affects how valuable the acquisition is rather than precisely when commitment occurs. Overall, the sensitivity results suggest that the acquisition decision is driven primarily by expectations regarding Altair's future economic and technological value, while exchange-rate exposure and integration costs act as secondary financial constraints on the decision.

4.5 Game-Theoretic Results

This section extends the baseline real-options model by incorporating bargaining between Siemens and Altair. The bargaining parameter measures the share of incremental technological surplus captured by the target through the negotiated acquisition price. Figure 4.4 summarizes the effects of bargaining on acquisition value, exercise timing, and the exercise boundary.

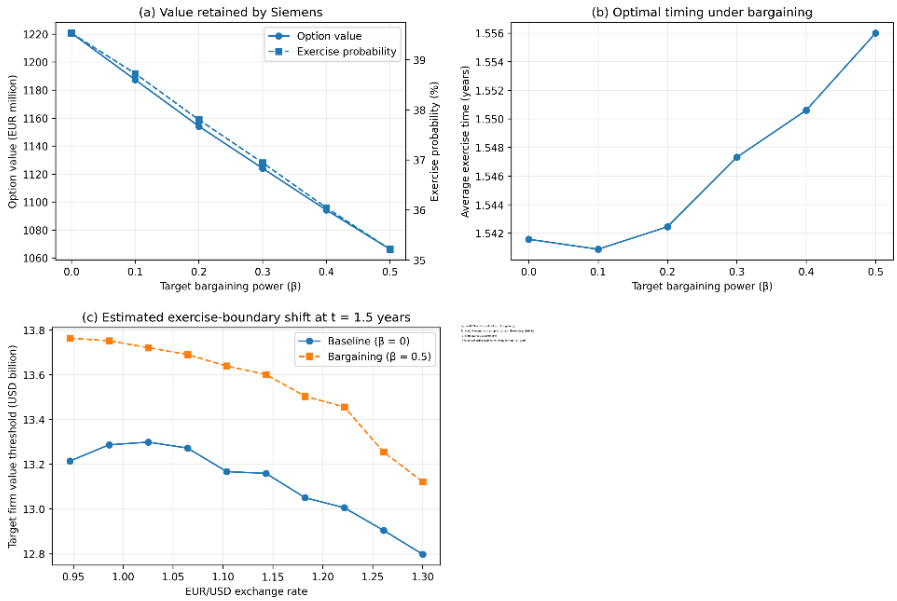


Figure 4.4 Game-Theoretic Effects of Bargaining

The results show that stronger target bargaining power mainly reduces the value retained by Siemens and lowers the likelihood of acquisition commitment, while its effect on timing remains limited. As β increases from 0 to 0.5, the acquisition option value declines from EUR 1,220.64 million to EUR 1,066.18 million, and the exercise probability declines from 39.53% to 35.21%. By contrast, the average exercise time increases only slightly from 1.5416 years to 1.5560 years. Bargaining therefore affects acquisition value and deal likelihood much more strongly than optimal timing.

This pattern is consistent with the Siemens–Altair transaction. Altair's capabilities in simulation, high-performance computing, data analytics, and AI increased both the strategic value of the acquisition and Altair's bargaining position. A larger share of the expected technological surplus can therefore be capitalized into the acquisition price rather than retained by Siemens. The resulting boundary shift toward more favorable target-value states indicates that stronger bargaining makes Siemens more selective. Thus, uncertainty primarily determines when acquisition becomes attractive, whereas bargaining determines how much surplus Siemens retains and how demanding the commitment condition becomes.

4.6 Robustness Check

This section examines whether the main findings are robust to alternative numerical specifications and technological-transition assumptions. The robustness analysis therefore considers two dimensions: numerical robustness, covering exercise frequency and LSMC regression specification, and calibration robustness, focusing on the technological-state transition probabilities.

Numerical Robustness

Numerical robustness is examined by varying both the frequency of exercise opportunities and the regression basis used to estimate continuation values. The baseline specification uses monthly exercise dates and a polynomial basis containing linear, squared, and interaction terms.

Test	Specification	Option value (EUR million)	Average exercise time (years)	Exercise probability
Baseline	Monthly; full basis	**1220.64**	**1.5416**	**39.53%**
Exercise frequency	Quarterly	1262.42	1.6866	38.59%
	Semiannual	1277.97	1.8052	38.01%
Regression basis	Simple	1218.70	1.5459	39.33%
	Extended	1257.60	1.7056	38.56%

Table 4.2 Robustness to Exercise Frequency and Regression Basis Specification

The results indicate that the main conclusions are not sensitive to numerical implementation. Coarser exercise grids produce later average exercise, while exercise probabilities remain relatively stable. The simple basis produces results close to the baseline, whereas the extended basis yields somewhat higher continuation values and later exercise. Overall, the monthly full-basis specification provides a reasonable balance between timing flexibility and nonlinear approximation.

Robustness to Technological Transition Probabilities

This robustness check addresses uncertainty in the technological-state calibration. Because technological transition probabilities cannot be estimated as directly as market-based parameters, the baseline assumptions are compared with lower and higher technological-improvement scenarios. This test is particularly relevant to the Siemens–Altair case, where acquisition value depends substantially on the realization of technological complementarities.

Scenario	Annual (p_{LH})	Annual (p_{HH})	Option value, EUR million	Average exercise time, years	Exercise probability
Low technological improvement	0.20	0.70	1,198.5	1.5367	38.96%
Baseline transition	0.30	0.80	1,220.64	1.5416	39.53%
High technological improvement	0.40	0.90	1,240.86	1.5454	40.10%

Table 4.3 Robustness to Technological Transition Probabilities

More favorable technological transitions increase option value and exercise probability, while average exercise time remains close to the baseline. For Siemens–Altair, this indicates that stronger prospects for realizing synergies in simulation, AI, and industrial software integration increase acquisition attractiveness without materially changing the timing mechanism. The main conclusion of substantial option value and delayed, selective commitment therefore remains robust to alternative technological-transition assumptions.

5 Conclusion

This paper develops an integrated framework combining real-options analysis and strategic bargaining to examine acquisition valuation and timing in cross-border M&A under uncertainty. The acquisition opportunity is modeled as a pre-commitment option to defer and solved using the Least-Squares Monte Carlo method. The framework incorporates target-firm value, exchange-rate dynamics, technological uncertainty, and bargaining over incremental technological surplus. The Siemens–Altair transaction provides an illustrative application to a high-technology cross-border acquisition.

Regarding RQ1, the results show that multiple sources of uncertainty generate substantial acquisition option value and state-dependent exercise behavior. The baseline option value is approximately EUR 1.22 billion, with an exercise probability of 39.53%. Sensitivity analysis indicates that firm-value growth, firm-value volatility, and technological synergy are the main drivers of acquisition value, whereas exchange-rate movements and integration costs play smaller financial roles. This suggests that the economic attractiveness of acquiring Altair was driven primarily by expectations about future firm value and technological complementarity rather than by short-term currency conditions.

Regarding RQ2, optimal acquisition timing is determined by the trade-off between immediate acquisition payoff and the continuation value of waiting. The baseline average exercise time is 1.5416 years, indicating that immediate commitment is generally not optimal. Instead, the acquirer benefits from preserving flexibility while information about firm value, technological fit, and market conditions is resolved. This interpretation is consistent with the role of due diligence, technological assessment and negotiation during the pre-commitment stage of a large technology acquisition.

Regarding RQ3, bargaining mainly affects how much value is retained by the acquirer and whether the acquisition remains sufficiently attractive to proceed. As target bargaining power increases from 0 to 0.5, Siemens' retained option value declines from EUR 1,220.64 million to EUR 1,066.18 million, while the exercise probability falls from 39.53% to 35.21%. Average exercise time changes only slightly, from 1.5416 to 1.5560 years. The exercise boundary also shifts toward more favorable target-value states. These results suggest that uncertainty primarily determines when acquisition becomes attractive, whereas bargaining affects how much surplus the acquirer retains and how selective the commitment decision becomes.

Robustness checks confirm that the main conclusions remain stable across alternative numerical and technological-transition assumptions, supporting the paper's central contribution of linking multidimensional uncertainty, optimal stopping, and strategic surplus sharing within a unified M&A decision framework.
Several limitations remain. The empirical application is based on a single case, and several parameters—particularly technological synergy, technological-transition probabilities, integration cost, and bargaining power—remain partly

scenario-based. The bargaining mechanism is also deliberately parsimonious rather than a fully dynamic negotiation model. Future research could extend the framework to multiple acquisitions, competing bidders, regulatory approval risk, staged payments, earnouts, or dynamic bargaining.

References

Longstaff F A, Schwartz E S (2001) Valuing American options by simulation: A simple least-squares approach. The review of financial studies 14(1): 113-147

Paudyal K, Thapa C, Koirala S, Aldhawyan S (2021) Economic policy uncertainty and cross-border mergers and acquisitions. Journal of Financial Stability 56: 100926

Bonaime A, Gulen H, Ion M (2018) Does policy uncertainty affect mergers and acquisitions?. Journal of financial economics 129(3): 531-558

Si D K, Li H X, Hu D, Li X L (2024) The uncertainty of RMB exchange rate expectation and foreign investment: Evidence from Cross-border M&As in China. Pacific-Basin Finance Journal 83: 102213

Li R, Zhang R (2024) How does exchange rate fluctuation affect the efficiency of Chinese enterprises' overseas mergers and acquisitions. Finance Research Letters 66: 105600

Li S, Ang J S, Wu C, Yang S (2021) Valuing technological synergies in mergers. The North American Journal of Economics and Finance 58: 101464

Wei J, Wang A, Ma X (2025) M&A and resources allocation in high-tech industries: Technology diffusion or market power effect. International Review of Economics & Finance: 104691

McDonald R, Siegel D (1986) The value of waiting to invest. The quarterly journal of economics 101(4): 707-727

Dixit A K, Pindyck R S (1994) Investment under uncertainty. Princeton university press

Smith J E (2005) Alternative Approaches for Solving Real-Options Problems: (Comment on Brandão et al. 2005). Decision Analysis 2(2): 89-102

Ahmadi R, Bratvold R B (2023) An exposition of least square Monte Carlo approach for real options valuation. Geoenergy Science and Engineering 222: 111230

Nadarajah S, Margot F, Secomandi N (2017) Comparison of least squares Monte Carlo methods with applications to energy real options. European Journal of Operational Research 256(1): 196-204

Najafi P, Aghaei M (2024) Real options analysis of seawater desalination with solar photovoltaic energy systems: A least-squares stochastic perspective. Desalination 586: 117823

Ebina T, Kumakura Y, Nishide K (2022) Hostile takeovers or friendly mergers? Real options analysis. Journal of Corporate Finance 77: 102292

Dunis C L, Klein T (2005) Analysing mergers and acquisitions in European financial services: An application of real options. The European Journal of Finance 11(4): 339-355

Barbopoulos L G, Cheng L T, Cheng Y, Marshall A (2019) The role of real options in the takeover premia in mergers and acquisitions. International Review of Economics & Finance 61: 91-107

Čirjevskis A (2020) Do synergies pop up magically in digital transformation-based retail M&A? Valuing synergies with real options application. Journal of Open Innovation: Technology, Market, and Complexity 6(1): 18

van den Honert R C, Stewart T J (1992) A game-theoretic model for mergers and acquisitions. European journal of operational research 59(2): 275-287

Yu J, Xu B (2011) The game analyses to price the target enterprise of merger and acquisition based on the perspective of real options under stochastic surroundings. Economic Modelling 28(4): 1587-1594

Lukas E, Reuer J J, Welling A (2012) Earnouts in mergers and acquisitions: A game-theoretic option pricing approach. European Journal of Operational Research 223(1): 256-263

Price of Inertia in Emergency Medical System Adjustment

Lýdia Gábrišová[1], Jaroslav Janáček[2], Marek Kvet[3]

Abstract. Emergency Medical Service systems play a crucial role in providing timely pre-hospital care, making the optimal deployment of ambulance stations essential for minimizing response times and improving accessibility. Traditional location models formulate this task as a weighted *p*-median problem and seek an entirely new optimal deployment of ambulances. However, these approaches generally overlook practical constraints associated with reorganizing existing systems, where relocating established stations is often costly, politically sensitive, or operationally infeasible. This paper introduces a novel optimization framework for evaluating the Price of Inertia in emergency service planning. The proposed approach preserves the current deployment of ambulance stations and determines the minimum number of additional ambulances required to achieve the service quality of a fully optimized network. The mathematical model is based on an enhanced radial formulation incorporating generalized disutility to account for the stochastic availability of ambulances and allowing multiple ambulances to be located at a single site. The Price of Inertia is interpreted as the operational penalty incurred by maintaining the existing infrastructure instead of implementing the optimal redesign. Unlike conventional measures based solely on performance loss, the proposed framework quantifies inertia in terms of the additional resources needed to compensate for relocation constraints. The methodology provides decision makers with a practical tool for strategic emergency medical service capacity planning by balancing service quality improvements against the costs and challenges of network reconfiguration.

Keywords: Location problems, emergency services, optimization, price of inertia

JEL Classification: C44, C61

1 Introduction

When people's lives or health are in imminent danger, the government established the Emergency Medical Service (EMS) to provide them with pre-hospital healthcare as soon as possible. As a component of the most advanced form of individualized patient care, resuscitation and intensive care, it is crucial and irreplaceable in the delivery of urgent care. Private EMS services and companies usually provide this service, but only if they have the proper licensing to run a medical facility. According to Brotcorne et al. (2003) and Ingolfsson et al. (2008), this system is included in the larger category of public service systems, which includes the integrated rescue system.

A public service system is defined by a group of served entities, namely clients with unique needs, and by service providers, which may encompass both private and public organizations, including corporations and governmental agencies. The clientele of a specific system is mostly located in a designated geographical area, although the resources that meet their needs are frequently centered in one or more places, which may exist both within and outside the supplied territory (Jánošíková, 2007). All endeavors to enhance the rescue system are unified by the imperative to attain the utmost availability and accessibility of service for all clients, while adhering to all restricting limitations. The proximity of the nearest available center from where the ambulance is dispatched to meet the patient's needs directly correlates with improved accessibility (Doerner et al., 2005). In emergency scenarios, distance and response time are undeniably critical factors. Identifying the ideal placement of service centers is a complex combinatorial problem with a nearly infinite array of viable options. Such issues cannot be resolved without the availability of advanced software help. Consequently, a mathematical model is typically developed to comprehend the combinatorial nature of the related decision-making process. The decision-making challenge involves identifying optimal sites for service centers to maximize a specified objective function, hence

[1] University of Žilina, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovakia, lydia.gabrisova@fri.uniza.sk
[2] University of Žilina, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovakia, jaroslav.janacek@fri.uniza.sk
[3] University of Žilina, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovakia, marek.kvet@fri.uniza.sk

enhancing service accessibility for clients. The accessibility of emergency services is typically measured by the average travel time or the geographical distance of the nearest ambulance to the emergency site. Consequently, the mathematical model of this issue can be represented as a weighted *p*-median problem, which has been extensively examined by numerous researchers regarding efficient exact or approximate solving methods. However, it is important to acknowledge that numerous alternative objectives may be optimized in lieu of the aforementioned one (Jánošíková et al., 2021; Reuter-Oppermann et al., 2017).

The associated branch-and-bound method solving algorithm typically fails with large instances when the mathematical model of the problem is described in the location-allocation form. This is because the algorithm introduces a matrix of binary variables to model the assignment of clients to service centers, which causes enormous computational requirements and memory demands. This inconvenience can be solved by substituting the original model with the radial formulation of the problem (Avella et al., 2007; García et al., 2011; Janáček, 2008). Since it is unnecessary to model the aforementioned assignment, it streamlines the solution of numerous integer-programming issues with big instances. Moreover, it may be readily modified to further expedite the computation (Janáček and Kvet, 2016; Kvet, 2015). Although initially conceived as an approximate method, it allows us to achieve the optimal solution rapidly.

Standard mathematical models derived from the weighted *p*-median problem may exhibit two significant shortcomings that indicate a considerable degree of simplification. The primary discomfort lies in the unpredictable behavior of real systems. Requests for intervention arise unpredictably, and the nearest EMS station may be momentarily inaccessible due to prior commitments. In this scenario, the ongoing emergency call is managed by the second nearest station or by the nearest available one. The proposed modeling method enhances the realism of the objective function and allows for a partial representation of the stochastic nature of real systems. This notion has been presented as the concept of *generalized disutility* (Jankovič, 2016; Kvet, 2014).

The served region's heterogeneity follows as the second drawback. It is overly rigid to assume that each network node can only host one service center. When we examine actual EMS networks, we find that certain cities have more stations than others. One way to meet this requirement is by allowing multiple facilities to be located in one location, or by decomposing the original network so that a small subnetwork replaces one node (clearly, the original weight of the node needs to be distributed somehow among all new nodes). It may be necessary to seek the assignment of a minimum number of ambulance vehicles to areas with higher population densities in order to meet anticipated demand.

Last but not least, it should be noted that any optimization results will bring a new design of the network structure to the existing system. In practice, this means that some stations will change their location in the network and move to where they are needed more, or where they will bring a more significant improvement in the optimized quality criterion. However, any changes are perceived by the public very sensitively and are often not accepted with understanding. Therefore, in this paper we address the question of how to achieve the required value of the average travel time while respecting the location of all current stations, minimizing the number of stations that need to be added to the network.

Speaking of the price of inertia from the paper title, it can be understood as follows: The *Price of Inertia* (*PoI*) quantifies the performance loss incurred by maintaining the existing EMS facility configuration instead of adopting the optimal configuration obtained by solving the location problem under current demand and network conditions. It is measured as the relative difference between the objective value of the existing configuration and that of the optimized configuration.

The principal contribution of this work is the introduction of a new optimization framework for evaluating the Price of Inertia in EMS systems. Unlike conventional ambulance location models that assume unrestricted relocation of existing stations, the proposed model preserves the current infrastructure and determines the minimum number of additional ambulances required to achieve the service quality of an optimally redesigned system. This provides decision makers with a quantitative measure of the operational cost of maintaining the existing EMS configuration and offers a practical tool for strategic capacity planning under relocation constraints.

The structure of this paper is divided into 4 basic parts. The introductory chapter is followed by a theoretical part, in which we describe the mathematical framework of the optimization problem. In the third part, we present the results of numerical experiments for the tested benchmarks. The last part summarizes the achieved goals and provides relevant conclusions.

2 Mathematical Formulation of the Problem

In order to provide the readers with the mathematical framework of the associated optimization problem, the following denotations are necessary to be introduced.

Let the symbol J denote the finite set of places of the serviced region, where demands for service are generated. Symbol w_j denotes the frequency of demands for service generated at place $j \in J$. Symbol I denotes the finite set of places, at which an ambulance can be located. Symbol I^c denotes the set of places, at which ambulances are currently located. Symbol y_i^c denotes the number of ambulances currently located at place $i \in I^c$. Symbol T^d denotes the demanded average response time of the system. The number r represents the maximal considered order of ambulance placement in order of increasing time distance from the demand location. In other words, the value r represents the number of nearest located stations, which are checked to assign the service request. If we were working only with the basic version of the weighted p-median problem, where a request is always assigned only to the closest station, then the value of r would be equal to 1. If we assume that a request could be assigned to at most the second closest station if the first one were unavailable, then r would take the value of 2. In general, r expresses the number of closest stations that can participate in satisfying a client request. For more details, see (Kvet, 2014). Symbol K denotes range 1, .., r}. Symbol q_k gives probability that the k-th nearest ambulance is the first available for $k \in K$. Symbol V denotes the finite set of radii used in the radial model of the location problem. Symbol e denotes the time difference between a pair of neighboring radii. Symbol u_i denotes an upper bound of the number of ambulances, which can be placed at location $i \in I$. Coefficient a_{ijv} gets its value from the range {0, 1} and the value 1 means that possible ambulance location $i \in I$.is inside the radius $v \in V$ from the place $j \in J$.

As far as the decision variables are concerned, two series of decision variables are to be introduced to determine the emergency medical system. Let y_i model the number of ambulances, which should be located at the place $i \in I$, and let $x_{jvk} \in \{0, 1\}$ take the value of 1 if the k-th closest ambulance is located outside the radius $v \in V$ from the demand location j. For easier orientation in symbols, constants and variables, we have summarized them into the following Table 1 and Table 2.

Symbol	Range	Meaning
I		finite set of places, at which an ambulance can be located
I^c		set of places, at which ambulances are currently located
y_i^c	$i \in I$	Number of ambulances currently located at i
J		finite set of places of the serviced region, where demands for service are generated
r		number of closest located ambulances that can cover current request
K		range {1, .., r}
u_i	$i \in I$	upper bound of the number of ambulances, which can be placed at location $i \in I$
w_j	$j \in J$	frequency of demands for service generated at place $j \in J$
q_k	$k \in K$	probability that the k-th nearest ambulance is the first available for $k \in K$
V		finite set of radii used in the associated radial model of the location problem
e		time difference between a pair of neighboring radii
T^d		demanded average response time of the system
a_{ijv}	$i \in I$, $j \in J$, $v \in V$	gets its value from the set {0, 1} and the value 1 means that possible ambulance location $i \in I$.is inside the radius $v \in V$ from the place $j \in J$.

Table 1 Used symbols and coefficients

Variable	Range	Values	Meaning
y_i	$i \in I$	N_0	number of ambulances, which should be located at the place $i \in I$
x_{jvk}	$j \in J$, $v \in V$, $k \in K$	$\{0, 1\}$	takes the value of 1 if the k-th closest ambulance is located outside the radius $v \in V$ from the demand location $j \in J$

Table 2 Summary of used decision variables

Before formulating a mathematical model of the relevant optimization problem, we must focus on the main criterion of design quality. This is the average time T_{avg} it takes for the nearest available crew to reach the scene of the incident. To express the true average arrival time, we must proceed from the concept of generalized disutility and the stochastic nature of the real system. In real conditions, the request for an ambulance is always assigned to the ambulance that is closest to the free one, which may not be the closest one. If we want to express this mathematically, then let's assume that the distance between nodes i and j will be denoted by the symbol d_{ij}. For the sake of completeness, it should be added that the min_k operator returns the distance to the k-th nearest service center, considering the fact that multiple stations can be in one network node. Then, the average response time can be expressed by the formula (1).

$$T_{avg} = \frac{\sum_{j \in J} w_j \sum_{k \in K} q_k \min_k \left\{ d_{ij} : i \in I,\ y_i \geq 1 \right\}}{\sum_{j \in J} w_j} \tag{1}$$

After all preliminaries the radial model of the optimization problem can be stated as follows:

$$Minimize \quad \sum_{i \in I} y_i \tag{2}$$

$$Subject\ to \quad \sum_{k \in K} x_{jvk} + \sum_{i \in I} a_{ijv} y_i \geq r \quad for\ j \in J,\ v \in V \tag{3}$$

$$\frac{e \sum_{j \in J} w_j \sum_{k \in K} q_k \sum_{v \in V} x_{jvk}}{\sum_{j \in J} w_j} \leq T^d \tag{4}$$

$$y_i \geq y_i^c \quad for\ i \in I \tag{5}$$

$$y_i \leq u_i \quad for\ i \in I \tag{6}$$

$$y_i \in Z_0^+ \quad for\ i \in I \tag{7}$$

$$x_{jvk} \in \{0, 1\} \quad for\ j \in J,\ v \in V,\ k \in K \tag{8}$$

The objective function (2) expresses the total number of deployed ambulances. If we want to evaluate the number of added stations, we can just subtract the value p expressing the current number of stations in the network. The series of constraints (3) ensures the relation among the decisions on locating ambulances and the auxiliary decision variables x_{jvk}. The left side of inequality (4) expresses the average response time of the system, and the constraints assure that the average response time is less than or equal to the demanded value T^d. The series of constraints (5) ensures that the original deployment stays unchanged except for the added ambulances. The series of constraints (6) limits the number of ambulances located at location i. The constraints (7) and (8) are obligatory constraints determining the feasible values of decision variables.

3 Numerical Experiments

3.1 Benchmarks and software tools

The numerical experiments reported in this small study were performed making use of the road network data. The customized radial model of the Emergency Medical Service system satisfying the demands for intervention was used to optimize the systems operated in the self-governing region of Žilina (ZA) in Slovakia and the Moravsko-slezský kraj (MSK) in the Czech Republic. The original number p of ambulances at disposal was 64 for the region MSK and 37 for the region ZA. The cardinalities of the candidate set I were 300 for the benchmark MSK and 300

for ZA. Mentioned set I consists of all cities and villages of the associated region. For the purpose of computational study, the set J was identical to the set I for all problem instances.

The problem (2)-(8) was solved by the optimization software FICO Xpress to obtain the optimal solution of the integer programming problem for the specific input model written in programming language Mosel. The associated computations were run on a PC equipped with the Intel® Core™ i7 8565U processor with the parameters: 1.80 GHz and 128 GB RAM. Solutions of described problems were found by the optimization software FICO Xpress 8.3 (64-bit, release 2017).

3.2 Analysis of current station deployment and results of performed experiments

The values of current and optimal average response times in minutes for MSK are $T^c = 7.076$ and $T^{opt} = 6.430$. The values of current and optimal average response times for ZA are $T^c = 6.819$ and $T^{opt} = 6.00$. The computational time CT to obtain the optimal unconstrained solution T^{opt} was 57.4 seconds for MSK and 30.4 seconds for ZA.

The demanded average response time T^d was set to the value $\alpha T^c + (1-\alpha)T^{opt}$ for α = 0.0, 0.1, …, 1.0. For completeness, the value of r was 5. The probabilities q_k were obtained in the following way: $q_1 = 0.798474$ was computed from real data. The values of the associated probability coefficients q_k for k = 2, …, 5 follow $q_2 = 0.160913$, $q_3 = 0.0324282$, $q_4 = 0.00653513$ and $q_5 = 0.00164939$. Since these parameters model probabilities, their sum over the index k takes 1.

The results for the territorial units ZA and MSK are plotted in Table 3. The symbol *NoAdd* denotes additional number of ambulances, *CT* stands for computational time in seconds and T^* is the resulting average response time in minutes. The column denoted by *PoM* (Price of Minute) gives the number of ambulances per one minute of decrease of the average response time.

α	Region MSK				Region ZA			
	NoAdd	*CT*	T^*	*PoM*	*NoAdd*	*CT*	T^*	*PoM*
0.0	8	75.0	6.430	12.4	7	181.2	6.000	8.5
0.1	7	75.2	6.473	11.6	6	136.0	6.026	7.6
0.2	5	61.1	6.543	9.4	5	154.9	6.119	7.1
0.3	4	56.0	6.584	8.1	4	117.2	6.213	6.6
0.4	3	37.7	6.627	6.7	4	135.9	6.213	6.6
0.5	2	47.0	6.691	5.2	3	100.7	6.331	6.1
0.6	1	34.9	6.761	3.2	2	168.5	6.463	5.6
0.7	1	34.3	6.761	3.2	2	167.1	6.463	5.6
0.8	1	44.0	6.761	3.2	1	56.6	6.601	4.6
0.9	1	37.6	6.761	3.2	1	66.8	6.601	4.6

Table 3 Results of numerical experiments

The reported computational times of the optimization are negligible with respect to the time horizon of the planning addressed by the optimization. Despite this fact, the radial is a suitable choice. Compared to the well-known location-allocation approach, it allows solving tasks of a larger scope. The corresponding radial mathematical model has a simpler structure and is less memory-intensive. From a practical point, it offers simple implementation.

It can be also noticed that the decrease in the number of added ambulances depends on decreased demanded average response time nonlinearly. This characteristic corresponds to nonlinearly increasing price of one minute of spared average response time.

4 Conclusions

This contribution was devoted to the evaluation of inertia associated with the maintenance of the Emergency Medical Service system. Inertia in this context means the tendency to maintain the state of the maintained system determined by previous developments. In general, inertia is justified by the costs that changes would require if the increase in system efficiency is not known.

In the present paper, the cost of inertia is evaluated by the number of ambulances that would have to be added to the current fleet to achieve a system efficiency that corresponds to a completely reorganized system. Performed computational study shows that the optimal average response time can be achieved by adding 12 or 19 percent of the size of the original fleet.

The main contribution of this research paper consists on one hand in the methodology that describes the mathematical apparatus for optimizing the number of stations that need to be added to the system to achieve the desired level of average response time. On the other hand, we also showed several numerical experiments that confirm the applicability of the solution in real-world conditions. In other words, this paper introduced a mathematical optimization framework that quantified the "Price of Inertia" in Emergency Medical Service systems by determining the minimum number of additional ambulances that must be deployed while preserving all existing station locations to achieve the same service quality as a fully optimized relocation of the current fleet. Instead of allowing unrestricted relocation of ambulance stations, the model keeps all existing stations fixed and optimally locates only additional ambulances. The objective was to minimize the number of added ambulances while achieving a prescribed average response time.

While previous EMS location models compare the current and optimal deployments, this paper goes one step further by interpreting inertia as the additional resources required to compensate for the inability (or unwillingness) to relocate existing stations. Therefore, the Price of Inertia is expressed in terms of additional ambulances rather than only as a percentage loss in response time.

Future research in this field may be aimed at developing other modeling techniques to optimize current ambulance deployment under the condition of keeping the average response time at an acceptable limit.

Acknowledgements

This work was supported by the following Slovak national research grants: VEGA 1/0220/25 “Regional equality in access to public rescue services provided from service centers on the transport network”, VEGA 1/0696/26 “Predictions of delays and vehicle occupancies to inform the planning of public transport”, and VEGA 1/0516/26 “Optimal design of charging infrastructure for electric vehicles and its efficient operation in urban mobility conditions”.

References

Avella P, Sassano A, Vasil'ev I (2007) Computational study of large scale p-median problems. Mathematical Programming 109: 89-114. https://doi.org/10.1007/s10107-005-0700-6

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. European Journal of Operations research 147: 451-463. https://doi.org/10.1016/S0377-2217(02)00364-8

Doerner K F, Gutjahr W J, Hartl R F, Karall M, Reimann M (2005) Heuristic Solution of an Extended Double-Coverage Ambulance Location Problem for Austria. Central European Journal of Operations Research 13: 325-340

García S, Labbé M, Marín A (2011) Solving large p-median problems with a radius formulation. INFORMS Journal on Computing 23: 546-556. https://doi.org/10.1287/ijoc.1100.0418

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. Health care management science 11: 262-274. https://doi.org/10.1007/s10729-007-9048-1

Janáček J (2008) Approximate Covering Models of Location Problems. In: Lecture Notes in Management Science: Proceedings of the 1st International Conference ICAOR, Yerevan, Armenia, pp 53-61

Janáček J, Kvet M (2016) Sequential approximate approach to the p-median problem. Computers & industrial engineering 94: 83-92. https://doi.org/10.1016/j.cie.2016.02.004

Jankovič P (2016) Calculating Reduction Coefficients for Optimization of Emergency Service System Using Microscopic Simulation Model. In: CINTI 2016. Conference Proceedings, pp 163-167. https://doi.org/10.1109/CINTI.2016.7846397

Jánošíková Ľ (2007) Emergency Medical Service Planning. Communications 9: 64-68

Jánošíková Ľ, Jankovič P, Kvet M, Zajacová F (2021) Coverage versus response time objectives in ambulance location. International Journal of Health Geographics 20: 1-16. https://doi.org/10.1186/s12942-021-00285-x

Kvet M (2014) Computational Study of Radial Approach to Public Service System Design with Generalized Utility. In: Digital Technologies 2014. Conference Proceedings, Žilina, Slovakia, pp 198-208

Kvet M (2015) Advanced Radial Approach to Resource Location Problems. In: Studies in Computational Intelligence: Developments and Advances in Intelligent Systems and Applications, pp 29-48. https://doi.org/10.1007/978-3-319-58965-7_3

Reuter-Oppermann M, van den Berg P L, Vile J L (2017) Logistics for Emergency Medical Service systems, Health Systems 6: 187-208. https://doi.org/10.1057/s41306-017-0023-x

Dynamic portfolio optimization based on volatility-adaptive LSTM and tail-dependence topological network

Qian Gao[1]

Abstract. Traditional portfolio optimization relies on covariance structures and has limited ability to capture co-movement under extreme market conditions and higher-order dependencies among assets. This study proposes a dynamic portfolio optimization framework integrating volatility-adaptive deep learning and topological network analysis. Its contribution lies in embedding forward-looking return prediction and higher-order network features into the asset allocation process. A volatility-adaptive LSTM model is developed to generate return forecasts responsive to market uncertainty. A tail dependence network is constructed by distinguishing left-tail and right-tail events in return series, and simplicial complexes are used to extract triangular and void structures among assets. Topological features are incorporated into the objective function as penalty terms, extending the traditional framework. Rolling out-of-sample backtesting compares strategies based on historical and prediction-derived information. Empirical results show that incorporating tail-based topology mitigates risk concentration and improves drawdown, while prediction-based strategies enhance risk-adjusted returns. The findings indicate that combining volatility-aware forecasting with higher-order dependency structures an provide a useful extension for asset allocation under extreme market conditions.

Keywords: long short-term memory, network topology, portfolio optimization, risk management, tail dependence

JEL Classification: C45, C53, G11, G17

1 Introdution

Prediction-based portfolio optimization models are capable of capturing short-term investment opportunities (Freitas et al, 2009). Existing studies indicate that machine learning algorithms can improve the estimation of input variables, thereby enabling the construction of more robust portfolios (Martínez-Barbero et al, 2024). In particular, long short-term memory (LSTM) networks are widely used for stock return prediction, and their integration with portfolio optimization using both predictive and historical information improves portfolio performance (Ma et al, 2023). Compared with alternative methods, LSTM demonstrates more stable performance in capturing complex temporal dependencies and is capable of handling long-term lag effects (Hochreiter and Schmidhuber, 1997), which makes it well-suited for prediction-driven portfolio optimization.

At the same time, an aspect of portfolio selection lies in characterizing the dependency structure among asset returns (Ricca and Scozzari, 2024). Recent studies increasingly incorporate complex network approaches to describe inter-asset relationships and apply them to portfolio optimization (Li et al, 2019). Financial networks based on correlation or distance measures reveal market structure and systemic risk. However, most approaches remain limited to pairwise relationships and are less suited to capturing structural features under extreme market conditions, making tail dependence networks an important complement (Wen et al, 2019). Furthermore, higher-order structures among multiple assets, such as triangular relations and cyclic dependencies, play an important role in risk transmission and aggregation (Wei et al, 2025). Related studies also indicate that combining predictive methods with complex network structures can enhance portfolio performance (Freitas and Bertini, 2023).

This study proposes a portfolio optimization framework that integrates prediction and network structures. A volatility-adaptive long short-term memory (VA-LSTM) model is used to generate return forecasts, and a tail dependence network is constructed to characterize risk relationships. Higher-order topological features are extracted and incorporated into the optimization process, enabling the joint modeling of predictive information and dependency-based risk. The results indicate that incorporating topological structures improves risk control and reduces drawdown, while prediction-based strategies enhance risk-adjusted performance. The remainder of the paper is organized as follows. Section two presents the methodological framework, section three provides the empirical analysis, and section four concludes the study and discusses future research directions.

[1] VSB - Technical University of Ostrava, Department of Finance, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czechia, qian.gao@vsb.cz

2 Methodology

This section clarifies the connection among return forecasting, tail-dependence topological structures, and portfolio optimization. The proposed framework is designed to extend standard prediction-based portfolio optimization and, more broadly, covariance-based allocation methods. In conventional minimum-variance or mean-variance allocation, the rolling covariance matrix is estimated only from historical returns. In prediction-based allocation, the same optimization problem is solved after replacing the historical return window with a predicted return window. The proposed model adds layers by constructing left-tail and right-tail topological structures from the same rolling input used to compute the covariance matrix. Therefore, the covariance matrix Σ, the triangle set $\mathcal{T}$, and the hole set $\mathcal{H}$ are all obtained from a clearly defined rolling return matrix, which can be either a historical return matrix or a predicted return matrix.

Let $\mathbf{R} = \{\mathbf{r}_t\}_{t=1}^{T}$ denote the matrix of daily asset returns, where $\mathbf{r}_t = (r_{1t}, \ldots, r_{Nt})^\top$ and N is the number of assets. For the prediction model, the supervised learning input is a sequence of past returns and the target is the next-day return vector,

$$\mathbf{X}_t = [\mathbf{r}_{t-L+1}, \ldots, \mathbf{r}_t], \qquad \mathbf{y}_t = \mathbf{r}_{t+1}, \tag{1}$$

where L is the sequence length. Thus, the output of both the standard LSTM benchmark and the proposed VA-LSTM is a vector of stock return forecasts. These quantities are subsequently estimated from the rolling window of predicted returns.

The VA-LSTM extends a standard LSTM by introducing an external volatility-adaptive gate after the recurrent hidden representation. For each input sequence, a market-wide rolling volatility indicator is computed as

$$v_t = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{W}\sum_{k=t-W+1}^{t}(r_{ik} - \bar{r}_{i,t})^2}, \tag{2}$$

where W is the local volatility window and $\bar{r}_{i,t}$ is the mean return of asset i within this local window. The LSTM produces hidden states $\mathbf{h}_t$, while the volatility gate maps v_t into a hidden-state modulation vector,

$$\mathbf{g}_t = \sigma(\mathbf{W}_g v_t + \mathbf{b}_g), \qquad \tilde{\mathbf{h}}_t = \mathbf{g}_t \odot \mathbf{h}_t, \qquad \widehat{\mathbf{r}}_{t+1} = \mathbf{W}_o\tilde{\mathbf{h}}_t + \mathbf{b}_o. \tag{3}$$

The hidden dimension is a latent representation size controlling the capacity of the recurrent model. It is not required to equal the number of stocks. The empirical section therefore reports a sensitivity check for hidden dimensions 16, 32, and 64, while 32 is used as the baseline specification because it provides a moderate capacity relative to the 30-stock universe and avoids an unnecessarily large model for a short financial time series.

At each rebalancing date s, the optimizer uses a rolling window of length M. The input window is denoted by

$$\mathbf{Z}_s^{(m)} = \begin{cases} [\mathbf{r}_{s-M}, \ldots, \mathbf{r}_{s-1}], & m = \text{historical}, \\ [\widehat{\mathbf{r}}_{s-M}, \ldots, \widehat{\mathbf{r}}_{s-1}], & m = \text{prediction}. \end{cases} \tag{4}$$

The covariance matrix used in the optimization problem is then estimated as

$$\Sigma_s^{(m)} = 252\ \mathrm{Cov}(\mathbf{Z}_s^{(m)}). \tag{5}$$

This definition makes the link between the LSTM and the optimization equation. Historical strategies estimate Σ_s from realized returns. 252 is the number of trading days in a year. Prediction-based strategies estimate Σ_s from the sequence of return forecasts generated by the LSTM model.

The topological component is constructed from the same rolling input matrix $\mathbf{Z}_s^{(m)}$. For the left-tail network, an extreme event occurs when asset i is below its α quantile. For the right-tail network, an extreme event occurs when asset i is above its $1-\alpha$ quantile. The empirical joint tail probability and the normalized tail distance are defined as

$$P_{ij,s}^{-,(m)} = \frac{1}{M}\sum_{t=s-M}^{s-1}\mathbb{I}(z_{it}^{(m)} \le q_{i,\alpha}^{(m)})\mathbb{I}(z_{jt}^{(m)} \le q_{j,\alpha}^{(m)}), \tag{6}$$

$$P_{ij,s}^{+,(m)} = \frac{1}{M}\sum_{t=s-M}^{s-1}\mathbb{I}(z_{it}^{(m)} \ge q_{i,1-\alpha}^{(m)})\mathbb{I}(z_{jt}^{(m)} \ge q_{j,1-\alpha}^{(m)}), \tag{7}$$

$$D_{ij,s}^{\pm,(m)} = \frac{(P_{ij,s}^{\pm,(m)} + \varepsilon)^{-1}}{\max_{a,b}(P_{ab,s}^{\pm,(m)} + \varepsilon)^{-1}}. \tag{8}$$

A smaller value of $D_{ij,s}^{\pm,(m)}$ indicates a stronger joint extreme event relation. A threshold graph is obtained by connecting pairs with distances below a threshold η_s. To avoid an arbitrary fixed threshold, η_s is selected as the empirical quantile of off-diagonal distances corresponding to a target edge density. The triangle set $\mathcal{T}_s$ is then defined as the set of three-cliques in this graph, while $\mathcal{H}_s$ is represented by the cycle basis after removing filled three-cycles. These two sets provide operational measures of higher-order tail co-movement and ring-like dependence structures.

The portfolio weights are obtained from the following long-only optimization problem,

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^\top \Sigma_s^{(m)} \mathbf{w} + \lambda_\Delta \bar{\sigma}_s^2 \sum_{\tau \in \mathcal{T}_s} \left(\sum_{i \in \tau} w_i \right)^2 + \lambda_H \bar{\sigma}_s^2 \sum_{h \in \mathcal{H}_s} \left(\sum_{j \in h} w_j \right)^2, \\ \text{s.t.} \quad & \sum_{i=1}^{N} w_i = 1, \\ & 0 \le w_i \le 1, \quad i = 1, \dots, N, \end{aligned} \tag{9}$$

where $\bar{\sigma}_s^2 = \mathrm{tr}(\Sigma_s^{(m)})/N$ scales the topological penalties to the magnitude of the variance term. If $\mathcal{T}_s$ and $\mathcal{H}_s$ are empty, the problem reduces to the standard minimum-variance portfolio. If Σ_s is replaced by a shrinkage covariance estimator, the model becomes a regularized covariance benchmark. If the expected return term is included as $\mathbf{w}^\top \Sigma_s \mathbf{w} - \gamma \boldsymbol{\mu}_s^\top \mathbf{w}$, the model becomes a classical mean-variance benchmark. These nested cases are used in the empirical comparison to separate the effects of covariance estimation, LSTM prediction, and tail-topological constraints.

Algorithm 1: Rolling prediction, topology construction and portfolio optimization

Data: Daily return matrix $\mathbf{R}$, sequence length L, rolling window M, rebalance frequency B, tail level α
Result: Out-of-sample portfolio returns and portfolio weights

```
Train the standard LSTM and VA-LSTM on the training sample using X_t and y_t from Equation 1;
Generate the full sequence of one-step-ahead return forecasts R̂;
for each rebalancing date s in the out-of-sample period do
    Select Z_s^(m) from historical returns or predicted returns according to the strategy;
    Estimate Σ_s^(m) from Equation 5;
    if the strategy uses tail topology then
        Compute left-tail or right-tail joint probabilities and distances using Equations 6 to 8;
        Select the graph threshold by the target edge density;
        Extract T_s and H_s from the threshold graph;
    end
    Solve Equation 9 and hold the portfolio until the next rebalancing date;
end
Evaluate all strategies on realized returns only;
```

3 Empirical Results

The empirical analysis uses daily data for the Dow Jones Industrial Average constituents. Prices are converted into simple daily returns before model estimation. The supervised prediction sample is constructed with a sequence length of 10 trading days, so that the model uses the previous 10 return vectors to predict the next return vector. The first 70% of the supervised sample is used for training. A 252-trading-day buffer is then reserved before the out-of-sample test period to ensure that the first portfolio decision has a complete rolling window for covariance and topology estimation. Portfolios are rebalanced every 21 trading days, and all out-of-sample returns are calculated using realized returns. Transaction costs are incorporated by subtracting 0.001 times portfolio turnover at each rebalancing date.

The baseline VA-LSTM specification uses two recurrent layers and a hidden dimension of 32. This hidden dimension is not tied to the number of assets. It represents the size of the latent temporal state used by the recurrent network. To address the sensitivity of this choice, hidden dimensions of 16, 32, and 64 are compared in Table 4. The LSTM output target is the vector of one-day-ahead stock returns. Volatility and tail-dependence quantities are not direct outputs of the neural network. They are estimated in the rolling portfolio stage from either realized returns or predicted returns, as described in the methodology section.

To strengthen the empirical comparison, the study evaluates several benchmark strategies. Table 1 summarizes the role of each strategy. The equal-weight strategy provides a simple allocation benchmark. Historical minimum variance and historical mean-variance represent classical covariance-based portfolio optimization. The shrinkage minimum-variance strategy uses a regularized covariance estimate. The inverse-volatility strategy is included as a simple risk-based allocation rule. The correlation-network strategy represents a pairwise network-constrained benchmark. The standard LSTM minimum-variance strategy is used to distinguish the contribution of the volatility-adaptive LSTM from a plain prediction-based allocation model. The proposed strategies combine VA-LSTM return forecasts with left-tail or right-tail topological penalties.

Strategy	**Rolling input for $\Sigma_s, \mathcal{T}_s, \mathcal{H}_s$**	**Purpose**
1/N Benchmark	Not estimated	Naive allocation benchmark
Hist Min Var	Historical returns	Classical minimum-variance benchmark
Hist Shrinkage Min Var	Historical returns with shrinkage covariance	Regularized covariance benchmark
Hist Inverse Volatility	Historical returns	Simple risk-based allocation benchmark
Hist Mean-Variance	Historical returns	Classical return-risk optimization benchmark
Hist Correlation-Network	Historical returns and correlation distance	Existing pairwise network-style benchmark
Standard LSTM Min Var	Standard LSTM return forecasts	Simpler prediction-based benchmark
VA-LSTM Min Var	VA-LSTM return forecasts	Prediction-based ablation without topology
Hist Left-Tail Topology	Historical returns and left-tail topology	Tail-topology ablation without LSTM forecasts
Hist Right-Tail Topology	Historical returns and right-tail topology	Tail-topology ablation without LSTM forecasts
Proposed VA-LSTM Left-Tail	VA-LSTM return forecasts and left-tail topology	Proposed crash-risk topology strategy
Proposed VA-LSTM Right-Tail	VA-LSTM return forecasts and right-tail topology	Proposed boom-state topology strategy

Table 1 Strategy definitions used in the empirical comparison

Table 2 reports the main out-of-sample performance indicators. This table is the primary evidence for portfolio performance and replaces the previous reliance on visual comparison only. It reports total return, annualized return, annualized volatility, Sharpe ratio, Sortino ratio, Calmar ratio, maximum drawdown, average turnover, and average Herfindahl concentration. These measures allow the strategies to be compared from the perspectives of return generation, risk exposure, downside risk, drawdown control, trading intensity, and positive days.

	Total Return	**Annual Return**	**Annual Volatility**	**Sharpe Ratio**	**Sortino Ratio**	**Calmar Ratio**	**Maximum Drawdown**	**Average Turnover**	**Positive Days**
1/N Benchmark	21.92%	18.83%	10.28%	1.833	2.605	3.389	-5.56%	**0.00%**	57.14%
Hist Min Var	17.68%	15.42%	8.88%	1.736	2.485	3.238	-4.763%	33.08%	56.78%
Hist Shrinkage Min Var	17.70%	15.43%	**8.84%**	1.747	2.503	3.242	**-4.761%**	27.56%	55.31%
Hist Inverse Volatility	22.85%	19.46%	9.59%	2.031	2.944	3.768	-5.17%	4.06%	56.78%
Hist Mean-Variance	59.23%	**47.24%**	29.23%	1.616	2.144	3.971	-11.90%	46.54%	57.88%
Hist Correlation-Network	17.27%	15.26%	10.47%	1.457	2.282	2.587	-5.90%	57.22%	56.41%
Standard LSTM Min Var	16.87%	16.38%	19.95%	0.821	1.125	1.830	-8.95%	21.45%	53.48%
VA-LSTM Min Var	24.30%	21.26%	15.35%	1.385	1.759	1.981	-10.74%	32.40%	53.85%
Hist Left-Tail Topology	22.16%	19.17%	11.79%	1.627	2.370	3.153	-6.08%	86.16%	56.41%
Hist Right-Tail Topology	2.65%	3.50%	14.68%	0.238	0.281	0.281	-12.45%	78.01%	52.38%
Proposed VA-LSTM Left-Tail	**61.43%**	45.75%	17.50%	**2.614**	**4.506**	**6.647**	-6.88%	56.02%	**58.61%**
Proposed VA-LSTM Right-Tail	39.04%	31.66%	15.65%	2.023	3.258	2.912	-10.87%	67.36%	53.11%

Table 2 Out-of-sample portfolio performance net of transaction costs

Figure 1 shows the trajectories of various strategies in the out-of-sample period. Its role is to show the timing of gains and drawdowns, while the formal comparison is based on the performance table. The predictive strategies are evaluated on realized returns, so the figure and the table reflect genuine out-of-sample portfolio performance. The comparison between the standard LSTM minimum-variance strategy and the VA-LSTM minimum-variance strategy isolates the effect of the volatility-adaptive gate. The comparison between the VA-LSTM minimum variance and the proposed VA-LSTM tail-topology strategies isolates the incremental effect of the topological penalties.

To examine whether the return differences are only visually observed, Table 3 reports paired tests of daily returns against the equal-weight benchmark. The paired t-test evaluates the mean difference in daily returns, while the Wilcoxon signed-rank test provides a nonparametric test that is less dependent on normality assumptions. These tests provide additional evidence on whether the daily return differences.

Table 4 reports a sensitivity check for the hidden dimension of the VA-LSTM. The hidden dimension determines the size of the latent temporal representation learned by the recurrent model and is therefore conceptually different from the number of assets in the portfolio universe. The values 16, 32, and 64 are examined to evaluate whether the empirical results are sensitive to this modelling choice. The baseline value of 32 is used as an intermediate

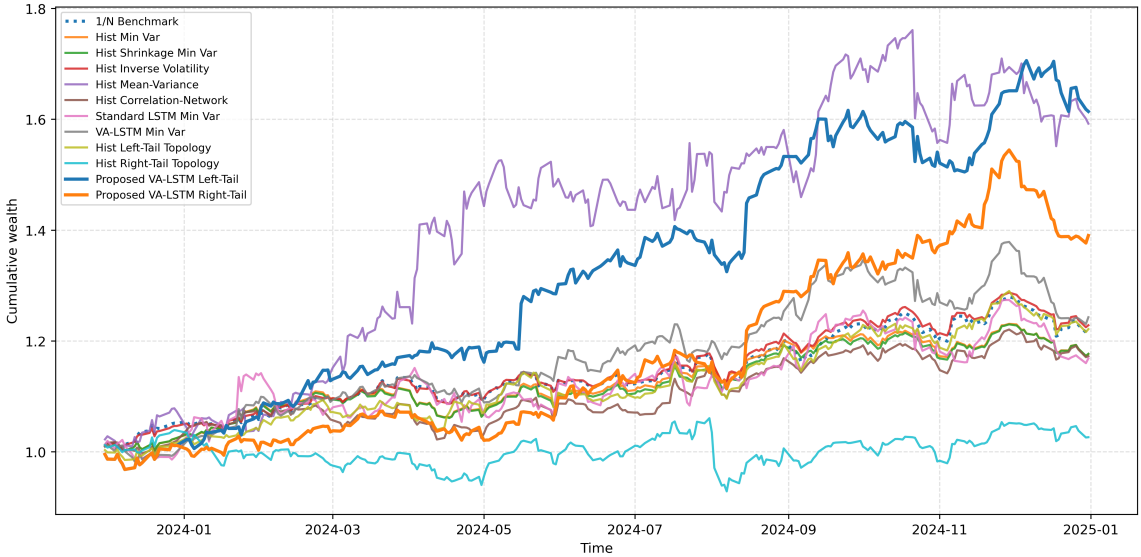


Figure 1 Out-of-sample cumulative wealth of benchmark and proposed strategies

specification, while the additional results help show whether the conclusions depend on a particular hidden-state size.

Strategy	Mean Daily Difference	Paired t p-value	Wilcoxon p-value
Hist Min Var	-0.0001	0.541	0.474
Hist Shrinkage Min Var	-0.0001	0.506	0.468
Hist Inverse Volatility	0.0000	0.668	0.858
Hist Mean-Variance	0.0011	0.256	0.059
Hist Correlation-Network	-0.0001	0.585	0.235
Standard LSTM Min Var	-0.0001	0.897	0.948
VA-LSTM Min Var	0.0001	0.868	0.583
Hist Left-Tail Topology	0.0000	0.967	0.855
Hist Right-Tail Topology	-0.0006	0.079	0.122
Proposed VA-LSTM Left-Tail	0.0011	0.107	0.117
Proposed VA-LSTM Right-Tail	0.0005	0.374	0.289

Table 3 Paired tests of daily returns against the equal-weight benchmark

Hidden Dimension	Forecast RMSE	Annual Return	Annual Volatility	Sharpe Ratio	Maximum Drawdown
16	0.0177	32.88%	16.64%	1.976	-10.72%
32	0.0211	45.75%	17.50%	2.614	-6.88%
64	0.0190	52.20%	17.11%	3.051	-7.56%

Table 4 Robustness check for the VA-LSTM hidden dimension

Table 5 reports the sensitivity of the proposed left-tail strategy to the topological penalty parameter. This check clarifies the role of the penalty term in Equation 9. When the penalty is zero, the strategy reduces to the prediction-based minimum-variance allocation. Increasing the penalty places stronger restrictions on weight concentration within tail-dependent topological structures.

The benchmark set shows whether the proposed strategy adds value beyond classical risk minimization, shrinkage covariance estimation, simple risk-based allocation, a pairwise network benchmark, and a standard LSTM forecast. The robustness checks further clarify the impact of hidden dimension selection and topological penalty calibration.

Lambda	Annual Return	Annual Volatility	Sharpe Ratio	Maximum Drawdown	Average HHI
0.0	21.26%	15.35%	1.385	-10.74%	0.6109
0.5	45.69%	17.45%	2.619	-6.85%	0.8964
1.0	45.75%	17.50%	2.614	-6.88%	0.9032
2.0	45.79%	17.53%	2.612	-6.90%	0.9068

Table 5 Sensitivity of the proposed left-tail strategy to the topology penalty parameter

4 Conclusion

This study develops a dynamic portfolio optimization framework that combines one-step-ahead return forecasting with tail-dependent topological regularization. The proposed VA-LSTM is used to generate forecasts of stock returns. At the same time, the covariance matrix and the tail-topological sets are estimated during the rolling portfolio stage from either historical returns or predicted returns. By incorporating triangular and hole structures extracted from left-tail and right-tail dependence networks, the model extends covariance-based allocation with higher-order information about asset co-movement under extreme market conditions.

The empirical analysis compares the proposed strategy with various strategies. The framework is most useful when tail co-movement and allocation concentration are important concerns, because the topological penalties explicitly restrict portfolio exposure to higher-order structures extracted from extreme-event networks. The findings suggest that prediction-based information can support risk-adjusted allocation, while tail-topological penalties help control concentration and downside exposure when higher-order dependence structures are relevant. Future research may extend the framework to other markets, alternative tail-dependence measures, and additional downside-risk objectives.

Acknowledgements

The research was supported by the Student Grant Competition of the Faculty of Economics, VSB - Technical University of Ostrava, project no. SP2026/003.

References

Freitas FD, De Souza AF, de Almeida AR (2009) Prediction-based portfolio optimization model using neural networks. Neurocomputing 72(10):2155–2170. https://doi.org/10.1016/j.neucom.2008.08.019

Freitas WB, Bertini JR (2023) Random walk through a stock network and predictive analysis for portfolio optimization. Expert Systems with Applications 218:119597. https://doi.org/10.1016/j.eswa.2023.119597

Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Li Y, Jiang XF, Tian Y, et al (2019) Portfolio optimization based on network topology. Physica A: Statistical Mechanics and its Applications 515:671–681. https://doi.org/10.1016/j.physa.2018.10.014

Ma Y, Wang W, Ma Q (2023) A novel prediction based portfolio optimization model using deep learning. Computers & Industrial Engineering 177:109023. https://doi.org/10.1016/j.cie.2023.109023

Martínez-Barbero X, Cervelló-Royo R, Ribal J (2024) Portfolio Optimization with Prediction-Based Return Using Long Short-Term Memory Neural Networks: Testing on Upward and Downward European Markets. Computational Economics 65(3):1479–1504. https://doi.org/10.1007/s10614-024-10604-6

Ricca F, Scozzari A (2024) Portfolio optimization through a network approach: Network assortative mixing and portfolio diversification. European Journal of Operational Research 312(2):700–717. https://doi.org/10.1016/j.ejor.2023.07.010

Wei Y, Watada J, Wang Z (2025) Topology Unveiled: A New Horizon for Economic and Financial Modeling. Mathematics 13(2). https://doi.org/10.3390/math13020325

Wen F, Yang X, Zhou WX (2019) Tail dependence networks of global stock markets. International Journal of Finance & Economics 24(1):558–567. https://doi.org/10.1002/ijfe.1679

Early Warning of Sovereign Credit Risk in BRIC: A Comparative GARCH-Based Analysis

Shikai Gao[1]

Abstract. This paper examines the volatility characteristics and early warning signals of sovereign credit risk in Brazil, Russia, India, and China (BRIC) using the GARCH model. This paper uses differences in 10-year sovereign bond yields relative to U.S. 10-year Treasury yields as a proxy for market-based sovereign credit risk and estimates country-level GARCH (1,1) models for BRIC. The study focuses on transforming GARCH conditional variance into risk warning signals based on quantile thresholds, thereby constructing a simple, transparent, and interpretable framework for monitoring sovereign credit risk. Empirical results indicate that changes in sovereign credit spreads across BRIC countries exhibit distinct characteristics of conditional volatility and cross-country heterogeneity. Further quantile-based early warning results also identify the months of extreme volatility in each country's financial markets. Overall, GARCH-based conditional volatility provides a straightforward early-warning tool for assessing sovereign credit risk in emerging economies. However, its interpretation should take into account cross-country heterogeneity and the limitations inherent in volatility-based modeling.

Keywords: Sovereign credit risk; Sovereign yield spreads; GARCH conditional volatility; BRIC economies; Quantile-based early warning

JEL Classification: C58, G15, G17, F34, H63

1 Introduction and literature review

Sovereign credit risk is a central issue in international finance and macrofinancial stability, as it reflects not only a government's ability to service its debt but also the broad interactions among fiscal space, external financing conditions, capital flows, and global investor sentiment. Edwards (1983) found in early research that sovereign risk premia in developing countries were closely associated with external debt burdens, debt-servicing capacity, and international reserves. Subsequent studies have further confirmed that macroeconomic fundamentals and their volatility largely explain the pricing of sovereign credit risk. For example, Hilscher and Nosbusch (2010) emphasized the role of macroeconomic fundamentals and fluctuations in terms of trade in explaining sovereign credit spreads in emerging markets , while Baldacci et al. (2011) demonstrated that fiscal and political risks are also important determinants of sovereign credit spreads in emerging economies. These studies suggest that sovereign credit risk should not be viewed as the result of a single fiscal variable, but rather as the combined result of domestic fundamentals, market expectations, institutional conditions, and the global financial environment.

For emerging markets, sovereign credit risk is particularly sensitive to external shocks and shifts in global risk appetite. Longstaff et al. (2011) used sovereign credit default swap (CDS) data to find that a significant portion of sovereign credit risk is correlated with global factors, suggesting that sovereign risk is not determined solely by country-specific factors. Aizenman et al. (2016) further demonstrate that the determinants of sovereign risk in emerging markets differ across the pre-crisis, crisis, and post-crisis phases, implying that sovereign risk pricing is state-dependent and varies with changes in the macro-financial environment. Similarly, Comelli (2012) found that the response of emerging market sovereign bond spreads to domestic and global factors varies across countries and changes over time. Recent evidence also points to nonlinear and crisis-sensitive dynamics. Cayirli et al. (2022) documented the state-dependent and asymmetric responses of emerging market sovereign spreads to global risk appetite and liquidity conditions. Daehler et al. (2021) noted that during the COVID-19 shock, traditional global and regional risk factors were insufficient to fully explain sovereign CDS spreads, while fiscal space, dependence on oil revenues, and pandemic-related information became more important. These findings suggest that sovereign credit risk in emerging markets tends to manifest not only in changes in credit spreads but also in abnormal volatility during periods of global financial stress.

[1] VSB - Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 70800 Ostrava-Poruba,shikai.gao@vsb.cz.

This paper focuses on the BRIC countries because these nations share the common characteristics of major emerging markets while also exhibiting significant cross-country differences. Brazil, Russia, India, and China differ markedly in terms of economic scale, dependence on commodities and energy, exchange rate regimes, financial openness, foreign exchange reserves, and exposure to geopolitical risks. Therefore, monitoring sovereign credit risk solely at the aggregate level of emerging markets may overlook the unique volatility patterns of individual countries. Within this context, sovereign bond spreads and sovereign credit default swap (CDS) spreads are widely used as market-based indicators of sovereign credit risk. Although CDS spreads are frequently employed in the literature, data availability for some BRIC countries is relatively limited across the entire sample period. Consequently, this study employs the spread between each country's 10-year sovereign bond yield and the U.S. 10-year Treasury yield as a consistent market proxy for measuring sovereign credit risk in BRIC countries. This finding is consistent with previous studies, which have all used the spread on sovereign bonds to capture the market's repricing of sovereign default and macro-financial risks. To ensure the comparability of GARCH estimates across countries, this analysis is limited to the BRIC countries. South Africa was excluded from the baseline empirical analysis because preliminary GARCH estimates indicated that its volatility dynamics exhibited boundary-type characteristics. Additionally, the diagnostic performance of the residuals was poor, which would make it difficult to interpret its conditional variance path within the same comparable-country GARCH framework. Therefore, the revised analysis focuses on Brazil, Russia, India, and China.

From a methodological perspective, this paper draws on the ARCH/GARCH literature, which provides a standard framework for modeling time-varying volatility and volatility clustering in financial time series. Engle (1982) proposed the ARCH model to capture conditional heteroscedasticity, while Bollerslev (1986) extended it into the GARCH framework by allowing current conditional variance to depend on both past shocks and past conditional variance. This modeling approach is particularly well-suited for sovereign risk analysis, as emerging market spreads tend to exhibit alternating periods of calm and sudden volatility. Recent research on sovereign risk has increasingly focused on volatility itself, not just the level of credit spreads. De Ferra and Mallucci (2022) highlight the time-varying volatility of interest rate spreads in the context of endogenous sovereign default risk, while Boeck et al. (2024) argue that the volatility of sovereign credit default swaps (CDS) can serve as a measure of economic uncertainty. Following this line of research, this paper does not aim to propose a new volatility model or directly estimate default probabilities.

2 Methodology

This section defines proxy variables for sovereign risk, presents a GARCH(1,1) model for a specific country, introduces diagnostic tests, and proposes an early warning rule based on quantiles.

2.1 Variable Definition

This paper focuses on the four BRIC countries: Brazil, Russia, India, and China. The key variable is the yield spread between each country's 10-year sovereign bond yield and the yield on 10-year U.S. Treasury bonds:

$$Spread_{i,t} = Yield10Y_{i,t} - US10Y_t \tag{1}$$

Where $Yield10Y_{i,t}$ denotes the yield on country i's 10-year sovereign bond at time t, and $US10Y_t$ denotes the yield on the U.S. 10-year Treasury bond over the same period. Sovereign bond spreads are typically viewed as the risk premium demanded by the market and can therefore serve as an empirical proxy for sovereign credit risk pressures (Edwards, 1983; Baldacci et al., 2011). Since this paper focuses on the marginal changes and volatility characteristics of risk pressure, the model uses the first-difference of sovereign yield spreads:

$$\Delta Spread_{i,t} = Spread_{i,t} - Spread_{i,t-1} \tag{2}$$

The sample consists of monthly data covering the period from January 2008 to March 2026. This timeframe encompasses the global financial crisis, the post-European debt crisis period, the impact of the Covid-19, and specific risk events in certain countries, making it suitable for examining the concentration of crises in the volatility of the main credit spread.

2.2 GARCH(1,1) model

The GARCH(1,1) model is used to capture the conditional volatility of the spread on the sovereign bond. This model is suitable for financial time series exhibiting volatility clustering, as it allows the conditional variance to be determined jointly by past shocks and past conditional variances. The mean equation is set as follows:

$$\Delta Spread_{i,t} = u_i + \phi_i \Delta Spread_{i,t-1} + \varepsilon_{i,t} \tag{3}$$

The innovation term is linked to the conditional variance as follows:

$$\varepsilon_{i,t} = \sqrt{h_{i,t}} z_{i,t} \tag{4}$$

This links the innovation term to conditional variance, where $\sqrt{h_{i,t}}$ represents conditional volatility and $z_{i,t}$ denotes the standardized innovation. Assume that $z_{i,t}$ follows a standardized Student-t distribution with zero mean and unit variance. This distribution assumption is adopted because changes in sovereign bond yield spreads during a crisis may exhibit heavy-tailed distributions and extreme observations. The shape parameter of the Student-*t* distribution is estimated jointly with the GARCH parameters.

The conditional variance equation is:

$$h_{i,t} = w_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i h_{i,t-1} \tag{5}$$

Here, $h_{i,t}$ represents the conditional variance of country i at time t, α_i measures the impact of short-term shocks on volatility, and β_i captures the effect of lagged conditional variance. Volatility persistence is assessed by $\alpha_i + \beta_i$, not by β_i alone. If $\alpha_i + \beta_i < 1$, the conditional variance process exhibits mean-reversion properties. If $\alpha_i + \beta_i$ is close to 1, volatility shocks decay slowly, and sovereign risk volatility exhibits high persistence. If this value approaches the stationarity boundary, one should exercise caution when applying standard GARCH model interpretations.

2.3 Model Diagnostics

To assess the suitability of the GARCH(1,1) model specification, this paper examines the standardized innovation terms obtained from country-specific models:

$$\hat{z}_{i,t} = \frac{\hat{\varepsilon}_{i,t}}{\sqrt{\hat{h}_{i,t}}} \tag{6}$$

where $\hat{z}_{i,t}$ denotes the standardized innovation. $\hat{\varepsilon}_{i,t}$ is the estimated innovation term. $\hat{h}_{i,t}$ is the estimated conditional variance. If the GARCH(1,1) model is appropriately specified, the standardized residuals should not exhibit significant serial correlation or conditional heteroscedasticity.

Three diagnostic checks are therefore applied. First, we perform a Ljung–Box test on $\hat{z}_{i,t}$ to check whether serial correlation still exists in the standardized innovation term. Second, we perform a Ljung–Box test on $\hat{z}_{i,t}^2$ to test whether volatility dependence persists after GARCH filtering. Third, the ARCH-LM test is employed to examine whether ARCH effects persist in the standardized innovation terms. The results of these diagnostic tests are reported in the empirical results section and are used to assess whether country-specific GARCH models are suitable for volatility-based early warning analysis.

2.4 Quantile-Based Early Warning Rules

To convert conditional variance into a risk monitoring indicator, this paper constructs the following early warning rules:

$$Warning_{i,t}^{q} = 1 \; if \; h_{i,t} > Q_q(h_i) \tag{7}$$

where, $Q_q(h_i)$ denotes the qth quantile of the conditional variance sequence for country i. This paper uses three thresholds—90%, 95%, and 99%—corresponding to general, high, and extreme alerts, respectively. It is important to note that these thresholds are calculated within each country's own conditional variance distribution. Therefore, the number of alert months will naturally be similar among countries with the same sample size. This paper focuses on the timing of the alert signals and the concentration of crises, rather than simply comparing the number of alert months across countries.

3 Empirical Results and Early-Warning Evidence

This section presents the estimated parameters of the GARCH(1,1) model for the BRIC countries, the results of diagnostic tests, conditional volatility paths, and quantile-based early-warning signals for the BRIC countries.

Table 1 reports the GARCH(1,1) estimation results for the time series of the sovereign spread differentials across the BRIC countries.

Country	w_i	α_i	β_i	$\alpha_i + \beta_i$	Descriptive explanation
Brazil	0.103	0.136	0.641	0.777	Below the high persistence threshold
Russia	0.0408	0.352	0.633	0.985	High volatility and persistence
India	0.0137	0.221	0.646	0.867	Below the high persistence threshold
China	0.0202	0.235	0.481	0.716	Below the high persistence threshold

Table 1 Parameters and volatility persistence of BRIC

The results show that Russia exhibits the highest volatility persistence, with a value of 0.985, indicating that shocks to its sovereign spread tend to persist. The values for India, Brazil, and China are 0.867, 0.777, and 0.716, respectively—all lower than Russia's—suggesting that volatility in these countries tends to revert to the mean to a greater extent.

Country	Observations	Ljung-Box residuals *p*-value	Ljung-Box squared residuals *p*-value	ARCH-LM *p*-value	Conclusion
Brazil	218	0.889	0.747	0.993	No remaining dependence
China	218	0.948	0.852	0.857	No remaining dependence
India	218	0.380	0.771	0.722	No remaining dependence
Russia	218	0.858	0.994	0.998	No remaining dependence

Table 2 Diagnostic Tests for Standardized Residuals

Note: Values are p-values. For the Ljung–Box tests, the null hypothesis is no serial correlation in the corresponding residual series. For the ARCH-LM test, the null hypothesis is no remaining ARCH effects.

To further assess the suitability of the estimated GARCH (1,1) model, diagnostic tests were conducted on the standardized residuals. Table 2 reports the results of the Ljung–Box test for the standardized residuals, the Ljung–Box test for the squared standardized residuals, and the ARCH-LM test. For all four BRIC countries, the p-values for these three diagnostic tests were higher than the conventional 5% significance level. These results indicate that there is no strong evidence of residual serial correlation, residual volatility dependence, or residual ARCH effects following GARCH filtering. Therefore, the estimated conditional variance can be used as a volatility-based early warning indicator for sovereign credit risk.

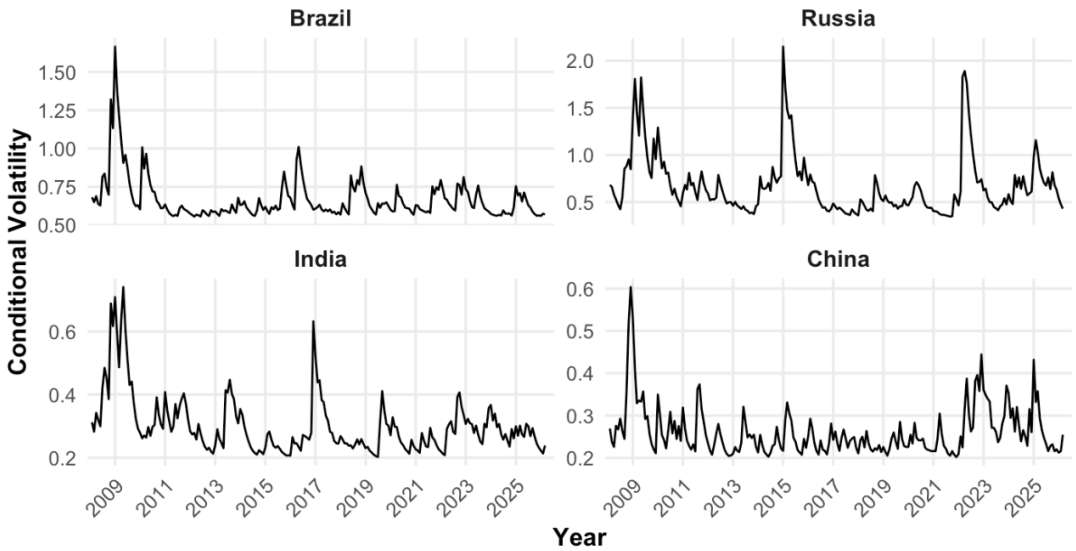


Figure 1 GARCH Conditional Volatility of Sovereign Yield Spreads Changes in BRIC Countries

Figure 1 shows the GARCH conditional volatility paths for each country. Brazil, Russia, India, and China all exhibit distinct volatility clustering patterns, meaning that periods of high volatility are concentrated in specific time periods. This is consistent with the conditional heteroskedasticity commonly observed in financial time series (Engle, 1982; Bollerslev, 1986). The periods of high volatility for Brazil, China, and India were primarily concentrated around the 2008–2009 global financial crisis. Russia exhibited distinct peaks in 2015 and 2022, indicating that its sovereign risk volatility is more characterized by country-specific shocks and geopolitical financial pressures.

Country	Months with a 99% extreme warning
Brazil	2008-11; 2009-01; 2009-02
Russia	2015-01; 2022-03; 2022-04
India	2008-11; 2009-01; 2009-05
China	2008-11; 2008-12; 2009-01

Table 3 Months with Extreme Warnings Based on the 99th Percentile

Table 3 lists the months in which conditional variance exceeded the 99th percentile threshold for each country. For Brazil, China, and India, these months of extreme warnings were concentrated primarily between 2008 and 2009, indicating that the GARCH conditional volatility measure was able to capture sovereign risk pressures during the global financial crisis. Russia's extreme warning months are concentrated in 2015 and 2022, consistent with its higher volatility persistence and spiked conditional volatility path. Early warning incidents in Russia can be linked to periods of particular strain in the country: the 2015 incident was consistent with the combined effects of falling oil prices, the depreciation of the rouble and financial pressures arising from sanctions, whilst the 2022 incident reflected a sharp adjustment in the pricing of Russia's sovereign risk following the escalation of geopolitical tensions and the imposition of related sanctions.

4 Conclusion

This paper constructs a volatility-based early warning framework for sovereign credit risk using a GARCH (1,1) model, based on changes in the yield spread between 10-year sovereign bonds from BRIC countries and 10-year U.S. Treasury bonds. The core aim of this paper is to examine whether GARCH-based conditional volatility can serve as a simple, transparent, and interpretable early warning signal for sovereign risk.

Empirical results indicate that changes in sovereign spreads across BRIC countries exhibit distinct characteristics of conditional volatility and cross-country heterogeneity. Russia exhibits high volatility persistence, indicating that shocks to its sovereign risk spread tend to be relatively long-lasting. The results for India, Brazil, and China suggest stronger mean-reverting characteristics in conditional volatility. Although these countries also experience temporary volatility spikes during periods of financial stress, these spikes are of shorter duration than those

observed in Russia. Diagnostic tests of the standardized residuals further indicate that the estimated GARCH (1,1) model does not exhibit significant residual serial correlation, residual volatility dependence, or residual ARCH effects, which supports the use of the estimated conditional variance as an early warning indicator.

Overall, the quantile-based early warning rule using conditional volatility estimated via the GARCH model provides a simple tool for monitoring sovereign risk pressures in emerging economies. Since the thresholds are calculated based on each country's own distribution of conditional volatility, the resulting signals are suitable for identifying periods of abnormal volatility within individual countries and comparing the timing of volatility events across countries, rather than ranking countries based on the absolute level of sovereign risk. Future research could further incorporate macrofinancial variables, global risk factors, multivariate GARCH models, or quantile-based connectivity methods to analyze the transmission mechanisms of sovereign credit risk more systematically.

Acknowledgements

The SGS Research Project SP2026/003 of VSB Technical University of Ostrava for financial support of this work.

References

Aizenman J, Jinjarak Y, Park D (2016) Fundamentals and sovereign risk of emerging markets. Pacific Economic Review 21(2): 151–177. https://doi.org/10.1111/1468-0106.12160

Baldacci E, Gupta S, Mati A (2011) Political and fiscal risk determinants of sovereign spreads in emerging markets. Review of Development Economics 15(2): 251–263. https://doi.org/10.1111/j.1467-9361.2011.00606.x

Boeck M, Feldkircher M, Raunig B (2024) A view from outside: Sovereign CDS volatility as an indicator of economic uncertainty. Macroeconomic Dynamics 28(7): 1423–1450. https://doi.org/10.1017/S1365100523000524

Bollerslev, T (1986) Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31(3): 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Cayirli O, Aktas H, Kayalidere K (2022) A closer look into the behavior of emerging market sovereign spreads: State-dependent and asymmetric behaviors. International Review of Economics & Finance 77: 522–548. https://doi.org/10.1016/j.iref.2021.10.017

Comelli F (2012) Emerging market sovereign bond spreads: Estimation and back-testing. Emerging Markets Review 13(4): 598–625. https://doi.org/10.1016/j.ememar.2012.09.002

Daehler T B, Aizenman J, Jinjarak Y (2021) Emerging markets sovereign CDS spreads during COVID-19: Economics versus epidemiology news. Economic Modelling 100: Article 105504. https://doi.org/10.1016/j.econmod.2021.105504

de Ferra S, Mallucci E (2022) Sovereign risk matters: Endogenous default risk and the time-varying volatility of interest rate spreads. Journal of International Economics 134: Article 103542. https://doi.org/10.1016/j.jinteco.2021.103542

Edwards S (1983) LDC's foreign borrowing and default risk: An empirical investigation (NBER Working Paper No. 1172). National Bureau of Economic Research. https://doi.org/10.3386/w1172

Engle R F (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50(4): 987–1007. https://doi.org/10.2307/1912773

Hilscher J, Nosbusch Y (2010) Determinants of sovereign risk: Macroeconomic fundamentals and the pricing of sovereign debt. Review of Finance 14(2): 235–262. https://doi.org/10.1093/rof/rfq005

Longstaff F A, Pan J, Pedersen L H, Singleton K J (2011) How sovereign is sovereign credit risk? American Economic Journal: Macroeconomics 3(2): 75–103. https://doi.org/10.1257/mac.3.2.75

Dynamic relationships in the green economy: renewable energy consumption, CO_2 emissions, real GDP and oil prices in European Union countries

Jana Hanclova[1], Lucie Klaubenschalk Chytilova[2], Jakub Illik[3]

Abstract. This paper examines short-run dynamic relationships among oil prices, real GDP per capita, CO_2 emissions per capita, and renewable energy consumption in 27 European Union countries over 1990–2024. A first-order panel vector autoregressive model (PVAR(1)) is estimated by the generalized method of moments (GMM), supplemented by panel Granger causality tests and orthogonalized impulse response functions. The results show that higher GDP per capita is followed by lower renewable energy consumption and higher CO_2 emissions in the short run. GDP and CO_2 emissions exhibit bidirectional feedback, whereas the GDP-to-renewable-energy relationship is only weakly significant in the causality test. Impulse responses are strongest initially and converge toward zero, confirming system stability. The findings indicate that economic growth does not automatically accelerate renewable-energy deployment and that policies supporting grids, storage, energy efficiency, and low-carbon investment remain essential for decoupling growth from emissions.

Keywords: PVAR, economic growth, CO_2 emissions, renewable energy, EU countries

JEL Classification: Q43, C33

1 Introduction

Oil-price shocks affect output, investment, energy demand, and environmental outcomes through production costs and substitution between energy sources. These links are especially relevant for the European Union (EU), given its dependence on imported fossil fuels, ambitious climate targets, and differing national energy structures. The period analysed also includes major oil-market and geopolitical shocks, including the COVID-19 pandemic and the Russia–Ukraine conflict (Hamilton, 2009; Kilian, 2009).

Higher oil prices may curb fossil-fuel demand and improve the competitiveness of renewable technologies, while lower prices may weaken incentives for decarbonisation (Apergis and Payne, 2014; Sadorsky, 2009). The direction and strength of these effects nevertheless depend on economic structure and policy.

This study analyses oil prices, real GDP per capita, CO_2 emissions per capita, and renewable energy consumption for 27 EU countries over 1990–2024. Unlike pairwise or static approaches, it models all four variables jointly and allows for endogenous feedback.

A Panel Vector Autoregression (PVAR) is appropriate because it treats the variables as endogenous and captures short-run dynamics and country heterogeneity. Impulse response functions (IRFs) then show the magnitude and persistence of shocks across the system (Love and Zicchino, 2006).

The study asks whether GDP growth affects renewable energy and CO_2 emissions, whether GDP and emissions exhibit bidirectional causality, and how shocks propagate through the stable PVAR system. Its contribution is the joint assessment of these relationships, including oil prices, in a broad EU panel.

Prior research often relies on pairwise or single-equation models and therefore cannot fully capture endogeneity, feedback, and short-run adjustment. Evidence jointly covering oil prices, emissions, renewable energy, and macroeconomic activity in EU countries remains limited.

[1] VSB-Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava Poruba, Czechia, jana.hanclova@vsb.cz.
[2]Czech University of Life Sciences Prague, Faculty of Economics and Management, Kamýcká 129, 165 00 Prague, Czechia, chytilova@pef.czu.cz, VSB-Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava Poruba, Czechia, lucie.klaubenschalk.chytilova@vsb.cz.
[3] VSB-Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava Poruba, Czechia, jakub.illik@vsb.cz.

The analysis tests three hypotheses: H1, GDP per capita has a negative short-run effect on renewable energy consumption; H2, GDP per capita positively affects CO_2 emissions in the short run; and H3, GDP per capita and CO_2 emissions exhibit a bidirectional relationship. Oil prices are included as a global source of variation affecting production costs, energy demand, and the relative attractiveness of renewable technologies.

The remainder of the paper is organised as follows. Section 2 reviews the theoretical and empirical literature. Section 3 presents the data and the PVAR methodology. Section 4 reports on the empirical results and discusses their implications. Section 5 concludes with policy recommendations and directions for future research.

2 Review of the Literature

PVAR models capture endogenous feedback, short-run adjustment, and cross-country heterogeneity in the energy–growth–environment nexus (Love and Zicchino, 2006). Empirical work generally finds that economic expansion raises emissions in the short run, especially where energy systems remain carbon-intensive (Apergis and Payne, 2014; Ajayi, 2024; Kazemzadeh et al., 2023).

Renewable-energy effects are more context-dependent. Short-run growth may initially raise demand for conventional energy because infrastructure adjusts slowly, while renewable deployment depends on policy support, finance, and technology (Sadorsky, 2009; Charfeddine and Kahia, 2019; Marra and Colantonio, 2021).

Oil prices add another transmission channel: higher prices may discourage fossil-fuel use and encourage efficiency or substitution toward renewables, but outcomes depend on the source of the shock and national economic structure (Kilian, 2009).

Existing multivariate studies typically focus on particular regions or omit at least one of oil prices, CO_2 emissions, and renewable energy (Charfeddine and Kahia, 2019; Kazemzadeh et al., 2023; Sepehrdoust et al., 2025). Evidence for the EU that places all four variables in one endogenous system is therefore limited.

3 Data and Methodology

The Panel Vector Autoregressive (PVAR) framework has emerged as a core econometric instrument for examining dynamic interdependencies.

3.1 Data description

Country-level data come from the World Bank Commodity Price Data (The Pink Sheet) and the World Development Indicators (World Bank, 2026a, 2026b). The balanced country set covers 27 EU members over 1990–2024: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, and Sweden.

Table 1 includes descriptive statistics for all variables used in the study, including the units of the variables:

- *Oil_price is Crude oil price (Oil_price)*, real in 2010 US dollar/bbl, World Bank Commodity Price Data (The Pink Sheet), CMO-Historical-Data-Annual;
- *GDPpc is real GDP per capita, constant 2015 USD (GDPpc),* dataset WB_WDI (World Bank – World Development Indicators), NY.GDP.PCAP.KD;
- *CO2pc represents CO_2 emissions per capita, (CO2pc)* total annual emissions of carbon dioxide (t CO2e/capita), WB_WDI, EN.GHG.CO2.PC.CE.AR5;
- *Rwenr denotes renewable energy consumption (Rwenr)* (% of total final energy consumption), dataset WB_WDI, EG.FEC.RNEW.ZS.

Variable	Obs	Mean	Std. dev.	Min	Max	Unit
Oil_price	945	51.766	23.909	17.463	96.818	US dollar/bbl
GDPpc	945	27549.7	20370.1	3537.4	112417.9	constant 2015 USD
CO2pc	945	8.080	3.959	3.004	32.194	t CO2e/capita
Rwenr	864	15.319	11.869	0.000	57.900	% of total final energy consumption

Table 1 Descriptive statistics of variables, source: own calculation in STATA

3.2 Methodology

Vector autoregression (VAR) models have been widely used to study oil price shocks and macroeconomic dynamics. Kilian (2009) demonstrates that distinguishing between different sources of oil price shocks is crucial for correctly identifying their effects on output and inflation. Panel VAR (PVAR) models combine the VAR structure with panel data techniques, allowing researchers to exploit both time-series dynamics and cross-sectional variation. Love and Zicchino (2006) provide a pioneering framework for estimating PVARs using GMM methods, while Abrigo and Love (2016) propose practical tools and implementation strategies widely adopted in applied research. Empirical PVAR studies on energy and the environment show that energy price shocks have persistent effects on emissions and renewable energy shares, but the magnitude and sign depend on economic structure and policy regimes (Apergis and Payne, 2014; Dong et al., 2024). To analyse the dynamic relationships between oil prices, economic growth, CO_2 emissions, and renewable energy consumption, this study employs a PVAR model. Let $i=1,\ldots,N$ denote countries and $t=1,\ldots,T$ time periods. The vector of endogenous variables is defined as $\boldsymbol{y}_{it}' = [\ln Oil_price_t \quad \ln GDPpc_{it} \quad \ln CO2pc_{it} \quad \ln Rwenr_{it}]$. The PVAR(p) model is specified as:

$$\boldsymbol{y}_{it} = \sum_{j=1}^{p} \boldsymbol{A}_j \boldsymbol{y}_{i,t-j} + \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{it} \tag{1}$$

where $\boldsymbol{A}_j$ are coefficient matrices, $\boldsymbol{\mu}_i$ capture unobserved country-specific fixed effects, $\boldsymbol{\varepsilon}_{it}$ is a vector of error terms.

Estimation strategy

The presence of lagged dependent variables and fixed effects leads to endogeneity, making standard estimators biased and inconsistent. To address this issue, the model is estimated using the Generalized Method of Moments (GMM) following the approach of Love and Zicchino (2006). First, fixed effects are removed using forward mean-differencing (Helmert transformation), which preserves orthogonality between transformed variables and lagged instruments. The transformed equation can be written as:

$$\tilde{\boldsymbol{y}}_{it} = \sum_{j=1}^{p} \boldsymbol{A}_j \tilde{\boldsymbol{y}}_{i,t-j} + \tilde{\boldsymbol{\varepsilon}}_{it} \tag{2}$$

Lagged values of endogenous variables are used as instruments to ensure identification. The optimal lag length p is determined based on information criteria adapted for panel VAR models (e.g., AIC, BIC), combined with economic considerations.

Pre-estimation diagnostics

Before estimating the PVAR, several diagnostic tests are conducted to ensure data adequacy: cross-sectional dependence and panel unit root tests. *Cross-sectional dependence* is tested using the Pesaran CD test:

$$CD = \sqrt{\frac{2T}{N(N-1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{ij} \tag{3}$$

where $\hat{\rho}_{ij}$ represents the pairwise correlation of residuals.

To ensure stationarity, *panel unit root tests* are conducted, including: Im–Pesaran–Shin (IPS), Fisher-type ADF, and Fisher-type Phillips–Perron tests. The IPS test is based on the regressiwhere $\hat{\rho}_{ij}$ represents the pairwise correlation of residuals. The null hypothesis assumes cross-sectional independence. The IPS test is based on the following regression:

$$\Delta \boldsymbol{y}_{it} = \boldsymbol{\alpha}_i + \boldsymbol{\beta}_i \boldsymbol{y}_{i,t-1} + \sum_{k=1}^{p} \boldsymbol{\gamma}_{ik} \Delta \boldsymbol{y}_{i,t-k} + \boldsymbol{\varepsilon}_{it} \tag{4}$$

where the null hypothesis $H_0 : \boldsymbol{\beta}_i = \mathbf{0}$ indicates all panels have a unit root. The alternative hypothesis is at least some panels are stationary. The IPS test allows *for heterogeneity across panel units*, meaning each country can have its own persistence. Fisher-type test based on ADF combines *p-values* from individual ADF tests across panel units and H_0: All panels contain a unit root and H_1: At least one panel is stationary. The Fisher-type test based on PP (Fisher–Perron) corrects for autocorrelation and heteroskedasticity non-parametrically and tests the hypotheses H_0: All series have unit roots and H_1: At least one series is stationary.

Stability analysis

A fundamental requirement for meaningful VAR interpretation is model stability. Stability ensures that impulse response functions can be economically interpreted and that dynamic interactions are not explosive. Model stability is evaluated by examining the *eigenvalues of the companion matrix*. The stability condition requires that $|\lambda_j| < 1 \quad \forall j$, where λ_j are the eigenvalues of the system. If all eigenvalues lie inside the unit circle, the PVAR system is stable and the impulse response functions are valid.

Impulse response functions (IRFs)

Impulse response functions trace the effect of a one-time shock on a variable on the future values of all variables in the system. The PVAR can be rewritten in its *moving average (MA) representation:*

$$\boldsymbol{y}_{it} = \sum_{k=0}^{\infty} \boldsymbol{\Phi}_k \varepsilon_{i,t-k} \tag{5}$$

where $\boldsymbol{\Phi}_k$ are impulse response matrices. Confidence intervals for IRFs are obtained using *bootstrap simulations.*

4 Empirical results

Cross-sectional Dependence test statistic tests whether residuals are correlated across panel countries. CD = 0.344 for lag(1) is close to zero. This implies low evidence of cross-sectional dependence, and therefore we do not reject the null hypothesis at the 5% level of significance. This result suggests that residuals of the PVAR(1) model are not significantly correlated across units; our model does not miss strong common shocks affecting all units, and standard PVAR(1) assumptions (independence across cross-sections) are reasonable. *Panel unit root tests* examine whether variables in panel data are stationary (stable over time), or non-stationary (contain a unit root). For the panel unit root tests, we used IPS, Fisher–ADF, and Fisher–PP tests. Table 2 clearly summarises the results of these tests for logarithmic transformations of the original variables (ln*variable*) or the first differences of these logarithmic variables (d_ln*variable).* We can conclude that all variables except *lnRwenr* are I(1) at the 5% significance level. Therefore, we will use differences of logarithmic variables for the PVAR model.

Variable	ADF	Fisher–ADF	Fisher–PP	variable	ADF	Fisher–ADF	Fisher–PP
lnOil_price	N	N	N	*d_lnOil_price*	S	S	S
lnGDPpc	N	N	N	*d_lnGDPpc*	S	S	S
lnCO2pc	N	N	S	*d_lnCO2pc*	S	S	S
lnRwenr	S	S	N	*d_lnRwenr*	S	S	S

Table 2 Results of panel unit root tests, source: own calculation in STATA. N = non-stationary panel data, S = stationary panel data

4.1 PVAR(1) Results

The PVAR model is estimated using the GMM system following Love and Zicchino (2006). Country fixed effects are removed via forward mean differentiating.

	d_lnOil_price	***d_lnGDPpc***	***d_lnCO2pc***	***d_lnRwenr***
d_lnOil_price(-1)	0.093**	0.002	-0.014	0.220
d_lnGDPpc (-1)	-0.366	0.285***	0.254**	-4.158*
d_lnCO2pc (-1)	0.539***	0.149***	0.038	-3.167*
d_lnRwenr (-1)	-0.0021	-0.0004	-0.0004	-0.151**

Table 3 GMM estimation of the PVAR(1) model in STATA
*Note: *** p < 0.01, ** p < 0.05, * p < 0.10*

In addition to country-specific fixed effects, time effects were initially considered to account for common shocks affecting all countries simultaneously. This issue is particularly relevant because the oil price variable is identical

across countries in a given year and varies only over time, reflecting developments in the global oil market. Therefore, year dummy variables were included in the initial specification to capture common time-specific influences beyond those represented directly by oil prices. However, the estimated coefficients associated with the time effects were not statistically significant. Consequently, year dummies were excluded from the final specification in order to preserve model parsimony without sacrificing explanatory power. The estimated PVAR(1) model in Table 3 provides insights into the short-run dynamic relationships among economic growth, CO_2 emissions, oil prices, and renewable energy consumption. The interpretation focuses primarily on statistically significant coefficients and established research hypotheses.

H1 receives qualified support: lagged GDP per capita has a negative, weakly significant effect on renewable energy consumption (−4.158*). In the short run, additional energy demand may therefore be met mainly by existing fossil-fuel infrastructure; renewable capacity responds more slowly and depends on policy, finance, and technological investment (Charfeddine and Kahia, 2019; Marra and Colantonio, 2021).

H2 is supported: lagged GDP per capita positively affects CO_2 emissions (0.254**), indicating that short-run growth remains linked to carbon-intensive production and energy use (Sepehrdoust et al., 2025).

H3 is also supported. Lagged CO_2 emissions positively affect GDP per capita (0.149***), while GDP raises emissions, revealing bidirectional short-run feedback between economic activity and carbon-intensive energy use.

GDP per capita is persistent (0.285***). Overall, the estimates suggest that short-run growth remains coupled with emissions and does not automatically expand renewable energy, underscoring the role of targeted transition policies.

4.2 Stability of the PVAR(1) System

Figure 1 and Table 4 show that all companion-matrix eigenvalues lie inside the unit circle. The PVAR(1) system is therefore dynamically stable, its impulse responses are well-defined, and shocks converge rather than becoming explosive.

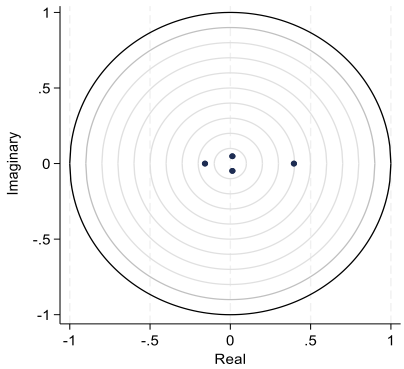


Figure 1 Roots of the companion matrix

Eigenvalue (Real)	Eigenvalue (Imaginary)	Modulus
0.396	0.000	0.396
-0.158	0.000	0.158
0.013	-0.048	0.050
0.013	0.048	0.050

Table 4 Analysis of the Eigenvalue Stability of the PVAR(1) Model

4.3 Granger Causality Analysis

Panel Granger causality tests confirm bidirectional feedback between GDP and CO_2 emissions: GDP Granger-causes emissions (χ^2 = 5.077**) and emissions Granger-cause GDP (χ^2 = 17.727***), supporting H2 and H3.

GDP Granger-causes renewable energy consumption only at the 10% level (χ^2 = 3.720*), so H1 receives weaker support here than in the PVAR estimate. The finding that CO_2 emissions Granger-cause oil prices (χ^2 = 7.782***) should be interpreted cautiously because oil prices are global and common across countries.

Excluded	*Equation d_lnOil_price*	*Equation d_lnGDPpc*	*Equation d_lnCO2pc*	*Equation d_lnRwenr*
d_lnOil_price	-	0.087	1.531	0.774
d_lnGDPpc	1.415	-	5.077**	3.720*
d_lnCO2pc	7.782***	17.727***	-	3.095*
d_lnRwenr	0.095	0.106	0.023	-
all	9.714**	18.926***	5.886	25.572***

Table 5 Panel Granger causality (Chi-square statistics), source: own calculation in STATA
*Note: *** p < 0.01, ** p < 0.05, * p < 0.10*

4.4 Impulse Response Function Analysis

The orthogonalized IRFs are consistent with the coefficient estimates. A positive GDP shock initially reduces renewable energy consumption (Figure 2) and increases CO_2 emissions (Figure 3), while an emissions shock temporarily raises GDP (Figure 4). All responses gradually converge toward zero, consistent with system stability.

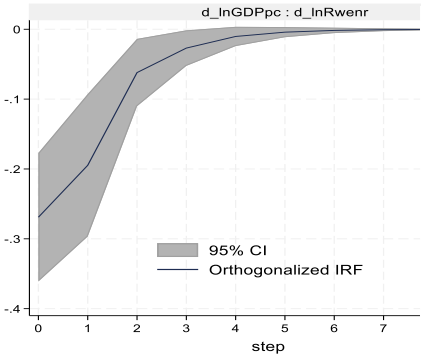


Figure 2 Response of *d_lnRwenr* to an impulse in *d_lnGDPpc*

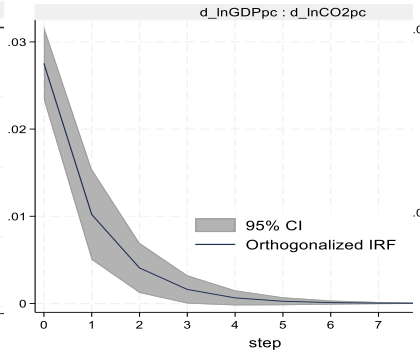


Figure 3 Response of *d_lnCO2pc* to an impulse in *d_lnGDPpc*

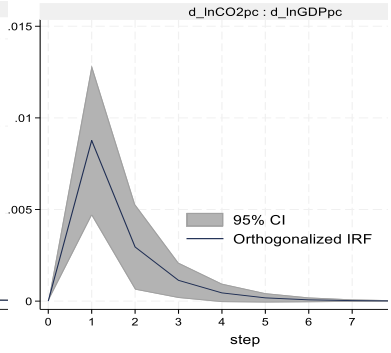


Figure 4 Response of *d_lnGDPpc* to an impulse in *d_lnCO2pc*

5 Discussion and Conclusions

Across the PVAR estimates, causality tests, and IRFs, the most robust result is bidirectional GDP–emissions feedback. The negative GDP–renewable relationship is less conclusive because its Granger-causality statistic is only marginally significant.

For the EU, the results distinguish long-run transition goals from short-run adjustment. Energy demand can rise faster than renewable generation, grids, and storage capacity; consequently, existing fossil-fuel infrastructure may absorb the initial increase in demand.

The study contributes by modelling GDP, CO_2 emissions, renewable energy, and oil prices simultaneously for 27 EU countries and by showing that the estimated shock effects are temporary. The findings complement earlier regional PVAR evidence while highlighting short-run adjustment asymmetry in the EU.

Policy should accelerate the short-run response of low-carbon supply through investment in grids and storage, faster renewable deployment, energy efficiency, industrial electrification, and effective carbon-pricing incentives. These measures can prevent additional demand from being met by fossil fuels.

In conclusion, EU growth remains associated with higher short-run CO_2 emissions, whereas renewable energy does not automatically expand with economic activity. Future research should model differences among member states, include climate-policy indicators, and distinguish among sources of oil-price shocks.

Acknowledgements

This article has been produced with the financial support of the European Union under REFRESH - Research Excellence For REgion Sustainability and High-tech Industries project number CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition and by the SGS project (SP2026/071) of the Faculty of Economics, VSB-Technical University of Ostrava.

References

Ajayi, T A (2024) Oil price, energy consumption and CO2 emissions as growth determinants: A PVAR system GMMapproach. International Journal of Energy Sector Management 18(6): 2086–2114. https://doi.org/10.1108/IJESM-09-2023-0013

Apergis, N, Payne J E (2014) Renewable energy, output, CO_2 emissions, and fossil fuel prices in Central America: Evidence from a nonlinear panel smooth transition vector error correction model. Energy Economics 42: 226-232. http://dx.doi.org/10.1016/j.eneco.2014.01.003

Aslan A, Ocal O, Ozsolak B, Ozturk I (2022) Renewable energy and economic growth relationship under the oil reserve ownership: Evidence from panel VAR approach. Renewable Energy 188: 402–410. https://doi.org/10.1016/j.renene.2022.02.039

Charfeddine L, Kahia M (2019) Impact of renewable energy consumption and financial development on CO2 emissions and economic growth in the mena region: A panel vector autoregressive (pvar) analysis. Renewable Energy 139: 198–213. https://doi.org/10.1016/j.renene.2019.01.010

Dong Q, Du Q, Min Du A (2024) Interplay between oil prices, country risks, and stock returns in the context of global conflict: A PVAR approach. Research in International Business and Finance 72: 102545. https://doi.org/10.1016/j.ribaf.2024.102545

Hamilton J D (2009). Causes and consequences of the oil shock of 2007–08. Brookings papers on economic activity 1: 215–283. https://doi.org/10.1353/eca.0.0047

Kazemzadeh E, Fuinhas J A, Koengkan M, Shadmehri M T A (2023) Relationship between the share of renewable electricity consumption, economic complexity, financial development, and oil prices: A two-step club convergence and PVAR model approach. International Economics 173: 260–275. https://doi.org/10.1016/j.inteco.2022.12.001

Kilian L (2009). Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. American Economic Review 99(3): 1053–1069. https://doi.org/10.1257/aer.99.3.1053.

Love I, Zicchino L (2006) Financial development and dynamic investment behavior: Evidence from panel VAR. Quarterly Review of Economics and Finance 46(2): 190–210. https://doi.org/10.1016/j.qref.2005.11.007

Marra A, Colantonio E (2021) The path to renewable energy consumption in the european union through drivers and barriers: A panel vector autoregressive approach. Socio-Economic Planning Sciences 76: 100958. https://doi.org/10.1016/j.seps.2020.100958

Omojolaibi J A, Egwaikhide F O (2014) Oil price volatility, fiscal policy and economic growth: A panel vector autoregressive (PVAR) analysis of some selected oil-exporting A frican countries. OPEC Energy Review 38(2): 127–148. https://doi.org/10.1111/opec.12018

Sadorsky P (2009). Renewable energy consumption and income in emerging economies. Energy Policy 37(10), 4021–4028. https://doi.org/10.1016/j.enpol.2009.05.003

Sepehrdoust H, Tartar M, Parvin S H (2025) Environmental impacts of macroeconomic variables: A case study of Iran and selected OPEC members. Chinese Journal of Population, Resources and Environment 23(2): 236–245. https://doi.org/10.1016/j.cjpre.2025.05.00

World Bank (2026a) Commodity Price Data (The Pink Sheet). Available at: https://www.worldbank.org/en/research/commodity-markets (accessed 30 July 2026)

World Bank (2026b) World Development Indicators. Available at: https://databank.worldbank.org/source/world-development-indicators (accessed 30 July 2026)

Structural Factors Influencing AI Copyright Literacy: A Quantitative Analysis from an Academic Survey Study

Jana Heckenbergerová[1], Barbora Kovářová[2]

Abstract. Artificial intelligence systems are increasingly integrated into academic work. Legal boundaries concerning copyright and AI use are still evolving within the EU and Czech legal context. This situation increases the risk of unintentional copyright violations.
The questionnaire survey, conducted primarily within the academic environment of the University of Pardubice, collected information on participants' AI literacy, their awareness of risk factors, and their knowledge of internal AI regulations.
The main objective is to quantify the effects of selected determinants on AI related behavior and compliance. Dependent variables include AI use intensity, copyright literacy scores, and compliance and disclosure indicators. Independent variables cover demographic attributes and structural characteristics of respondents.
The findings indicate that copyright-related AI literacy in the academic environment is not significantly influenced by respondents' individual characteristics, but rather by structural factors, especially disciplinary orientation. Policy measures should therefore focus on discipline specific guidelines and institutional frameworks.

Keywords: AI Copyright Literacy, AI use intensity, Quantitative Survey Methods, Non-parametric Inference Analysis

JEL Classification: C12, C83, I23, K24

1 Legal and Research Context of AI Use in Higher Education

Artificial intelligence (AI) has moved from a specialized technological field into routine academic practice. Generative AI tools are now used for drafting, summarizing, translation, coding, literature searches, preparation of teaching materials and preliminary data analysis. In higher education these activities directly intersect with copyright-sensitive processes: creating scholarly and teaching texts, reusing third-party materials, producing presentations and course content, working with databases and datasets, and publishing or sharing outputs through institutional platforms. The question is therefore not whether AI should be present in universities, but under what conditions it can be used in a legally sound, transparent and academically responsible manner.

The European legal framework does not regulate academic AI use through copyright law alone. Regulation (EU) 2024/1689 (the AI Act) defines AI systems and introduces, among other duties, a general requirement of AI literacy for providers and deployers of AI systems. For universities this requirement is relevant because staff and students increasingly act as users of AI systems in institutional contexts. At the same time, copyright questions remain governed by the national copyright framework and by harmonized EU copyright law, including Directive (EU) 2019/790 on copyright in the Digital Single Market, which introduced specific text-and-data mining exceptions relevant for research. In the Czech context, these rules operate together with Act No. 121/2000 Coll., the Copyright Act, which remains decisive for the protection of works, lawful uses, employee works and school works (European Union, 2019; European Union, 2024; Czech Republic, 2000).

AI copyright literacy can therefore be understood as a specific intersection of AI literacy and copyright literacy. It covers the ability to recognize when the use of AI may affect copyright-protected materials, to distinguish between human authorship and machine-generated assistance, to assess whether an input, dataset or prompt contains protected or confidential content, and to decide when AI use should be disclosed. In academic practice, this literacy also includes knowledge of institutional rules, citation and attribution practices, license conditions, the limits of text-and-data mining, and the risk that generated outputs may reproduce or transform protected expression. The doctrinal literature on AI-assisted creation stresses that the copyright status of outputs depends on human creative choices expressed in the final result, not on the mere fact that an AI tool was used (Hugenholtz and Quintais, 2021).

[1] University of Pardubice, Faculty of Economics and Administration, Studentská 95, 532 10 Pardubice, jana.heckenbergerova@upce.cz
[2] University of Pardubice, Faculty of Economics and Administration, Studentská 95, 532 10 Pardubice, barbora.kovarova@upce.cz

Existing research provides several foundations for such an approach. The AI literacy literature conceptualizes AI literacy as a multidimensional competence that combines understanding, use, evaluation and ethical reflection. Long and Magerko (2020) identify AI literacy as a set of competencies needed to interact with AI critically, while Ng et al. (2021) propose four broader aspects: knowing and understanding AI, using and applying AI, evaluating and creating with AI, and addressing ethical issues. UNESCO similarly emphasizes that generative AI in education and research requires human-centered governance, institutional validation of tools and capacity building for teachers, learners and researchers (UNESCO, 2023). These frameworks, however, usually address AI competence broadly and do not focus in detail on copyright-specific risks.

Empirical studies on generative AI in higher education show that students and staff are already using AI tools, but often under conditions of uncertainty. Chan and Hu (2023), based on a survey of 399 students in Hong Kong, report generally positive perceptions of generative AI, alongside concerns about accuracy, privacy and ethical issues. Johnston et al. (2024), in a survey of 2,555 students at the University of Liverpool, found that awareness of generative AI was widespread and that students supported clear institutional policies rather than a complete ban. Gruenhagen et al. (2024) showed, based on a survey of Australian university students, that more than one third had used chatbots for assessment assistance and did not necessarily perceive such use as a breach of academic integrity. Brown et al. (2025) further demonstrate that AI use, attitudes and self-reported AI literacy may differ between students and staff and may be influenced by socio-demographic factors, while participants also expressed a need for clearer guidance.

A parallel strand of research examines institutional responses. Comparative analyses of university AI policies show that higher education institutions increasingly address academic integrity, assessment design, ethical use, privacy and training, but that policy frameworks remain uneven and often leave responsibility to individual teachers or departments (Jin et al., 2025; An et al., 2025). These findings are important for the present study because they suggest that compliance cannot be explained only by individual knowledge. The organizational environment, the clarity of rules and the disciplinary context in which AI is used may significantly shape behavior.

Finally, research on copyright literacy predates the current wave of generative AI and shows that copyright knowledge in academic and information environments is uneven. Morrison and Secker (2015) identified gaps in copyright literacy among librarians and cultural heritage professionals in the United Kingdom, while the multinational study by Todorova et al. (2017) linked copyright literacy to institutional policy, professional training and the role of copyright advisors. In the Czech context, Kovářová (2019) found that library and information science students recognized the importance of copyright education but also showed knowledge gaps, for example in relation to databases and unpublished works. Recent work on copyright anxiety in higher education further suggests that uncertainty about copyright may lead to risk-averse behavior and may negatively affect teaching and research practices (Wakaruk et al., 2025).

The existing literature thus points to a clear research gap. AI literacy, generative AI use, academic integrity and copyright literacy have been studied, but rarely as an integrated problem of AI copyright literacy in the everyday environment of a university. The present survey responds to this gap by treating AI copyright literacy not merely as abstract legal knowledge, but as a measurable set of awareness, risk perception and compliance-related behaviors shaped by both individual and structural factors.

2 An Academic Survey Study

2.1 Research Objective

The objective of the questionnaire survey is to quantify how individual, disciplinary and institutional factors are associated with AI-related behavior, copyright literacy and compliance in the academic environment. The survey focuses primarily on the University of Pardubice and operationalizes the legal and ethical issues described above into measurable indicators, including the intensity of AI use, awareness of copyright-related risks, knowledge of internal rules, disclosure of AI assistance and selected forms of compliant or non-compliant behavior.

The questionnaire survey was originally developed as part of the doctoral research of one of the authors, Barbora Kovářová, conducted at the Faculty of Law, Charles University, within a dissertation project focused on artificial intelligence and copyright issues in the academic environment. This background explains the legal and institutional orientation of the study and its focus on the practical implementation of copyright-related requirements in university practice. At the same time, the present article develops this dissertation-based empirical material into a broader quantitative analysis conducted in cooperation with colleagues from the math department.

The study is designed to provide empirical evidence for institutional decision-making. Its purpose is not to resolve doctrinal questions of copyright law, but to identify which factors relate to higher or lower levels of AI copyright literacy and whether differences are driven mainly by individual characteristics of respondents or by structural features of the academic environment, especially disciplinary orientation and institutional guidance. The results should support the development of targeted university policies, training activities and discipline-specific recommendations for responsible AI use.

2.2 Questionnaire Structure

The questionnaire contains 47–62 items, depending on the version and vignette scope, and is organized into several thematic blocks. The first block captures demographic characteristics, including gender, age, academic status, study level, field of study, and home institution. The second block (AI-USE INTENSITY) measures the intensity of AI use through items U0–U8. Item U0 assesses overall frequency. Items U1–U7 measure the frequency of AI use across specific tasks, such as brainstorming, summarization, translation, text generation, visual generation, data analysis, and presentation design. Item U8 captures risky behavior, defined as entering sensitive or unpublished materials into AI tools. An **AI-Use index** is computed as the sum of U1–U7 answers scaled as 0–6. Item U8 is used to construct a binary risk indicator; respondents reporting such behavior at least 2–3 times per month were classified as high-risk. The third block (COPYRIGHT-LITERACY) consists of a 10-item knowledge test (L1–L10) covering key copyright domains relevant to academia. The **CL-index** is defined as the sum of correct responses (range 0–10). "Don't know" responses are coded as incorrect, leading to a conservative estimate of literacy. The fourth block (COMPLIANCE & DISCLOSURE) assesses adherence to general rules of AI use, such as disclosure of AI assistance and verification of outputs. A **CD-index** is calculated as the sum of six items (C1–C6) on a 0–4 scale. The fifth block (POLICY AWARENESS) examines awareness and perceives clarity of institutional AI policies (P1–P4), based on answers the **PA-index** is evaluated. The sixth block includes two parallel modules for students (S1–S10) and academic staff (A1–A10). Each module concluded with vignette-based scenarios (S11–S12; A11–A12), which tests the ability to apply legal rules in concrete AI-related situations.

2.3 Respondents structure

A total of 897 responses were collected; one test entry was excluded, yielding a final dataset of 896 unique respondents. Data quality checks were performed, including deduplication, straight-lining analysis, verification of time logs, and consistency checks of key variables. Identified deviations remained within acceptable limits and allowed further analysis. The sample size is adequate for exploratory statistical analysis. While a simple heuristic (the square root rule) would imply a substantially larger underlying population (approximately 800000), the actual population of university students and academic staff in the Czech Republic is considerably smaller, at roughly 335000.

The sample structure shows a strong predominance of students (867; 96,8%) over academic staff (30; 3,2%). Females prevail (58,1%) over males (40,3%). The age distribution is dominated by respondents under 22 years (72,5%), reflecting the largely student-based composition of the sample. Table 1 summarizes absolute and relative frequencies of individual characteristics in a contingency table. Geographically, the sample is concentrated mainly at the University of Pardubice (84,4%), followed by Charles University (13,2%) and other institutions (2,4%). In the field of study, illustrated by the pie chart in Figure 1, economic and business disciplines cover approximately half of the sample, followed by social sciences and humanities, STEM fields without IT, IT and data sciences, and other areas. This concentration constitutes a limitation for generalizing the findings to the Czech academic sector.

Age			**Academic status**		**Gender**	
	AF - n	**RF** in **%**	**Student - AF**	**A. Staff - AF**	**Male - RF** in **%**	**Female - RF** in **%**
≤ 22 years	650	72,5	650	0	28,65	42,47
23-30 years	178	19,8	176	2	7,69	12,15
31-45 years	40	4,5	29	11	2,45	2,01
46+ years	28	3,1	11	17	1,56	1,45

Table 1 Contingency table of absolute and relative frequencies for Age, Gender and Academic status

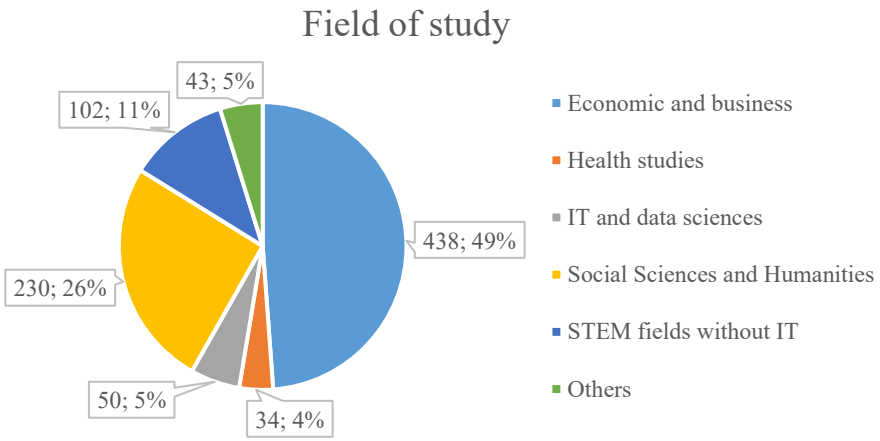


Figure 1 Field of study frequencies visualized by pie chart

2.4 Quantitative analysis of survey outcomes

The statistical distributions of the survey outcomes (AI-Use index, CL-index, CD-index and PA-index) are visualized by histograms in Figure 2, and their basic sample characteristics are summarized in Table 2.

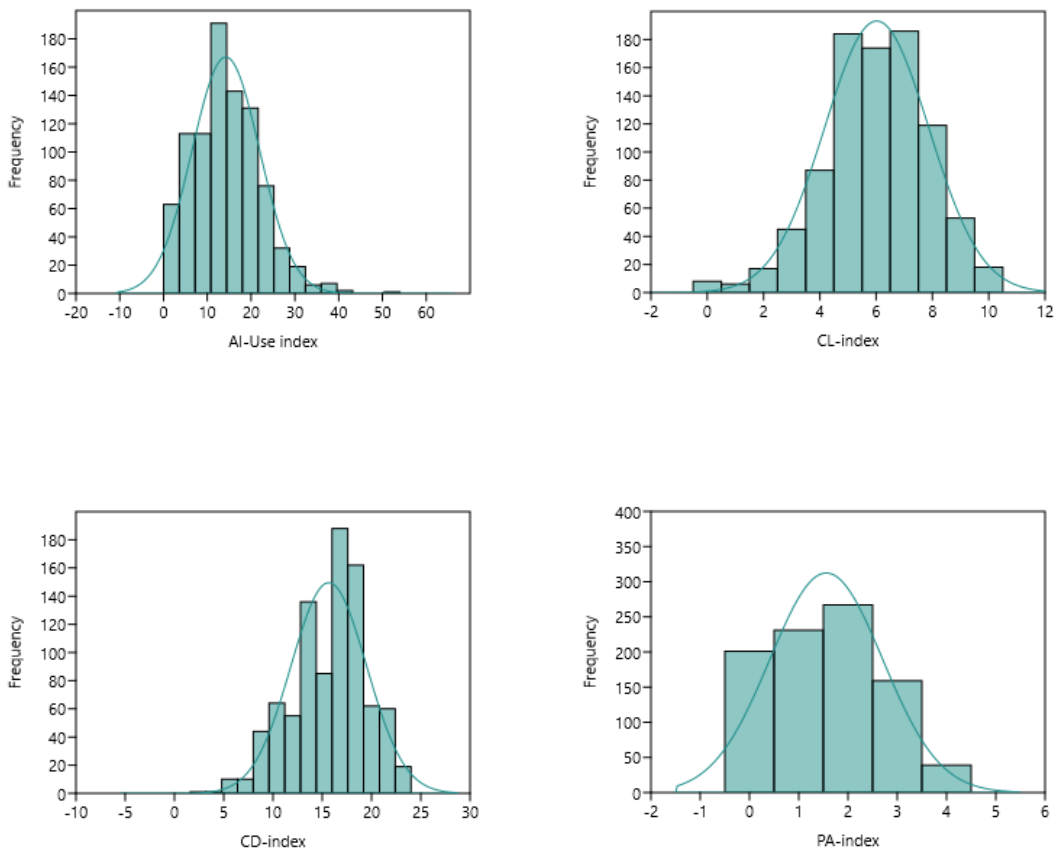


Figure 2 Frequency histograms of survey outcomes
(AI-Use index, CL-index, CD-index and PA-index)

From the outputs of the second block (AI-USE INTENSITY) of the survey, the results indicate that generative AI is widely adopted in academia. It has become a routine part of both study and work. In total, 73,8% of respondents

report using AI at least once a week. Higher intensity of use is observed among academic staff compared to students. In terms of applications, AI is mainly used as a support tool for text-related tasks. The most common utilizations include idea generation, outlining, summarizing academic sources, and language editing or translation. From a risk perspective, answers on item U8 show that 21,9% of respondents upload unpublished materials to AI tools at least 2–3 times per month or more frequently.

Sample characteristics	AI-Use index	CL-index	CD-index	PA-index
average	14,22	6,02	15,66	1,56
median	14	6	16	2
modus	17	7	17	2
dispersion	59,37	3,43	14,65	1,31
std deviation	7,71	1,85	3,83	1,15
coef. of variation	0,542	0,308	0,244	0,735
skewness	0,57	-0,37	-0,36	0,20
kurtosis	0,84	0,31	-0,14	-0,86
variation range	54	10	24	4

Table 2 Sample characteristics of survey outcomes (AI-Use index, CL-index, CD-index and PA-index)

The average level of copyright literacy, measured by the CL-index, can be considered as moderate. Respondents achieved an average score of 6,02 out of 10 questions, corresponding to 60,2% correctness. The distribution is slightly left-skewed, indicating the presence of a subgroup with higher proficiency alongside a non-negligible share with low levels of knowledge. A detailed analysis reveals substantial variability in item-level performance. High success rates are associated with traditional academic situations. Medium levels are observed in items related to the online environment and AI. Low performance is concentrated in areas concerning contractual restrictions and collective rights management. Moreover, approximately 37,6% of respondents consider output generated solely by AI to be protected by copyright.

The mean compliance index is 15,66 with a standard deviation of 3,83, and the coefficient of variation is equal to 0,244 indicating moderate variability. This suggests a moderately elevated level of reported compliance, but with noticeable dispersion. The CD-index shows a relatively wide distribution with slight left-skewness. A gap appears between general attitudes and actual practices, especially in AI disclosure. Among students, transparency is problematic; about one-fifth do not disclose AI even in substantial use, and over one-third report uncertainty about disclosure obligations. Academic staff show higher compliance, mostly reporting consistent disclosure. This aligns with stricter ethical standards and institutional policies requiring transparent AI disclosure.

The distribution of the PA-index has an average of 1,56, a standard deviation 1,15, and a coefficient of variation 0,735, which indicates high relative variability. The slightly positive skewness and negative kurtosis suggest a weak right-skewed and relatively flat distribution with a broad range. These results are consistent with the survey findings, which show a very low level of awareness of AI usage policies. About half of the respondents either do not know the rules or do not know their content, while only a small proportion demonstrates at least basic understanding. This contrasts with international practice and indicates a substantial deficit in policy awareness. It supports the need for systematic implementation of clearly defined internal regulations.

3 Inference Analysis of the Survey Outcomes

3.1 Factors affecting AI use intensity

Age: Age is a categorical variable, while the AI-Use index can be treated as continuous, so ANOVA can be initially considered. However, normality within subgroups cannot be assumed; therefore, the non-parametric Kruskal–Wallis test has to be applied with the level of significance $\alpha = 0,05$.

The null hypothesis states that medians of the AI-Use index do not differ across age groups. The alternative hypothesis assumes that at least one group median significantly differs, and thus age has a significant effect on AI-Use index value.

The test statistic H = 2,322 with a corresponding p-value of 0,5076; therefore, the null hypothesis cannot be rejected as the p-value is higher than the selected α. Consequently, age category does not significantly affect the AI-Use index. Nevertheless, minor differences between groups may still exist, but they are not statistically detectable. Thus, **age is not an essential determinant of AI use intensity**.

Gender: Gender is treated as a binary categorical variable with two categories: male and female. Other responses are excluded from the analysis. The AI-Use index can be therefore divided into two independent samples based on gender and two sample testing procedures could be applied. Normality is not assumed in either subgroup; therefore, the non-parametric Mann–Whitney U test is used with a significance level of α = 0,05.

The null hypothesis H0 assumes that AI-Use index medians are equal in man and female data group. The alternative hypothesis H1 contradicts that both medians are significantly different and therefore gender has a significant effect on AI-Use index value.

The resulting test statistic U = 222630 with a corresponding p-value of 0,3136. Therefore, the null hypothesis cannot be rejected on given significance level. Gender does not significantly influence the AI-Use index and **gender is not a crucial factor of AI use intensity**.

Academic status: Analogically to the previous paragraph, academic status can be considered as a binary categorical variable with two categories: student and academic staff. The AI-Use index dataset is therefore divided into two independent samples, nevertheless normality still cannot be assumed. The non-parametric Mann–Whitney U test is then performed with a significance level of α = 0,05.

The null hypothesis H0 states that AI-Use index medians are equal in students and academics data group. The alternative hypothesis H1 claims that both medians significantly differ and that academic status has a significant effect on AI-Use index.

The U test statistic is equal to 9550 with a corresponding p-value of 0,0132. At the chosen significance level, the null hypothesis H0 has to be rejected in favor of the alternative hypothesis. Therefore, it can be concluded that there is a statistically significant difference between students and academic staff. Academic staff achieve, on average, higher AI-Use index values by 5 points, indicating more frequent use of different AI tools in this group. **Academic status is a key determinant of AI use intensity.**

Results of all testing procedures are summarized in following Table 3.

Factor	test	statistic	p-value	H0 decision	significance
Age	Kruskal-Wallis	H = 2,322	0,5076	cannot be rejected	**NO**
Gender	Mann–Whitney U test	U = 222630	0,3136	cannot be rejected	**NO**
Academic status	Mann–Whitney U test	U = 9550	0,0132	has to be rejected	**YES**
Field of study	Kruskal-Wallis	H = 41,86	$6{,}071 \cdot 10^{-8}$	has to be rejected	**YES**

Table 3 Summary of statistical inferences with AI-Use index

Field of study: Finally, let us quantify the frequency of AI tools utilization within various fields of study. AI-Use index dataset is divided into six subcategories: Economic and business (EB), Health studies (HS), IT and data sciences (IT), Social Sciences and Humanities (SSH), STEM fields without IT (STEM) and others (O). Based on the tests of normal distribution, normality cannot be assumed within all mentioned subgroups. Therefore, nonparametric inference methods such as Kruskal-Wallis test has to be performed with a significance level of α = 0,05.

The null hypothesis states that medians of the AI-Use index are equal across all six fields of study. The alternative hypothesis contradicts that at least one subcategory median significantly differs from others, and thus field of study has a significant effect on AI-Use index value.

The test statistic reaches H = 41,86 with a corresponding p-value of $6{,}071 \cdot 10^{-8}$. As the p-value is lower than α, the null hypothesis has to be rejected at the given significance level. It can therefore be concluded that field of study substantially affects the AI-Use index. Mann-Whitney pairwise comparisons indicate that IT with highest median differs from others, followed by EB, SSH and O. Median values in Health Studies and STEM fields are

approximately 8 points lower than in IT, as illustrated in Figure 3 (Left). The results indicate that **field of study is a significant factor of AI use intensity**.

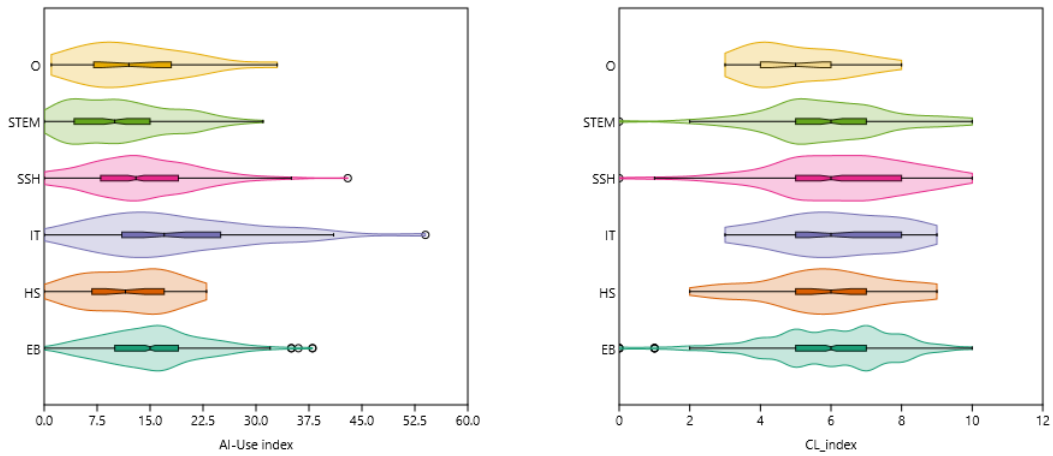


Figure 3 Left - Violin Boxplots showing distributions of AI-Use index within Fields of study
Right - Violin Boxplots showing distributions of CL-index within Fields of study

3.2 Factors affecting AI Copyright Literacy

In the following section, the effect of input factors on the CL-index will be analyzed. Since the selected procedures, tests, and hypotheses are identical to those used in the inference analyses of the AI-Use index, all results can be summarized in Table 4.

Factor	test	statistic	p-value	H0 decision	significance
Age	Kruskal-Wallis	H = 1,69	0,6281	cannot be rejected	**NO**
Gender	Mann–Whitney U test	U = 226550	0,63096	cannot be rejected	**NO**
Academic status	Mann–Whitney U test	U = 12025	0,495	cannot be rejected	**NO**
Field of study	Kruskal-Wallis	H = 14,87	0,009	has to be rejected	**YES**

Table 4 Summary of statistical inferences with CL-index

Field of study is an essential determinant of AI Copyright Literacy. Substantial differences within subcategories are revealed by Mann-Whitney pairwise comparisons and visualized by Violin Boxplots in Figure 3 (Right). The highest average CL-index is in Social Sciences and Humanities, followed by IT, EB, STEM and HS. Other areas mainly covered by transportation students have significantly worse results, implying H0 rejection and field of study significance.

3.3 Factors affecting Compliance and Disclosure Indicators

Inference analyses results revealing significant determinants of CD-index are summarized in Table 5. Gender is the only factor that does not affect Compliance and Disclosure indicators significantly. Pairwise comparisons between age groups indicate that respondents aged under 22 years exhibit lower CD-index values and create a substantial subgroup. Academic staff average CD-index is 3 points higher than student average and shows less variance, as illustrates Violin Boxplot in Figure 4. Fields of study can be divided into two distinct groups. The first group includes IT, SSH, and STEM subgroups, with a median of CD_index equal to 17, while the second group comprises EB and HS students, with a median CD_index that is 2 points lower.

Factor	test	statistic	p-value	H0 decision	significance
Age	Kruskal-Wallis	H = 33,97	$1{,}811 \cdot 10^{-7}$	has to be rejected	**YES**
Gender	Mann–Whitney U test	U = 224260	0,4306	cannot be rejected	**NO**
Academic status	Mann–Whitney U test	U = 6958	$1{,}3699 \cdot 10^{-5}$	has to be rejected	**YES**
Field of study	Kruskal-Wallis	H = 36,8	$5{,}876 \cdot 10^{-7}$	has to be rejected	**YES**

Table 5 Summary of statistical inferences with CD-index

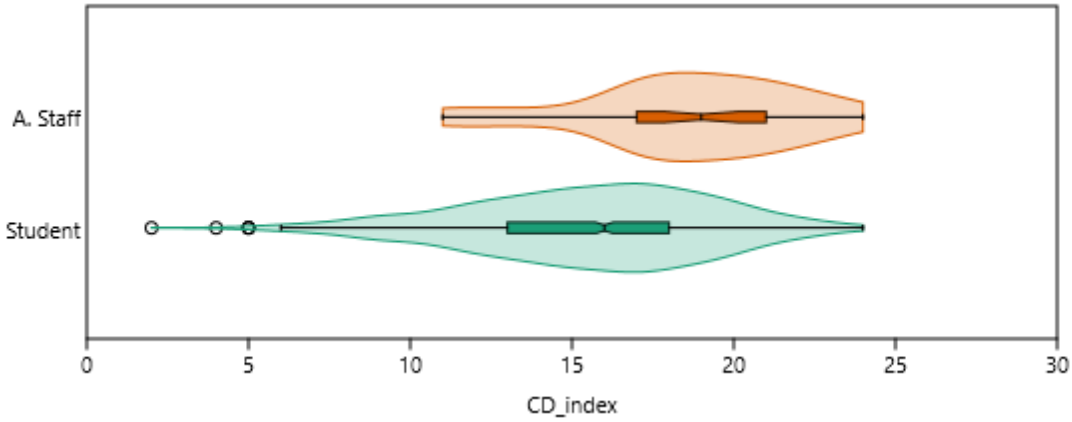


Figure 4 Violin Boxplots showing distributions of CD-index for students and academic staff

4 Discussion, Conclusions and Recommendations

The present study shows that generative AI is already a routine part of academic work, while AI copyright literacy remains only moderate. Compliance and disclosure indicators are slightly above average, but policy awareness is low and highly variable. Risk-related behavior remains prevalent: 21,9% of respondents report uploading unpublished materials to AI tools at least 2-3 times per month, and 37,6% misclassify outputs generated solely by AI as copyright-protected. These results point to a structural imbalance between high AI adoption and insufficient legal preparedness.

In comparison with survey studies conducted abroad, our findings align with general patterns reported in the literature. Similarly to Chan and Hu (2023), Johnston et al. (2024) and Gruenhagen et al. (2024), AI tools are not perceived as exceptional technologies but rather as routine instruments used under conditions of uncertainty. The key distinction lies in the focus of the present study, which identifies copyright-specific risks in everyday academic AI use, including uncertainty about the legal status of outputs, disclosure obligations, and the handling of protected or unpublished materials.

The results also partly overlap with Brown et al. (2025), who reported differences in AI use and literacy between students and staff. In our sample, academic staff use AI more intensively and show higher compliance; however, copyright literacy is not significantly associated with academic status, age or gender. The most pronounced differences occur across fields of study, suggesting that AI copyright literacy is shaped primarily by structural and disciplinary conditions.

Low and uneven awareness of institutional rules corresponds to findings that university AI policy frameworks remain heterogeneous and insufficiently embedded in practice (An et al., 2025; Jin et al., 2025). Consequently, formal rules cannot effectively guide users if they are unclear or not translated into discipline-specific contexts.

Recommendations should therefore emphasize stronger policy awareness and clearer communication of institutional rules. Scenario-based legal training is needed to support decision-making, particularly in disciplines with lower literacy or compliance levels. Risk mitigation should include safer handling of unpublished and protected materials, the use of institutionally approved tools, and standardized disclosure rules. Such measures are essential for maintaining academic integrity and reducing copyright-related uncertainty in higher education.

Acknowledgements

This contribution has been supported by the institutional support of the University of Pardubice, Czech Republic.

References

An Y, Yu J H, James S (2025) Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. International Journal of Educational Technology in Higher Education 22:10. https://doi.org/10.1186/s41239-025-00507-3

Brown R, Sillence E, Branley-Bell D (2025) AcademAI: Investigating AI Usage, Attitudes, and Literacy in Higher Education and Research. Journal of Educational Technology Systems 54(1):6–33. https://doi.org/10.1177/00472395251347304

Chan C K Y, Hu W (2023) Students' voices on generative AI: perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education 20:43. https://doi.org/10.1186/s41239-023-00411-8

Gruenhagen J H, Sinclair P M, Carroll J A, Baker P R A, Wilson A, Demant D (2024) The rapid rise of generative AI and its implications for academic integrity: Students' perceptions and use of chatbots for assistance with assessments. Computers and Education: Artificial Intelligence 7:100273. https://doi.org/10.1016/j.caeai.2024.100273

Jin Y, Yan L, Echeverria V, Gašević D, Martinez-Maldonado R (2025) Generative AI in higher education: A global perspective of institutional adoption policies and guidelines. Computers and Education: Artificial Intelligence 8:100348. https://doi.org/10.1016/j.caeai.2024.100348

Johnston H, Wells R F, Shanks E M, Boey T, Parsons BN (2024) Student perspectives on the use of generative artificial intelligence technologies in higher education. International Journal for Educational Integrity 20:2. https://doi.org/10.1007/s40979-024-00149-4

Kovářová P (2019) Copyright Literacy of LIS Students in the Czech Republic. Information Literacy in Everyday Life. ECIL 2018. Communications in Computer and Information Science, vol 989. Springer, Cham, pp 585–593. https://doi.org/10.1007/978-3-030-13472-3_55

Long D, Magerko B (2020) What is AI Literacy? Competencies and Design Considerations. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, pp 1–16. https://doi.org/10.1145/3313831.3376727

Morrison C M, Secker J (2015) Copyright Literacy in the UK: a survey of librarians and other cultural heritage sector professionals. Library and Information Research 39(121):75–97. https://doi.org/10.29173/lirg675

Ng D T K, Leung J K L, Chu S K W, Qiao M S (2021) Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence 2:100041. https://doi.org/10.1016/j.caeai.2021.100041

Todorova T Y, Kurbanoğlu S, Boustany J, Doğan G, Saunders L, Horvat A, Terra AL, Landøy A, Repanovici A, Morrison C, Sanchez Vanderkast E J, Secker J, Rudzioniene J, Kortelainen T, Koltay T (2017) Information professionals and copyright literacy: a multinational study. Library Management 38(6/7):323–344. https://doi.org/10.1108/LM-01-2017-0007

UNESCO (2023) Guidance for generative AI in education and research. UNESCO, Paris. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Wakaruk A, Secker J, Morrison C (2025) Copyright Anxiety and Legal Chill in Higher Education: A Comparison of Canada and the United Kingdom (UK). Journal of Copyright in Education & Librarianship 8(1). https://doi.org/10.17161/jcel.v8i1.23058

Mitigating Conflict in Fair Sharing via Resource Design

Robert Hlavatý[1]

Abstract. The fair sharing problem concerns dividing resources among multiple parties so that each party feels it received a fair share according to its own valuation. We study the problem in a specific setting in which resource availability is interval-based rather than a fixed amount. The aim is then to find the ideal amounts of resources within given interval bounds that would preferably eliminate or mitigate the conflict between receiving parties. The resources are often held by an authority which from their own perspective are subject to given (unit) costs and therefore the authority seeks to set up such an arrangement of the resources within its available budget. We formulate the problem as a multiobjective optimisation problem and consider different fair-sharing problem settings: with or without classic envy-freeness among parties, and with different goals – utilitarian or egalitarian. We present an illustrative computational example on the synthetically generated data. The results suggest that, by identifying an appropriate resource structure within the admissible bounds, it is possible to improve the attained level of fairness relative to the original configuration.

Keywords: fair sharing, multiobjective optimisation, resource design

JEL Classification: C61

1 Introduction

Fair division is concerned with allocating scarce resources among agents with heterogeneous preferences in such a way that the resulting allocation is regarded as fair according to some normative criterion (Amanatidis et al, 2023). In the classical literature, the problem was studied mainly for divisible resources, where fairness was associated with concepts such as envy-freeness, proportionality, and equitable Pareto-efficient divisions (Varian, 1974; Weller, 1985). The continuous version of the problem has also given rise to constructive cake-cutting procedures, including envy-free protocols for heterogeneous divisible goods (Brams and Taylor, 1995), while more recent work has revisited the continuous setting from an algorithmic viewpoint, providing an efficient algorithm with fast convergence to a fair solution (Blank and Polyakov, 2024).

By contrast, the fair division of indivisible goods is inherently more difficult, since exact envy-freeness and proportionality may fail to exist and one must often work with relaxations or alternative fairness guarantees (Amanatidis et al, 2023). This has led to research on ordinal preferences, incomplete preferences, and algorithmic fairness notions for indivisible allocation problems (Bouveret et al, 2010). Recent surveys also show that the area has expanded beyond the traditional divisible/indivisible dichotomy toward mixed settings involving divisible and indivisible goods (Liu et al, 2024). A related line of work studies the compatibility of fairness and efficiency. In particular, Cole and Tao, 2021 show that although envy-freeness and Pareto efficiency may fail simultaneously in deterministic allocations, ex ante mixed allocations can recover both properties under suitable conditions.

A further challenge arises when the allocation problem is affected by uncertainty. Existing studies have considered, among others, uncertain resource availability (Buermann et al, 2020), uncertain needs (Xue, 2018), partial information about agents' valuations (Halpern and Shah, 2021), and robust or stochastic fair allocation under uncertain utilities (Karsu et al, 2025; Lobo et al, 2024). Taken together, these contributions indicate that fairness, efficiency, and uncertainty are tightly interrelated, but they still predominantly treat the available resource structure as fixed.

In this paper, we study a different perspective. We assume that the availability of each resource type is not fixed in advance, but can be adjusted within predefined lower and upper bounds under a budget constraint imposed by a central authority. The problem is therefore not only to allocate resources fairly, but also to identify a resource configuration that better matches the heterogeneous preferences of the participating agents and thereby mitigates potential conflict among them. We formulate this task as a multiobjective optimisation problem and examine several variants, including utilitarian and egalitarian formulations as well as settings with and without classical envy-freeness. In this sense, the proposed model moves from fair allocation of a fixed system toward fair redesign of the resource structure itself.

[1] CZU Prague, Department of Systems Engineering, Kamýcká 129, 16500 Prague - Suchdol, hlavaty@pef.czu.cz

2 Methods

In this section, we formulate the fair sharing problem as a multiobjective optimisation problem with two different forms. The basic idea is that each agent represents one criterion through their individual utility, and the aim is to identify such a feasible allocation that the resulting distribution of utilities satisfies a chosen objective.

Let $N = \{1, \dots, n\}$ denote the set of agents and $M = \{1, \dots, m\}$ the set of resource types. For each agent $i \in N$ and resource $j \in M$, let $v_{ij} \geq 0$ denote the unit utility of agent i from one unit of resource j. Further, let $c_j \geq 0$ be the unit cost of resource j, let L_j and U_j denote the lower and upper bounds on the admissible availability of resource j, and let B denote the total (resource) budget available to the central authority that allocates the resources among agents. The decision variables are defined as follows:

- $x_{ij} \geq 0$: amount of resource j allocated to agent i,
- $y_j \geq 0$: total available amount of resource j to be established within the predefined bounds.

We assume simple additive utilities with no complementarities, substitutions or synergies. The utility attained by agent i is

$$u_i(\boldsymbol{x}) = \sum_{j \in M} v_{ij} x_{ij}, \qquad i \in N. \tag{1}$$

The basic feasibility set $\mathcal{F}$ is given by the constraints

$$\sum_{i \in N} x_{ij} \leq y_j, \qquad j \in M, \tag{2}$$

$$L_j \leq y_j \leq U_j, \qquad j \in M, \tag{3}$$

$$x_{ij} \geq 0, \qquad i \in N,\ j \in M. \tag{4}$$

The budget feasibility set $\mathcal{F}^B$ is given by the constraint set

$$\sum_{j \in M} c_j y_j \leq B, \tag{5}$$

and in addition, we consider a variant in which the allocation must satisfy the classical envy-freeness condition. This means that each agent evaluates their own bundle at least as worthy as the bundle of any other agent, i.e.,

$$\sum_{j \in M} v_{ij} x_{ij} \geq \sum_{j \in M} v_{ij} x_{kj}, \qquad i, k \in N,\ i \neq k. \tag{6}$$

and we denote this envy-free feasibility set as $\mathcal{F}^{EF}$. We further denote $\mathcal{F}^* = \mathcal{F} \cap \mathcal{F}^{EF}$ or $\mathcal{F}^* = \mathcal{F}$ as relevant feasible set, depending on whether the envy-freeness is imposed. From the multicriteria point of view, the problem may be written in the general form

$$\max \left(u_1(\boldsymbol{x}), u_2(\boldsymbol{x}), \dots, u_n(\boldsymbol{x}) \right), \tag{7}$$

subject to constraint set $\mathcal{F}^* \cap \mathcal{F}^B$. Since the formulation (7) does not represent a unique solution in general, we work with two scalarisations (8), (9) corresponding to two different viewpoints.

2.1 Utilitarian formulation

The first formulation adopts the utilitarian perspective and seeks to maximise the aggregate utility over all agents:

$$\max \sum_{i \in N} u_i(\boldsymbol{x}) = \max \sum_{i \in N} \sum_{j \in M} v_{ij} x_{ij}. \tag{8}$$

This formulation favours resource configurations that maximise total realised utility, regardless of how evenly this utility is distributed among agents.

2.2 Egalitarian formulation

The second formulation adopts a more fairness-oriented perspective and seeks to maximise the utility of the worst agent. To this end, we introduce an auxiliary variable t that represents the minimum utility attained across all agents. The corresponding model is

$$\max t \tag{9}$$

subject to

$$\sum_{j \in M} v_{ij} x_{ij} \geq t, \qquad i \in N. \tag{10}$$

This formulation ensures that the minimum individual utility is as large as possible and can therefore be interpreted as an egalitarian baseline model. In contrast to the utilitarian formulation, it explicitly protects agents with relatively weak allocations and promotes a more balanced distribution of resources across the agent set.

3 Design-oriented reformulation

We further use both baseline scalarised formulations to develop a design-oriented reformulation inspired by the metaoptimisation framework developed by Zeleny, 1986. The proposed model can be viewed as an extension of classical fair-sharing formulations: while standard models focus on the fair allocation of a fixed resource structure, here the resource structure itself becomes partially decision-dependent through the redesign of resource availabilities within predefined bounds. This is why the resource-design perspective inspired by Zeleny, 1986 is methodologically relevant in the present setting. The main principle is not to take the currently available resource structure as fixed, but rather to identify what resource configuration, within the predefined lower and upper bounds, would be required to support an improved utility level at the minimum possible cost. We present the following three-phase framework in order to achieve the above goal.

Phase 1: Identification of the target level. In the first phase, an ideal performance level is sought. In the utilitarian variant, this target is given by the componentwise ideal point obtained from the original vector of agent utilities. In the egalitarian variant, the target is the ideal value of the auxiliary variable t, representing the highest jointly attainable minimum utility.

Utilitarian variant

$$u_i^* = \max_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{F}^* \cap \mathcal{F}^B} \sum_{j \in M} v_{ij} x_{ij}, \qquad i \in N, \tag{11}$$

which yields the ideal point

$$\boldsymbol{u^*} = (u_1^*, \ldots, u_n^*). \tag{12}$$

Egalitarian variant

$$t^* = \max\ t \tag{13}$$

subject to

$$\sum_{j \in M} v_{ij} x_{ij} \geq t, \qquad i \in N, \tag{14}$$

$$(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{F}^* \cap \mathcal{F}^B. \tag{15}$$

Phase 2: Minimum-cost attainment of the target. In the second phase, the target identified above is imposed as a required performance level, and the objective is to determine the least costly resource configuration capable of attaining it.

Utilitarian variant

$$\min \sum_{j \in M} c_j y_j \tag{16}$$

subject to

$$\sum_{j \in M} v_{ij} x_{ij} \geq u_i^*, \qquad i \in N, \tag{17}$$

$$(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{F}^*. \tag{18}$$

Egalitarian variant

$$\min \sum_{j \in M} c_j y_j \tag{19}$$

subject to

$$\sum_{j \in M} v_{ij} x_{ij} \geq t^*, \qquad i \in N, \tag{20}$$

$$(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{F}^*. \tag{21}$$

If the minimum budget obtained in **Phase 2** exceeds the originally available budget B, the target level cannot be attained directly under the given resource limit [1]. In that case, the target must be transformed by means of an

[1] This can only happen in the case of utilitarian variant.

optimum-path ratio in the spirit of Shi, 1995. Let B^* denote the minimum budget required to attain the target identified in **Phase 1**. If $B^* > B$, we define the ratio $\rho = \frac{B}{B^*}$, $0 < \rho < 1$ and replace the original target by its scaled attainable counterpart. In this way, the resource design is projected back onto the admissible budget level while preserving the proportional structure of the desired performance target.

Phase 3: Adjustment in the infeasible case. If the target level obtained in **Phase 1** is not jointly attainable, it must be replaced by an achievable one. This can be done through various techniques of multiobjective optimisation (e.g. STEP method, as in Hlavatý and Brožová, 2023) since the task is basically finding some efficient solution that is, in a given measure, "close" to an ideal solution. In our paper, we apply an auxiliary deviation-minimisation problem and then repeat the minimum-cost design step (**Phase 2**) for the adjusted target. Note that this step may only be necessary in the case of the utilitarian variant. No additional infeasibility-adjustment phase is needed in the egalitarian case, since the target value t^* is itself obtained as an optimal feasible value to the problem (13)–(15).

Utilitarian variant

Introduce deviations $\delta_i \geq 0$ and solve

$$\min \sum_{i \in N} \delta_i \tag{22}$$

subject to

$$\sum_{j \in M} v_{ij} x_{ij} + \delta_i \geq u_i^*, \qquad i \in N, \tag{23}$$

$$(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{F}^*. \tag{24}$$

Then define

$$\bar{u}_i = u_i^* - \bar{\delta}_i, \qquad i \in N, \tag{25}$$

and solve the analogue to **Phase 2** with a new target

$$\min \sum_{j \in M} c_j y_j \tag{26}$$

subject to

$$\sum_{j \in M} v_{ij} x_{ij} \geq \bar{u}_i, \qquad i \in N, \tag{27}$$

$$(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{F}^*. \tag{28}$$

The resulting formulation shifts the emphasis from allocating a fixed resource system to identifying a cost-efficient resource configuration that supports fairness-oriented target utility levels as much as possible. For clarity, the three-phase redesign procedure can be summarised as Algorithm 1. Note that the infeasibility-adjustment step is relevant only for the utilitarian variant.

Algorithm 1 Summary of the redesign procedure

```
Solve Phase 1 and identify the target level.
if the target is jointly infeasible then
    Activate Phase 3 and construct an attainable target level.
    Solve Phase 2 for the adjusted target.
else
    Solve Phase 2 and compute the minimum required budget B*.
    if B* ≤ B then
        The target is directly attainable.
    else
        Compute the optimum-path ratio ρ = B/B*.
        Scale the target accordingly.
    end if
end if
```

4 Illustrative example

4.1 Baseline scenario

Let us demonstrate the proposed methodology on a synthetic example. For simplicity, we assume a case with three agents $N = \{1, 2, 3\}$ and three generic resources $M = \{1, 2, 3\}$ and the individual unit utilities v_{ij} are: $(v_{11}, v_{12}, v_{13}) = (2, 2, 3), (v_{21}, v_{22}, v_{23}) = (2, 3, 2), (v_{31}, v_{32}, v_{33}) = (2, 6, 5)$ generated as $v_{ij} \sim \mathcal{U}_{(1,10)}$. The baseline model is solved with fixed capacities $\bar{y} = (68, 21, 12)$, i.e. $\sum_{i \in N} x_{ij} \leq \bar{y}_j,\ j \in M$, generated as $y_j \sim \mathcal{U}_{(1,100)}$. Since the resource availability is fixed, we do not assume any restrictive bounds as in (3). Note that such scenario is fixed on the input and we will refer to it as the *baseline scenario*. We solve the problem in utilitarian setting (i.e. solving (8) on $\mathcal{F}^* \cap \mathcal{F}^B$ with fixed capacities $\bar{y}$) and then in egalitarian setting (i.e. solving (9) on $\mathcal{F}^* \cap \mathcal{F}^B$ with fixed capacities $\bar{y}$) and obtain the results (rounded, where applicable) in Table 1. The models were implemented in *MS Excel Opensolver* and solved using *COIN-OR CBC* solver.

Utilitarian variant (w/o envy-freeness)

	Resource 1	Resource 2	Resource 3	Utility
Agent 1	68			136
Agent 2				0
Agent 3		21	12	186
Total	68	21	12	322

Egalitarian variant (w/o envy-freeness)

	Resource 1	Resource 2	Resource 3	Utility
Agent 1	29		12	94
Agent 2	39	5.33		94
Agent 3		15.67		94
Total	68	21	12	282

Utilitarian variant (w/ envy-freeness)

	Resource 1	Resource 2	Resource 3	Utility
Agent 1	29.93		5.43	76.14
Agent 2	38.07			76.14
Agent 3		21	6.57	158.86
Total	68	21	12	311.14

Egalitarian variant (w/ envy-freeness)

	Resource 1	Resource 2	Resource 3	Utility
Agent 1	31.81		8.74	89.85
Agent 2	36.19	5.83		89.85
Agent 3		15.17	3.26	107.33
Total	68	21	12	287.03

Table 1 Baseline results for the four model variants.

The utilitarian model dominates the egalitarian one in the total utility because the former chases the total utility maximum regardless of the individuals. Imposing envy-freeness in a utilitarian setting causes a decrease in total utility because *Agent 2* obviously cannot have zero utility in such a setting. On the other hand, the envy-free egalitarian variant shows higher total utility by sacrificing the portion of its equitable distribution among agents.

4.2 Resource structure redesign

Next, we show how to redesign the resource structure in accordance with the approach proposed in Section 3. In addition to baseline scenario input, we assume resource unit costs $(c_1, c_2, c_3) = (9, 3, 4)$, the total available budget corresponding to the original resource structure (i.e. $\boldsymbol{c}^\top \boldsymbol{y}$) which is 723 and resource bounds set to $L_j = 10,\ U_j = 50,\ \forall j$ [2]. Following the steps in Section 3, the different scenarios yield the results (rounded, where applicable) in Table 2. The solutions reported in Table 2 are obtained by the proposed target-based redesign procedure in section 3. In the two egalitarian variants, only **Phase 1** and **Phase 2** were active, whereas in the utilitarian variants the target-based redesign additionally involved **Phase 3** in both cases.

Upon comparing the results before (Table 1) and after (Table 2) resource redesign, it is possible to observe a significant improvement in the total utility for all scenarios. Also, the individual utilities of the agents are significantly better (with the exception of the first scenario, which does not focus on just distribution of utilities). Note that the improved utility levels can be achieved for the same budget $B = 723$ that was used in the fixed scenarios in Table 1. Whereas in case of Rubchinsky, 2010 the conflict resolution is essentially reduced to finding a proportional, equitable, and efficient split of items that are already given. By contrast, our model not only asks how to allocate a fixed resource system fairly, but how to redesign resource availabilities themselves—within lower and upper bounds and under a budget constraint—so that the resulting allocation better fits the heterogeneous preferences of all participants. In this sense, conflict is mitigated not only through the final allocation rule but also through the endogenous design of the underlying resource structure.

[2] The proposed approach is functional even without any resource bounds, however, then it often happens that some of the resources are completely omitted in favour of the others. Setting these bounds forces the use of every resource which is more realistic in practical applications.

Utilitarian redesign (w/o envy-freeness)

	Resource 1	Resource 2	Resource 3	Utility
Agent 1				0
Agent 2	45.19			90.38
Agent 3		45.19	45.19	497.06
Total	45.19	45.19	45.19	587.44

Egalitarian redesign (w/o envy-freeness)

	Resource 1	Resource 2	Resource 3	Utility
Agent 1	1.58		50	153.16
Agent 2	39.87	24.47		153.16
Agent 3		25.53		153.16
Total	41.45	50	50	459.48

Utilitarian redesign (w/ envy-freeness)

	Resource 1	Resource 2	Resource 3	Utility
Agent 1			36.97	110.91
Agent 2	45.19	10.27		121.18
Agent 3		34.92	8.22	250.59
Total	45.19	45.19	45.19	482.68

Egalitarian redesign (w/ envy-freeness)

	Resource 1	Resource 2	Resource 3	Utility
Agent 1	11.3		37.67	135.62
Agent 2	30.14	25.11		135.62
Agent 3		24.89	12.33	210.96
Total	41.44	50	50	482.2

Table 2 Results of the target-based resource redesign procedure for the four model variants.

5 Conclusion

In this paper, we proposed a fair-sharing framework in which resource availabilities are not treated as fixed, but can be redesigned within predefined bounds under a budget constraint. The illustrative results suggest that such a redesign may improve the resulting allocation both from the efficiency-oriented and from the fairness-oriented perspective. Possible application areas include the budget- constrained redesign of shared service capacities, such as laboratory facilities, diagnostic slots, or shared computing resources. Further work should therefore include more extensive computational experiments, a broader parametric analysis, and a systematic investigation of how the utilitarian and egalitarian formulations behave under different budget levels. In particular, it would be of interest to study the threshold effects between the two approaches.

Acknowledgements

This paper was supported by the Czech Science Foundation (GAČR) project No. 25-15714S.

References

Amanatidis G, Aziz H, Birmpas G, et al (2023) Fair division of indivisible goods: Recent progress and open questions. Artificial Intelligence 322:103965. https://doi.org/10.1016/j.artint.2023.103965

Blank ML, Polyakov MO (2024) Elementary Solution to the Fair Division Problem. Problems of Information Transmission 60(1):53–70. https://doi.org/10.1134/S003294602401006X

Bouveret S, Endriss U, Lang J (2010) Fair Division under Ordinal Preferences: Computing Envy-Free Allocations of Indivisible Goods. In: Frontiers in Artificial Intelligence and Applications. IOS Press, https://doi.org/10.3233/978-1-60750-606-5-387

Brams SJ, Taylor AD (1995) On envy-free cake division. Journal of Combinatorial Theory, Series A 70(1):170–173. https://doi.org/10.1016/0097-3165(95)90088-8

Buermann J, Gerding EH, Rastegari B (2020) Fair allocation of resources with uncertain availability. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '20, p 204–212, https://doi.org/10.65109/VEHI8120

Cole R, Tao Y (2021) On the existence of Pareto Efficient and envy-free allocations. Journal of Economic Theory 193:105207. https://doi.org/10.1016/j.jet.2021.105207

Halpern D, Shah N (2021) Fair and Efficient Resource Allocation with Partial Information. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, pp 224–230, https://doi.org/10.24963/ijcai.2021/32

Hlavatý R, Brožová H (2023) Designing an efficient transportation system under a constrained budget and cost uncertainty. In: Proceedings of the 41st International Conference on Mathematical Methods in Economics (MME). University of Economics Prague, Prague, Czech Republic, pp 105–110

Karsu O, Elver IE, Kınık TA (2025) Finding robustly fair solutions in resource allocation. Omega 131:103208. https://doi.org/10.1016/j.omega.2024.103208

Liu S, Lu X, Suzuki M, et al (2024) Mixed Fair Division: A Survey. Journal of Artificial Intelligence Research 80:1373–1406. https://doi.org/10.1613/jair.1.15800

Lobo E, Payan J, Cousins C, et al (2024) Fair and Welfare-Efficient Constrained Multi-matchings under Uncertainty. https://doi.org/10.48550/arXiv.2411.02654

Rubchinsky A (2010) Brams–Taylor model of fair division for divisible and indivisible items. Mathematical Social Sciences 60(1):1–14. https://doi.org/10.1016/j.mathsocsci.2010.03.004

Shi Y (1995) Studies on optimum-path ratios in multicriteria de novo programming problems. Computers & Mathematics with Applications 29(5):43–50. https://doi.org/10.1016/0898-1221(94)00247-I

Varian HR (1974) Equity, envy, and efficiency. Journal of Economic Theory 9(1):63–91. https://doi.org/10.1016/0022-0531(74)90075-1

Weller D (1985) Fair division of a measurable space. Journal of Mathematical Economics 14(1):5–17. https://doi.org/10.1016/0304-4068(85)90023-0

Xue J (2018) Fair division with uncertain needs. Social Choice and Welfare 51(1):105–136. https://doi.org/10.1007/s00355-018-1109-5

Zeleny M (1986) Optimal system design with multiple criteria: De Novo programming approach. Engineering Costs and Production Economics 10(2):89–94. https://doi.org/10.1016/0167-188X(86)90002-9

Assessing SDG 11 Efficiency in Czech Regions: An Iterative BoD-DEA Approach with Weight Restrictions

Eva Hrabovská[1], Markéta Matulová[2]

Abstract. Assessing Sustainable Development Goals (SDGs) at the regional level requires robust quantitative tools capable of aggregating multidimensional indicators into meaningful composite indices. This paper focuses on SDG 11 (Sustainable Cities and Communities) and evaluates the technical efficiency of the 14 Czech regions over the 2016–2023 period. Given the broad framework of SDG 11—encompassing housing, transportation, environment, and public services—constructing a robust model necessitates a systematic selection of representative indicators from an extensive initial dataset. The indicators were selected based on a literature review and categorized into seven sub-domains reflecting targets 11.1–11.7. To maintain the model's discriminatory power, an algorithm was employed to systematically select one variable from each sub-domain, generating all possible indicator combinations. These combinations served as input data for a Benefit-of-the-Doubt (BoD) DEA model assuming constant returns to scale (CRS). Each combination formed an individual model, with the final efficiency score calculated as the arithmetic mean of all iterations. To ensure a balanced perspective, weight restrictions were incorporated, limiting the influence of each indicator to a range of 5% to 30% of the overall assessment. Empirical analysis reveals high efficiency levels in the Karlovy Vary Region and widening gaps between the most and least successful regions, as well as distinct disparities between eastern and western parts of the country. In 2023, the Karlovy Vary Region achieved the highest efficiency (0.97), while the Olomouc Region recorded the lowest (0.37). This study contributes to the discussion on enhancing the discriminatory power of models in the field of sustainable development and provides a framework for data-driven regional management.

Keywords: Sustainable Development, SDG 11, DEA, Benefit of the Doubt, regional efficiency, Czech Republic

JEL Classification: C44

1 Introduction

Since the second half of the 20th century, the concept of sustainable development has evolved from an initially environmental focus into a complex approach striving for balance between ecological stability, social justice, and economic growth. Modern definitions, based on the classic concept of the Brundtland Commission, understand sustainability as responsible behavior that meets the needs of the present without compromising the ability of future generations to meet their own needs (Ministry of Regional Development of the Czech Republic, 2025; Klarin, 2018). This vision was formalized in 2015 with the adoption of Agenda 2030, which defines 17 Sustainable Development Goals (SDGs). Goal 11 (Sustainable Cities and Communities) holds a special significance within this framework, as the current rate of urbanization brings fundamental challenges in terms of emission burden, resource consumption, and quality of life in urban areas (United Nations, 2015; United Nations Information Centre Prague, 2025).

Successful implementation of Agenda 2030, however, requires robust tools for monitoring progress. In international comparisons, the Czech Republic has long performed very well—ranking 8th in the global index in 2021, although the latest Voluntary National Review from 2025 indicates a drop to 12th position (Office of the Government of the Czech Republic and Ministry of the Environment, 2017–2025). Government reports attribute this decline to the accumulation of external shocks, such as the COVID-19 pandemic and the energy crisis. However, academic literature also points to deeper methodological aspects of evaluation. Authors such as Issever Grochová and Litzman (2021) point out that commonly used indices often ignore the baseline conditions of countries, and

[1] Mendel University in Brno, Department of Statistics and Operation Analysis, Zemědělská 1, 613 00 Brno, Czech Republic, xhrabov4@node.mendelu.cz

[2] Mendel University in Brno, Department of Statistics and Operation Analysis, Zemědělská 1, 613 00 Brno, Czech Republic, marketa.matulova@mendelu.cz

therefore recommend the use of Data Envelopment Analysis (DEA) to measure relative success in fulfilling the SDGs.

Specifically, non-parametric methods such as DEA and its specific variant, *Benefit of the Doubt* (BoD), are becoming the standard for creating composite indicators in the field of sustainability (Nežinský, 2020; Mardani et al, 2017). Their main advantage is the ability to aggregate multidimensional and incommensurable data without the need to subjectively determine the weights of individual goals in advance. This objective approach to weight setting was used in an international context, for example, by Guo et al (2024), who took into account the specific priorities of individual countries. Unlike traditional aggregation methods (such as simple averages or Principal Component Analysis) that apply fixed, uniform weights across all evaluated units, the combinatorial approach used in this study naturally respects regional differences.

However, when applying these models to the regional level of the Czech Republic (NUTS 3 level), a fundamental mathematical problem known as the *curse of dimensionality* arises. For a comprehensive evaluation of SDG 11, it is essential to monitor a wide range of indicators, from housing affordability to air quality and waste. Given that the Czech Republic comprises only 14 regions (so-called Decision-Making Units, DMUs), the use of a high number of observed variables leads to an inevitable loss of the model's discriminatory power. Under such conditions, most units mathematically emerge as fully efficient. Traditional solutions in the form of drastic data reduction then lead to unacceptable simplification and loss of information about regional specificities.

This article responds to this methodological challenge and builds on the shift presented by Aparicio et al (2026), who used a modified DEA method based on repeated calculations over reduced datasets to evaluate SDGs across EU countries. The main goal of this study is to apply an innovative iterative BoD-DEA approach with weight restrictions to evaluate the technical efficiency of the 14 Czech regions in the field of SDG 11 during the 2016–2023 period. This approach allows for maintaining the complexity of all SDG 11 sub-targets while ensuring sufficient discriminatory power of the model.

2 Methodology and Data

Sustainable development, and specifically SDG 11, is a highly complex area. To cover the issue in its full breadth, it is necessary to monitor a large number of indicators. Because official indicators are not fully available at the regional level, proxy indicators had to be systematically selected. The selection of these variables was based on an extensive review of existing literature focused on urban sustainability and regional efficiency. Metrics previously validated by various academic authors were included (Piña and Pardo Martínez, 2016; Chen et al, 2024; Bartniczak and Raszkowski, 2019; Kourtit et al, 2020), as well as the official methodological framework of Agenda 2030 (Ministry of the Environment of the Czech Republic, 2022). Given the multifaceted nature of the goal, the variables were first categorized into 7 thematic groups based on their affiliation with SDG 11 sub-targets (sub-targets 11.a and 11.b were not included due to their lack of relevance at the regional level). An overview of the initial variables, their assignment to sub-targets, and abbreviations are given in Table 1.

Sub-target	Indicator	Abbreviation
11.1	Completed dwellings per 1,000 mid-year population	Dwellings
	Aggregated social exclusion index	ISV
	Number of monthly salaries needed to buy 1 m^2 of a flat	Housing_Affordability
	Number of housing allowances per 1,000 population	Housing_Allowance
11.2	Road length per km^2 (total)	Roads
	Fatalities per 1,000 accidents	Road_Fatalities
	Number of traffic accidents per 100,000 population	Accidents
	Bus and train km traveled according to the timetable (km/inhabitant)	PT_Km
	Costs for domestic scheduled road transport (CZK/inhabitant)	PT_Costs
11.3	Average registered number of employees per 1,000 persons aged 15–64	Employees
	Built-up area and courtyards / total public greenery area	Urban_Greenery
	Ecological stability coefficient (as of Dec 31)	KES

Continued on next page...

Sub-target	Indicator	Abbreviation
11.4	Expenditure on monument protection and cultural heritage care (CZK/inhabitant)	Monument_Exp.
	Expenditure on nature and landscape protection (CZK/inhabitant)	Nature_Exp.
	Environmental protection investment per inhabitant	Env_Investments
11.5	Technical emergencies per 1,000 population	Tech_Emergencies
	Direct fire damages (CZK/mil. CZK GDP)	Fire_Damages
	Killed and injured persons in fires per 1,000 population	Fire_Fatalities
	Number of fires per 1,000 population	Fire_Incidents
11.6	Specific REZZO 1–4 emissions – solid pollutants (t/km^2)	Solid_Emissions
	Municipal waste production (kg/inhabitant)	Waste
	Proportion of population living in houses connected to sewerage (%)	Houses_Sewage
	Proportion of treated wastewater from water discharged into sewerage (%)	Sewage_System
11.7	Protected areas (% of the region)	Protected_Areas
	Population per 1 doctor	Doctors
	Expenditure on education (thous. CZK/pupil)	Education_Exp.
	Crime rate (per 100,000 population)	Crime_Rate
	Beds in special regime homes and homes for the elderly (per 1,000 pop. 65+)	Retirement_Homes

Table 1 Overview of selected indicators for evaluating SDG 11

As is evident from Table 1, the initial dataset contains a large number of indicators for such a complex evaluation. If they were used simultaneously for the analysis of only 14 regions, the already mentioned discriminatory power of the model would be severely compromised, as the standard rule requires the number of evaluated units to be at least twice as high as the number of variables (Aparicio et al, 2026).

For the calculations, we used the Benefit of the Doubt model that simplifies the standard input-output framework by assigning a single constant input equal to one to all units, thereby focusing exclusively on aggregating outputs (variables were, where necessary, naturally transformed to exhibit a maximization character). The model was further extended by the Andersen-Petersen concept of super-efficiency (excluding the evaluated unit from the reference set). This step was included to further increase the discriminatory power of the model, as it allows for the mutual ranking even of units that would otherwise lie on a common efficiency frontier.

According to the SDG methodology, all sub-targets should be represented in a balanced way. To combine the main advantage of the DEA method—the endogenous assignment of weights—with the need for a multidimensional assessment balancing across the seven dimensions, weight restrictions were incorporated into the model. Instead of strictly fixing the contribution of each of the seven thematic sub-groups to an equal $1/7$ share, this requirement was relaxed by setting flexible lower and upper bounds. Specifically, for each sub-target, its proportional contribution to the overall efficiency score was restricted to between 5% and 30%. These bounds ensure a necessary degree of proportionality: they allow units to capitalize on their relative strengths, while guaranteeing that no single indicator is completely ignored or entirely dominates the final evaluation.

To address this dimensionality issue without permanently discarding valuable regional data, an iterative algorithm was applied, which systematically selected exactly one variable from each thematic group. This created all possible combinations of indicators, which served as input datasets of exactly 7 variables, resulting in 14,400 iterations for each observed year. In the iterative process, each combination m ($m = 1, \ldots, M$, where $M = 14,400$) defines a unique subset of selected outputs $R_m = \{r_1, \ldots, r_7\}$, such that $r_k \in C_k$ for each $k \in K$. For each separate iteration m, the partial super-efficiency score is determined by the following model:

$$
\begin{aligned}
\max \quad & h_o^m = \sum_{r \in R_m} u_r y_{ro} \\
\text{subject to:} \quad & \sum_{r \in R_m} u_r y_{rj} \leq 1 && \forall j \in J,\ j \neq o \\
& L_k \sum_{r \in R_m} u_r y_{ro} \leq u_r y_{ro} \leq U_k \sum_{r \in R_m} u_r y_{ro} && \forall k \in K,\ r \in R_m \cap C_k \\
& u_r \geq 0 && \forall r \in R_m
\end{aligned} \tag{1}
$$

where J represents the set of all evaluated units (DMUs), $K = \{1, \ldots, 7\}$ is the set of thematic sub-groups of outputs, and C_k denotes the set of all variables belonging to the k-th thematic group. For a given iteration m, R_m represents the specific combination of exactly seven selected output variables. The term y_{rj} represents the value of the r-th output for the j-th unit, and u_r denotes the optimal weight assigned to that output. The parameters L_k and U_k define the lower and upper bounds for the proportional share of the k-th group (represented by its single selected variable) in the total virtual output; in our study, these are relaxed to $L_k = 0.05$ and $U_k = 0.30 \quad \forall k \in K$.

The partial results h_o^m obtained from all M iterations are subsequently aggregated by calculating the arithmetic mean to determine the final composite efficiency score for a given region in a specific year:

$$
H_o = \frac{1}{M} \sum_{m=1}^{M} h_o^m \tag{2}
$$

3 Results and Discussion

The results of the modified iterative model are shown in Figure 1.

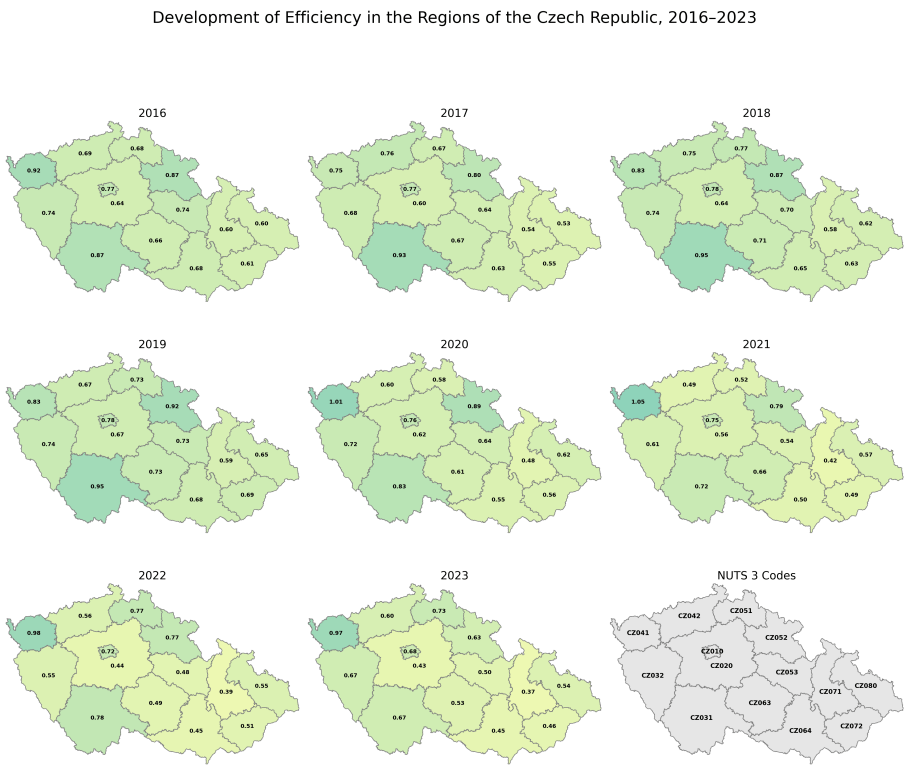


Figure 1 Development of efficiency of Czech regions in the years 2016–2023 with a NUTS 3 reference map.

The results indicate that the South Bohemian Region (CZ031) maintained the leading position until 2020, after which it was overtaken by the Karlovy Vary Region (CZ041). Interestingly, Prague (CZ010) ranks fourth in most years, just behind the Hradec Králové Region (CZ052). The Moravian regions consistently rank at the bottom of the distribution. The Olomouc (CZ071) and South Moravian (CZ064) regions share, especially in the last observed years, the last three spots in the ranking together with the Central Bohemian Region (CZ020). A pronounced

west-east divide is evident nationwide, with western regions consistently demonstrating higher efficiency than those in the east throughout the monitored period.

To identify the strengths and weaknesses of the regions, the average weights of individual indicators are displayed in Figure 2 using 2023 as a representative year.

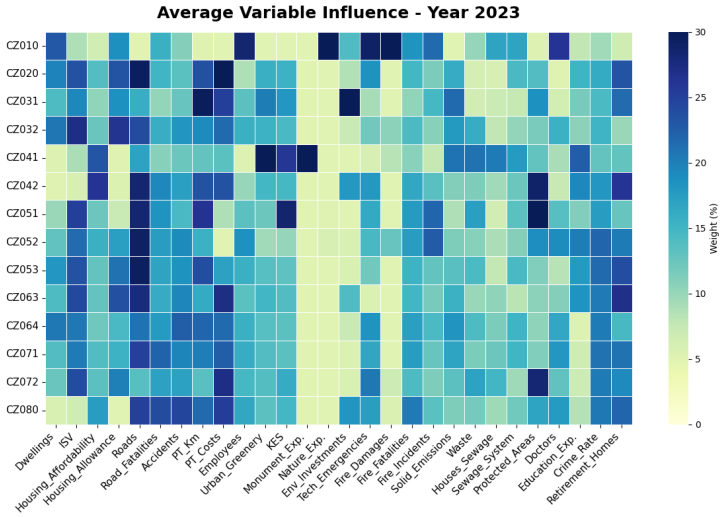


Figure 2 Average influence of individual variables on the overall evaluation of regions in 2023

A detailed analysis of the weights confirmed the long-term uniqueness of the capital city. While the DEA model assigned high weights to similar variables across most regions, Prague (CZ010) deviates significantly from this pattern. Variables that were assigned weights approaching the upper limit of the allowed interval in most regions are minimized in Prague's case, and vice versa. This phenomenon confirms the distinctness of this region from the rest of the Czech Republic. This structural uniqueness logically aligns with the specific metropolitan character of Prague, which naturally possesses a completely different socio-economic and demographic baseline than standard administrative regions.

The specific weight structure showed a contrast between Prague and other regions; however, its uniqueness did not lead to achieving the highest level of efficiency. This structural divergence highlights the non-uniform nature of development paths, where metropolitan capitals follow entirely different trajectories as mentioned in Jamoussi et al (2022). Specifically, Prague's unique position can be interpreted through the lens of a "double delinking" perspective. While large urban agglomerations strive to decouple environmental degradation from both economic wealth and physical scale, they face complex trade-offs that heavily alter their efficiency dynamics (Kourtit et al, 2020). This explains why Prague's exceptional metropolitan profile does not automatically reach the environment-driven frontier of the Karlovy Vary Region. The Karlovy Vary Region has long held the role of national leader. Since 2020, this region has continuously dominated the evaluation and thus forms the efficiency frontier.

In a detailed look at this region, it is clear that the model consistently identified large natural areas, associated low emission levels, and above-average expenditure on monument protection as its main relative strengths. The Karlovy Vary Region historically builds its development on the spa industry, the prerequisite of which is a well-maintained environment, clean air, and systematic care for cultural and historical heritage. The most efficient areas also include the South Bohemian and Hradec Králové regions. However, their dynamics are opposite. While until 2020 their efficiency increased, in the second half of the observed period, it decreased. Despite this, both regions still belong to the more successful group of DMUs even in 2023. As for the internal structure of their efficiency, each of them gains its score from different sources. For the South Bohemian Region, the main pillars are a favorable emission situation and accessible public transport. In contrast, the Hradec Králové Region builds its result primarily on completely different dominant dimensions, namely an extensive road network and a high level of security associated with low crime.

Looking closely at the structurally weakest regions, it is evident that the model repeatedly assigned the highest allowed weights to the group of indicators associated with the development and accessibility of public transport. However, it must be emphasized that high weights for these indicators define success only in terms of the region's internal profile, not in an absolute inter-regional comparison.

4 Conclusion

This study demonstrates the use of advanced optimization techniques for measuring sustainable development. From a mathematical point of view, the problem of insufficient discriminatory power of basic DEA models has been successfully overcome. The introduction of an iterative BoD model with strict weight restrictions ensured mathematical robustness and led to a full ranking of regions without the need for losing valuable input information. The observed differences in efficiency thus do not stem from the model settings but reflect actual socio-economic and environmental disparities, providing a stable basis for regional policy decision-making based on data.

Acknowledgements

This paper was supported by the grant IGA26-PEF-TP-003.

References

Aparicio J, Kapelko M, Monge JF, et al (2026) Enhancing the benefit of the doubt model through 'ensemble-dea': achieving the sustainable development goals. Expert Systems with Applications 296:129010. https://doi.org/10.1016/j.eswa.2025.129010

Bartniczak B, Raszkowski A (2019) Implementation of the sustainable cities and communities sustainable development goal (SDG) in the European Union. Sustainability 11(20):5887. https://doi.org/10.3390/su11205887

Chen S, Shao C, Yu H, et al (2024) Navigating urban human settlement sustainability: A multi-indicator assessment based on sustainable development goal 11. Journal of Cleaner Production 472:143509. https://doi.org/10.1016/j.jclepro.2024.143509

Guo Y, Yu MM, See KF (2024) Developing a sustainable development goals index for OECD countries: An effectiveness-based hierarchical data envelopment analysis. Environmental Science & Policy 160:103836. https://doi.org/10.1016/j.envsci.2024.103836

Issever Grochová L, Litzman M (2021) The efficiency in meeting measurable sustainable development goals. International Journal of Sustainable Development & World Ecology 28(8):709–719. https://doi.org/10.1080/13504509.2021.1882606

Jamoussi HBO, Hichri SM, Nijkamp P, et al (2022) Territorial development at the crossroads of attractiveness and sustainability. Journal of Sustainable Development 15(6):24. https://doi.org/10.5539/jsd.v15n6p24, https://ccsenet.org/journal/index.php/jsd/article/view/0/47910

Klarin T (2018) The concept of sustainable development: From its beginning to the contemporary issues. Zagreb International Review of Economics and Business 21(1):67–94. https://doi.org/10.2478/zireb-2018-0005

Kourtit K, Nijkamp P, Suzuki S (2020) Are global cities sustainability champions? a double delinking analysis of environmental performance of urban agglomerations. Science of The Total Environment 709:134963. https://doi.org/10.1016/j.scitotenv.2019.134963

Mardani A, Zavadskas EK, Streimikiene D, et al (2017) A comprehensive review of data envelopment analysis (DEA) approach in energy efficiency. Renewable and Sustainable Energy Reviews 70:1298–1322. https://doi.org/10.1016/j.rser.2016.12.030

Ministry of Regional Development of the Czech Republic (2025) Homepage. https://mmr.gov.cz/cs/uvod, (In Czech). Accessed: 2025-07-30

Ministry of the Environment of the Czech Republic (2022) Sustainable development goals: SDG-data. Available online: https://www.sdg-data.cz/, accessed: 2026-05-07

Nežinský E (2020) DEA modely pre sociálno-ekonomickú analýzu a podporu hospodársko-politického rozhodovania. Wolters Kluwer, Prague, (In Slovak)

Office of the Government of the Czech Republic and Ministry of the Environment (2017–2025) Series of national reports and voluntary national reviews on the implementation of the 2030 agenda in the czech republic. Includes reports from 2017 (ISBN 978-80-7440-184-8), 2021, and 2025 (ISBN 978-80-7212-676-7). (In Czech)

Piña WHA, Pardo Martínez CI (2016) Development and urban sustainability: An analysis of efficiency using data envelopment analysis. Sustainability 8(2):148. https://doi.org/10.3390/su8020148

United Nations (2015) Transforming our world: the 2030 agenda for sustainable development. Tech. rep., New York

United Nations Information Centre Prague (2025) United nations information centre prague homepage. https://osn.cz/, (In Czech). Accessed: 2025-07-31

Asymmetric evaluation of experts opinions considering their relevancies to the evaluated proposals

Miroslav Hudec[1], Miljan Vučetić[2], David Ulmann [3], Hanna Kristín Skaftadóttir [4]

Abstract. In multi-criteria decision-making, criteria weights are typically assigned by experts. On the other hand, we have situations that should reflect differences in experts' competence, reliability, or prior achievements in relation to the alternatives or candidates under consideration. Addressing this issue is both important and challenging, particularly because conventional aggregation operators, such as the weighted arithmetic mean, may be insufficient for capturing such expert-dependent distinctions. In this paper, we propose an approach in which experts are divided into two categories: top experts and regular ones. Their opinions, votes, or assigned scores are aggregated by means of an asymmetric conjunction operator. Under this approach, the agreement of a majority of top experts is treated as a primary requirement, while the agreement of a majority of regular experts provides additional support. The applicability of the proposed method is demonstrated through voting and evaluation examples. The paper concludes by outlining the advantages of the approach, its limitations, and possible directions for further research.

Keywords: Alternatives, asymmetric aggregation, experts categories, experts votes

JEL Classification: C44, D70

1 Introduction

The importance or suitability of experts is usually expressed with weights. This topic is covered by a variety of articles tackling problems from different angles, see, e.g., Bilbao-Terol et al 2026; Liu et al 2015; Bolger and Rowe 2015. In this work, the focus is on matching experts with evaluating alternatives (e.g., project proposals or candidates for a higher degree), i.e., on the level of experts achievement required for each particular evaluated alternative.

For illustrative purposes, let us say that a group of experts evaluates whether to support a (scientific) project proposal for a climbing expedition to a mountain at an altitude of 6,000 meters above sea level. A similar structural case is voting on whether a candidate with a particular number of top-tier articles, supporting letters, research results, pedagogical results, and the like should be promoted to a higher degree. In the former, we might have experts with 20 years of climbing experience up to 4,000 meters and experts with 10 years of climbing experience above 8,000 meters. It might also hold for the board of experts in a scientific committee. In a business environment, the chief executive officer might ask the board of experts to evaluate marketing proposals or new target markets proposed by several marketing workers. Or generally, we might have top experts which have significantly higher achievements and regular experts exhibiting weaker achievements (but above minimal level for joining committee) than the evaluated candidate.

An immediate solution might be weights for experts. We should be careful because this is often an oversimplified solution on one side and problematic as a set of weights should be assigned. In consequent aggregation by averaging functions due to the compensative effect (especially the fully compensative arithmetic mean), the multiplication of low weight and high points compensates for the multiplication of high weight and low points Dujmović 2018.

In this work, the primary aim is on asymmetric aggregation of computing experts opinions belonging to two (or more) categories. Hence, we propose the logical aggregation *the majority of top experts,* ***and if possible****, the majority of regular experts positively evaluate or vote*. This aggregation is based on the theoretical works of asymmetric aggregation Hudec and Mesiar 2020; Bosc and Pivert 2012. A proposal or candidate is successful when the majority of top experts in the field agree, and if possible, the majority of regular experts or those not

[1] VSB-TU Ostrava, Faculty of Economics, Dept. of Systems Engineering and Informatics, Ostrava, Czech Republic, miroslav.hudec@vsb.cz
[2] VSB-TU Ostrava, Faculty of Economics, Dept. of Systems Engineering and Informatics, Ostrava, Czech Republic; VLATACOM Institute of High Technologies, Artificial Intelligence Department, Belgrade, Serbia; and Faculty of Informatics and Computing, Singidunum University, Belgrade, Serbia , miljan.vucetic@vsb.cz
[3] VSB-TU Ostrava, Faculty of Economics, Dept. of Systems Engineering and Informatics, Ostrava, Czech Republic, david.ulmann@vsb.cz
[4] Bifröst University, Dept. of Business, Bifröst, Iceland, hannak@bifrost.is

fully related to the evaluated field agree. The regular experts should also be considered because top experts might overlook some weak points. Hence, we avoid weights for each expert, only assign relevance to each group.

The remainder of this article is organized as follows. Firstly, the preliminaries of asymmetric conjunctive aggregation are introduced in Section 2, followed by the proposed model in Section 3, which covers the logic model, classification of experts, and an illustrative example. Section 4 is devoted to discussions of the model, results, and future extensions to flexible expert classifications. Finally, Section 5 concludes this work.

2 Asymmetric Conjunction Preliminaries

A project proposal or candidate should attract a relevant proportion of top experts (let us mark this subset of experts by m_1). This is a crucial requirement; at the same time, it is desirable to satisfy a relevant proportion of *regular* experts (lower achievements or from other fields). Let us mark this subset by m_2.

To meet this evaluation requirement, we adapt *AND IF POSSIBLE* function known as asymmetric conjunction Hudec and Mesiar 2020:

$$C(x, y) = \min(x, A(x, y)) \tag{1}$$

where x is the proportion of votes *yes* or points assigned by experts from the subset m_1, while y is the proportion of votes *yes* or points assigned by experts from the subset m_2, and A is an averaging aggregation function. As a typical example of *the AND IF POSSIBLE* operator is as follows Bosc and Pivert 2012

$$C(x, y) = \min(x, k \cdot x + (1 - k)y) \tag{2}$$

where $k \in (0, 1]$. For the extremal value $k = 0$, we get the conjunction expressed by the *MIN* function: $\min(x, y)$. When $k = 0.5$, the second argument becomes the arithmetic mean, i.e., the function $C(x, y) = \min(x, \frac{1}{2}(x + y))$ is obtained. For A in asymmetric conjunction (1) we can use any averaging (weighted) function. For simplicity, we adopt the weighted arithmetic mean with $0.5 < k < 1$.

Due to *the MIN* function, where the first projection covers the votes of top experts, when a proposal does not reach the threshold, it is rejected regardless of the opinions of regular experts. On the other hand, when a proposal does not reach the threshold of regular experts but fully meets the criteria of top experts, it is not rejected, but the score is reduced.

3 Proposed Model

Firstly, we develop logic aggregation. Consequently, we classify experts into two subsets: m_1 and m_2.

3.1 Logic aggregation

We start by calculating x and y (2) for the case when experts vote binary: *yes* or *no*. The proportion of votes for *yes* in set m_1 is

$$x = \frac{|m_1^+|}{|m_1|} \tag{3}$$

where $|m_1^+|$ is the number of experts (cardinality) in m_1 voting for *yes*, while $|m_1|$ is the cardinality of set m_1, so $x \in [0, 1]$. Analogously, the proportion of votes for *yes* for set m_2 is

$$y = \frac{|m_2^+|}{|m_2|} \tag{4}$$

where $|m_2^+|$ is the number of experts in m_2 voting for *yes*, so $y \in [0, 1]$.

The question is how to assign the value for k in (2). In this work, we assigned $k = 0.75$ because it indicates that the set m_1 is significantly more relevant. In the opposite case of being very lightly relevant, we have $k = 0.25$; while for $k = 0.5$, it is just more relevant.

When x *AND IF POSSIBLE* $y > 0.5$ occurs, the candidate is successful, as is the case in a usual evaluation or voting when a majority is required. Three cases for $k = 0.75$ in (2) where the simple majority (more than 50 % votes) is sufficient are shown. In Table 1, significant support by m_1 in comparison to m_2 causes a change from *no* to *yes*, while in Table 2, significant support by m_2 in comparison to low support by m_1 causes no change in decision

no. In Table 3, even though support by set m_1 is recorded to be a little above average, significantly below average support by set m_2 causes no change in decision (*no*), despite the cardinality of m_1 being higher than that of m_2.

votes / experts	**top**	**regular**	**total**
total	10	10	20
yes	8	2	10
no	2	8	10
proportion	0.8	0.2	**0.5**

Table 1 Examples of votes by two categories of experts where by arithmetic mean candidate is not successful (borderline case, no majority of votes), but with (2 for $k = 0.75$) the result is 0.65 or candidate is successful

votes / experts	**top**	**regular**	**total**
total	10	10	20
yes	2	8	10
no	8	2	10
proportion	0.2	0.8	**0.5**

Table 2 Examples of votes by two categories of experts where by arithmetic mean candidate is not successful (borderline case), but with (2 for $k = 0.75$) result is 0.2 or candidate is significantly failed

votes / experts	**above**	**below**	**total**
total	12	8	20
yes	7	1	8
no	5	7	12
proportion	0.583	0.125	**0.4**

Table 3 Examples of votes by two categories of experts where by arithmetic mean candidate is not successful, where the same decision by (2) result is 0.469

The next examined case is assigning points to the proposals. Instead of the proportion of votes, we count the proportions of points as

$$x = \frac{q}{\sum_{i=1}^{m_1} q_i} \tag{5}$$

for all experts belonging to m_1, where q is the sum of points from these experts and the denominator is the sum of all possible points of this group of experts. In the same way, we get

$$y = \frac{r}{\sum_{i=1}^{m_2} r_i} \tag{6}$$

for all experts belonging to m_2, where r is the sum of points from these experts, and the denominator is the sum of all possible points of this group of experts. The other aspects of computing (2) remain the same.

3.2 Classifying Experts

Classifying experts into two subsets is the relevant part of this voting and evaluation model. We recognize two cases: (i) proposals and candidates are evaluated only by the quality of the proposals, and (ii) candidates provide a list of supporters (opinions from additional experts that it is a reasonable proposal, e.g., for academic positions). In the latter, we divide experts by the following condition:

$$\begin{array}{ll} e < p \wedge e < s & \text{then } m_2 \\ \text{otherwise} & \text{then } m_1 \end{array} \tag{7}$$

where e is the achievement of the expert, p is the achievement of the candidate, and s is the achievement of the supporter, while in the former we have only e and p, i.e.,

$$\begin{array}{ll} e < p & \text{then } m_2 \\ \text{otherwise} & \text{then } m_1 \end{array} \tag{8}$$

We emphasize that the achievements of computing experts opinions and, therefore, their classification is a dynamic process, i.e., adjusted to each candidate or proposal.

3.3 Illustrative example

Let us have two candidates c_1 and c_2 who provided supporting letters from two experts s_1 and s_2, respectively. The board consists of seven experts e_1, ... e_7. The achievements in four fields are a_1, a_2, a_3, and a_4. Data related to the achievements are in Table 4, while the division of experts into two subsets based on (7) is in Table 5. Votes are shown in Table 6 where expert e_7 has not provided a vote, or this vote has been marked as non-valid.

	a_1	a_2	a_3	a_4
c1	20	7	8	10
c2	7	7	6	8
s1	26	12	12	15
s2	6	4	9	10
e1	25	15	15	16
e2	18	9	9	10
e3	21	10	9	12
e4	16	4	10	6
e5	7	3	9	6
e6	6	4	10	9
e7	8	3	10	6
min	**5**	**2**	**8**	**6**

Table 4 Candidates, experts, supporters and their achievements

	p_1		p_2
e1	m1	e1	m1
e2	m1	e2	m1
e3	m1	e3	m1
e4	m2	e4	m1
e5	m2	e5	m2
e6	m2	e6	m1
e7	m2	e7	m2

Table 5 Dividing experts into two subsets based on (7)

	p_1	p_2
e1	yes	no
e2	yes	no
e3	no	no
e4	yes	yes
e5	no	yes
e6	no	yes
e7	-	yes

Table 6 Votes for candidates

Applying (2) for $k = 0.75$, we obtain the final scores. Candidate c_1 receives 0.5625 (above 0.5, therefore accepted), while considering the non voting expert as a non-supporter or neutral; whereas candidate c_2 receives 0.4 and is rejected. Experts with higher achievements prefer candidate c_1, which is reflected in the decision.

4 Discussion

This proposal addresses the following cases:

- **voting candidates independently** - each candidate who reaches the required proportion of votes (usually higher than 0.5) is selected,
- **voting candidates in a competitive way** - the candidate who reaches the highest proportion is elected, or several highly ranked candidates are selected.

This model mitigates but does not fully cover cases when non-presented experts (or experts who did not provide their opinion) are counted. For instance, if the number of experts on the university board is 50, while the number of voting experts (present at the meeting) is 26, the candidate should persuade 100 % of them; however, when all experts are present, the candidate should persuade 52 % of them. That is, the candidate succeeds with 24 votes against. It is a topic for future research to assign this observation to asymmetric conjunction.

Concepts of flexibility and uncertainty in general voting, such as *yes/no* responses and intensities of *maybe*, have been explored in the Basel (Switzerland) experiment Schwarz et al 2024. Similarly, the division of stakeholders into subgroups and the handling of uncertainty in their support for proposals have been applied in an urban planning context in Geuensee, Lucerne (Switzerland) Emmenegger et al 2025.

The next question is the crisp classification of experts into sets m_1 and m_2. It might be problematic for experts near the border of the two sets. The solution is formalizing membership in sets m_1 and m_2. The model is robust; only the computation of values x and y is modified. In classical sets, we count elements or sum the intensities of membership, which are always equal to 1. Hence, we can express voting (3) as

$$x = \frac{\sum_{i=1}^{m_1} \phi(e_i) * v_i}{\sum_{i=1}^{m_1} \phi(e_i)} \tag{9}$$

where $\phi(e_i)$ indicates the crisp belonging of expert e_i to set m_1, and v_i is the expert vote; i.e., *yes* gets a value of 1, while *no* gets a value of 0. Apparently, for fuzzy sets, we get

$$x = \frac{\sum_{i=1}^{m_1} \mu(e_i) * v_i}{\sum_{i=1}^{m_1} \mu(e_i)} \tag{10}$$

where $\mu(e_i)$ indicates the matching degree (or intensity of belonging) of expert e_i to set m_1. The other parameters have the same meaning. Analogously, this holds for set m_2. To avoid inconsistent voting, the sum of intensities belonging to two fuzzy sets m_1 and m_2 should equal 1. It is one of the criteria when formalizing fuzzy sets and the overlapping of neighboring fuzzy sets; see, e.g., Alonso et al 2021.

A further extension introduces flexibility by allowing experts to express their preferences beyond binary decisions, indicating a degree of inclination towards the affirmative option, where $v_i = 1$ denotes a definite *yes*, $v_i = 0$ denotes a definite *no*, and intermediate values $v_i \in (0, 1)$ capturing levels of support. In this case, $v_i \in \{0, 1\}$ are replaced by $v_i \in [0, 1]$ in (9) for crisp categories m_1 and m_2 or in (10) when categories m_1 and m_2 overlap.

Another idea could be the Dempster-Shafer method Chen et al 2014. Instead of *stating that when x and if possible y > 0.5 the candidate is successful*, a fixed threshold of 0.5 can be replaced with the outputs: *accepted*, *undecided*, and *rejected*. That might solve the problem of borderline cases (about 0.5), non voting experts, and the relevance of experts.

The future work should also consider cases when non-present experts are counted indirectly. If all members are present, a candidate for a higher degree or a proposal for a project should persuade slightly more than 50% of experts. When the number of attending experts is slightly above 50% of the total number of experts, the candidate should persuade almost all of them, regardless of whether the experts belong mainly to set m_1 or m_2. We might record such votes as *Maybe* and apply the Dempster-Shafer method of evaluation.

A further extension of the proposed model should concern the increasing use of AI-based tools in expert evaluation processes. AI systems may be used to summarize documents, compare proposals, identify inconsistencies, or provide preliminary assessments of quality, novelty, or relevance. However, such tools should not be treated as equivalent to human experts. AI systems lack responsibility, institutional accountability, and contextual

understanding of the case under evaluation. Their outputs may provide useful diagnostic support, but they should not override the judgment of experts. This distinction can be naturally represented through an asymmetric conjunction. In addition to distinguishing between top experts and regular experts, the model may include AI-based evaluative tools as a third auxiliary category. The corresponding aggregation logic could be expressed as: *the proposal should be supported by the majority of top experts and, if possible, by the majority of regular experts and, if possible, by AI-based evaluative tools*. In this construction, AI-generated assessments can strengthen or weaken the overall evaluation only in a secondary sense. This is particularly relevant in contexts where AI tools are used to assist evaluation, but where final decisions must remain grounded in human expertise, transparency, and accountability.

5 Conclusions

This work proposes the logic of aggregating the experts votes or opinions, considering experts achievements relative to the evaluated candidates or proposals or their achievements in the field. The *and if possible* asymmetric aggregation has benefits when the goal is to ensure that agreement among top experts is more relevant while also considering the opinions of regular experts. This model also works when the border between these two categories of experts is vague, where intensities of belonging to these overlapping categories are included in the aggregation.

In the case of the overlapping boundary between sets m_1 and m_2, as well as in extension to flexible voting, i.e., expressing inclinations towards *yes* or *no*, the basis of the asymmetric model remains the same; the counting of votes changes to include the counting of intensities of belonging to sets and inclinations of answers. This model can be further extended by AI based evaluation, which can strengthen or weaken the overall evaluation only in a secondary sense. A relevant aspect is sensitivity analysis, which is together with real-world applications left for future research. The main task was proposing model which can be developed into diverse directions.

Acknowledgements

This article has been produced with the financial support of the European Union under the REFRESH – Research Excellence For REgion Sustainability and High-tech Industries project number CZ.10.03.01/00/22_003/0000048 via the Operational Programme Just Transition and the SGS project no. SP2026/071.

References

Alonso J, Castiello C, Luis M, et al (2021) Explainable fuzzy systems: Paving the way from interpretable fuzzy systems to explainable AI systems. Springer, Cham

Bilbao-Terol A, Cañal-Fernández V, González-Pérez C (2026) Improving the assessment of social and environmental impacts of re projects for financial institutions. Operational Research 26(2):47

Bolger F, Rowe G (2015) The aggregation of expert judgment: Do good things come to those who weight? Risk Analysis 35(1):5–11

Bosc P, Pivert O (2012) On four noncommutative fuzzy connectives and their axiomatization. Fuzzy sets and Systems 202:42–60

Chen Q, Whitbrook A, Aickelin U, et al (2014) Data classification using the dempster–shafer method. Journal of Experimental & Theoretical Artificial Intelligence 26(4):493–517

Dujmović J (2018) Soft Computing Evaluation Logic: The LSP Decision Method and Its Applications. IEEE-John Wiley & Sons, Hoboken

Emmenegger B, Zapletal F, Hudec M, et al (2025) Weighted multi-criteria decision-making with coalition strategies: a framework for citizens public participation: B. emmenegger et al. Central European Journal of Operations Research pp 1–21

Hudec M, Mesiar R (2020) The axiomatization of asymmetric disjunction and conjunction. Information Fusion 53:165–173

Liu D, Xu W, Du W, et al (2015) How to choose appropriate experts for peer review: An intelligent recommendation method in a big data context. Data Science Journal 14:16–16

Schwarz D, Emmenegger B, Fivaz J, et al (2024) Effects of fuzzy voting on citizen participation: empirical evidence from switzerland. In: 2024 Tenth International Conference on eDemocracy & eGovernment. IEEE, pp 1–8

Spatial Patterns of Household Internet Access in EU Regions: An ESDA Approach

Michaela Chocholatá[1]

Abstract. This paper uses Exploratory Data Analysis (EDA) and Exploratory Spatial Data Analysis (ESDA) to analyse regional disparities in household internet access across European Union regions in 2025. The aim of the paper is to identify spatial patterns, evaluate spatial autocorrelation, and detect clusters of regions with similar levels of internet accessibility. The analysis is based on Eurostat regional data and was carried out using the GeoDa software environment. The methodological framework combines descriptive visualisation techniques with spatial statistical methods, including the Moran's *I* statistic, the Moran scatterplot and Local Indicators of Spatial Association (LISA). The results reveal significant territorial disparities in household internet access and confirm statistically significant positive spatial autocorrelation (Moran's I ≈ 0.441), indicating moderate to strong spatial clustering of internet accessibility across EU regions. High-access regions are concentrated in Northern and Western Europe, while low-access clusters are primarily located in Southern and Eastern Europe. These findings suggest that digital inequalities within the European Union have a strong geographical dimension, highlighting the importance of spatial analytical approaches for evaluating regional digital cohesion and territorial disparities.

Keywords: spatial autocorrelation, EU regions, household internet access, regional disparities

JEL Classification: C21, R12

1 Introduction

In the 21st century, digital connectivity constitutes a fundamental determinant of economic competitiveness, social inclusion, educational attainment, and regional resilience. Home internet access is increasingly recognized not simply as a technological convenience, but as an essential condition for participation in the digital economy and contemporary society. The COVID-19 pandemic notably intensified the relevance of digital infrastructure and exposed disparities in the capacity of households and regions to engage in remote work, online education, digital public services, and e-commerce.

Within the European Union (EU), digital transformation represents one of the strategic priorities of long-term socio-economic development. The European Commission emphasizes digital inclusion through initiatives such as the Digital Decade 2030 (European Parliament and Council of the European Union, 2022), stressing equal access to digital infrastructure and reducing territorial disparities. Despite substantial progress in broadband infrastructure and digital policy implementation, significant regional disparities in household internet access persist across the EU. In this context, spatial analytical methods provide valuable tools for identifying regional patterns and spatial dependence in digital development indicators. Traditional descriptive statistics alone are insufficient for capturing the geographical dimension of socio-economic phenomena, as they do not account for spatial dependence and spatial interactions between regions (Anselin, 1995; Cliff & Ord, 1981; Furková, 2016; Tislenko, 2024; Kazar & Kazar, 2026). Therefore, Exploratory Data Analysis (EDA) and Exploratory Spatial Data Analysis (ESDA) enable a deeper understanding of the spatial concentration, clustering, and autocorrelation of household internet access across regions.

The issue of digital inequalities and the digital divide have attracted growing attention in regional science and socio-economic research. Previous studies have shown that disparities in internet accessibility are closely related to regional economic development, infrastructure quality, income levels, and technological readiness (Billon et al., 2009; Vicente & López, 2011). More recent research confirms that substantial digital inequalities persist across

[1] Bratislava University of Economics and Business, Faculty of Economic Informatics, Department of Operations Research and Econometrics, Dolnozemská cesta 1, 852 35 Bratislava, michaela.chocholata@euba.sk.

the EU despite ongoing digital transformation policies. Recent studies point to considerable differences in digital inclusion between Northern and Western European countries and several Southern and Eastern European states, highlighting the continuing relevance of territorial digital disparities (Gomes & Dias, 2025). De Clercq et al. (2023) demonstrated that regional disparities in broadband accessibility continue to persist across European regions, particularly between urban and rural territories, with important implications for regional economic growth and territorial cohesion. Tislenko (2024) confirms that digital inequalities in the European Union exhibit clear spatial patterns and are influenced by neighbourhood effects, as demonstrated by spatial autocorrelation analysis using Moran's *I*. Kazar and Kazar (2026) highlight that digitalisation processes in the European Union are strongly affected by spatial spillovers, indicating that regional digital performance is not independent but influenced by neighbouring regions, thereby reinforcing the core–periphery divide.

Despite the growing body of literature on digital inequalities, relatively limited attention has been devoted to the explicit analysis of spatial clustering of household internet access at the regional level using ESDA techniques. Most existing studies focus either on national-level comparisons or on determinants of digitalisation, while the spatial dimension of regional clustering at NUTS 2 level remains underexplored.

The aim of this paper is to analyse regional disparities in household internet access across EU regions in 2025 using EDA and ESDA techniques. The study focuses on identifying spatial concentration patterns, evaluating the existence of spatial autocorrelation, and detecting significant clusters of regions with high and low levels of internet access. The analysis is based on Eurostat regional data and implemented in the GeoDa software environment. By incorporating spatial analytical methods, the paper contributes by explicitly analysing spatial clustering and local spatial patterns of digital accessibility at the NUTS regional level within the EU, thereby providing a clearer understanding of the territorial structure of digital inequalities.

This paper is organized as follows. Section 2 presents the methodological framework and describes the applied exploratory and spatial analytical methods; section 3 describes the data used for analysis. Section 4 provides the empirical analysis and discusses the obtained results of EDA and ESDA techniques. The final section concludes the paper and outlines possible directions for future research.

2 Methodology

The application of EDA and ESDA methods is particularly appropriate in regional analysis, where spatial dependence and geographical heterogeneity play a crucial role. Unlike traditional statistical approaches, ESDA enables the identification of spatial patterns without imposing restrictive model assumptions, making it especially suitable for exploratory analysis of regional disparities (Anselin, 1995). Building upon the theoretical background related to regional digital inequalities and spatial dependence, the empirical part of the paper applies exploratory and spatial analytical methods in order to identify spatial patterns in household internet access across EU regions.

The methodological framework of the paper is based on the application of EDA and ESDA, which together provide a comprehensive approach for examining both the statistical and spatial characteristics of regional data. EDA represents an important preliminary step in empirical analysis, enabling the identification of the basic structure of the dataset, variability, distributional properties, and potential outliers through descriptive statistics and graphical visualization techniques (Tukey, 1977). In regional studies, commonly applied EDA tools include histograms, boxplots, and choropleth maps, which facilitate the interpretation of territorial disparities and spatial heterogeneity among observed regions.

While EDA focuses primarily on the statistical distribution of variables, ESDA extends the analytical framework by explicitly incorporating spatial relationships among territorial units (Anselin, 1995; Furková, 2016; Kopczewska, 2021; Tislenko, 2024; Kazar & Kazar, 2026). ESDA methods enable the examination of spatial dependence, spatial clustering, and spatial outliers, thereby providing a deeper understanding of geographically conditioned socio-economic processes. The incorporation of spatial effects is particularly important in regional analysis because neighbouring regions often exhibit similar developmental characteristics as a consequence of geographical proximity and mutual interactions.

A key instrument of ESDA is the analysis of spatial autocorrelation, which evaluates the extent to which similar values are spatially concentrated across a study area. In this context, the global Moran's *I* statistic is among the most widely used indicators for measuring spatial autocorrelation in regional and socio-economic research (Moran, 1950; Cliff & Ord, 1981; Kopczewska, 2021):

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} \quad (1)$$

where x_i represents the observation in the region i, $\overline{x}$ denotes the average of all regions studied, N is the number of regions in the data set and w_{ij} are the elements of the spatial weight matrix **W** of dimension $N \times N$.[2]

Moran's I assesses whether neighbouring regions tend to display similar or dissimilar values of the analysed variable compared to a random spatial distribution. Positive values of Moran's I indicate positive spatial autocorrelation and the existence of spatial clustering, whereas negative values suggest spatial dispersion.

An important complementary tool within ESDA is the Moran scatterplot, which graphically illustrates the relationship between standardized values of the analysed variable and their corresponding spatial lag. The scatterplot enables the identification of different types of spatial association among regions, specifically High–High, Low–Low, High–Low, and Low–High spatial relationships. These categories facilitate the interpretation of regional clustering processes and the identification of territories that significantly differ from their neighbouring regions.

To complement the global assessment of spatial autocorrelation, Local Indicators of Spatial Association (LISA) were applied in order to identify statistically significant local clusters and spatial outliers (Furková, 2016). LISA cluster maps therefore provide a more detailed understanding of the territorial structure of regional disparities and spatial heterogeneity within the EU.

The combination of EDA and ESDA methods provide a comprehensive analytical framework for examining regional disparities, as it enables the simultaneous assessment of statistical variability and spatial dependence among territorial units.

3 Data

The analysis is based on the Eurostat indicator "*Households with access to the internet at home*" (in %) for the year 2025 (Eurostat, 2025). The spatial dataset consists of 194 territorial units of the EU. The dataset is primarily composed of NUTS 2 regions; however, Germany and Greece are represented at the NUTS 1 level due to the unavailability of comparable NUTS 2 data. In order to ensure the construction of a territorially contiguous spatial weights matrix[3] in GeoDa, island and geographically isolated regions were excluded from the analysis.

4 Empirical Results

The initial EDA revealed relatively high levels of household internet access across the majority of EU regions. Nevertheless, the results indicate the continued existence of considerable regional disparities. The spatial distribution of the analysed variable is presented in Figure 1, which illustrates the geographical differentiation of household internet accessibility across EU regions.

The box map demonstrates visible geographical disparities in household internet access across EU regions. Higher levels of internet accessibility are predominantly concentrated in Northern and Western Europe, whereas comparatively lower values are more frequently observed in parts of Southern and Eastern Europe. The lowest values were identified in Greek and Croatian regions, respectively.

At the same time, the spatial distribution also suggests the existence of differences between metropolitan and more peripheral or rural regions. Capital city regions and highly urbanised areas generally achieve higher levels of household internet access, which may be associated with better digital infrastructure, higher population density, stronger economic performance, and greater technological readiness. In contrast, lower values are more frequently observed in less urbanised and geographically peripheral territories, where weaker infrastructure and lower digital inclusion may persist.

The observed pattern therefore indicates that digital inequalities within the EU are shaped not only by broader east–west and north–south territorial differences, but also by the urban–rural dimension of regional development. This finding is consistent with recent studies emphasizing the persistence of the European urban–rural digital divide and the unequal spatial distribution of digital infrastructure across regions (de Clercq et al., 2023; Kazar & Kazar, 2026).

[2] For more information about the spatial weights matrix **W** see e.g., Furková (2016), Kopczewska (2021).
[3] Spatial weights matrix of queen type was used – see e.g., Furková (2016), Kopczewska (2021).

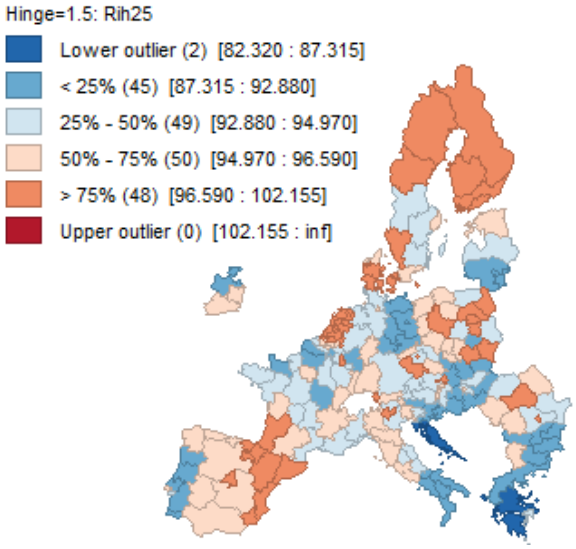


Figure 1 Spatial distribution of households with internet access at home across EU regions in 2025, Source: own calculation in GeoDa

The descriptive statistical characteristics of the dataset are further illustrated by the boxplot presented in Figure 2.

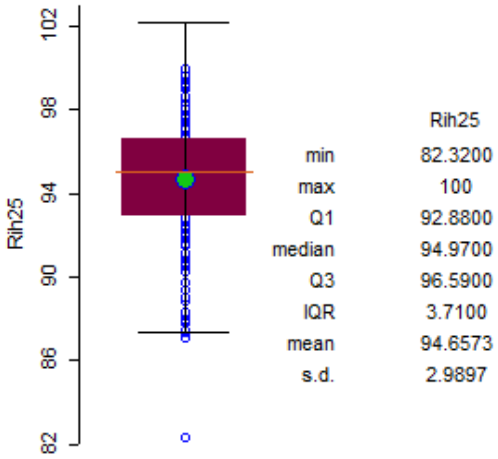


Figure 2 Boxplot of household internet access across EU regions, Source: own calculation in GeoDa

The boxplot indicates that the majority of EU regions achieve relatively high levels of household internet access, with the median value reaching approximately 94.97%. The interquartile range suggests relatively moderate variability among regions, while the overall distribution remains concentrated at higher levels of internet accessibility. At the same time, the boxplot reveals the presence of several regions with substantially lower values, indicating potential outliers and the persistence of territorial disparities in digital accessibility. The observed distribution therefore confirms that, despite considerable progress in digital infrastructure development within the EU, regional heterogeneity in household internet access continues to exist.

To further investigate the geographical dimension of regional disparities, ESDA techniques were subsequently applied. The analysis of global spatial autocorrelation was conducted using Moran's *I* statistic. The obtained Moran scatterplot is presented in Figure 3 (*left*).

The Moran scatterplot demonstrates a positive relationship between standardized regional values and their spatial lag, indicating the presence of positive spatial autocorrelation. A substantial proportion of regions are concentrated within the *High–High* and *Low–Low* quadrants, suggesting that neighbouring regions tend to exhibit similar levels of household internet access. The calculated value of Moran's *I* reached approximately 0.441, indicating moderate to strong positive spatial autocorrelation among EU regions. This result implies that regions with high internet accessibility are geographically concentrated near other highly digitalized regions, while regions with lower

internet access tend to cluster together spatially. The statistical significance of the observed spatial autocorrelation was verified using a permutation test[4].

To identify local spatial patterns, Local Indicators of Spatial Association (LISA) were subsequently applied. The resulting LISA cluster map is presented in Figure 3 (*right*). The LISA analysis identifies several statistically significant spatial clusters. This spatial pattern can be explained by the concentration of economic activity, higher income levels, and more advanced digital infrastructure in Northern and Western Europe, which reinforce spatial spillover effects and regional convergence in digital development.

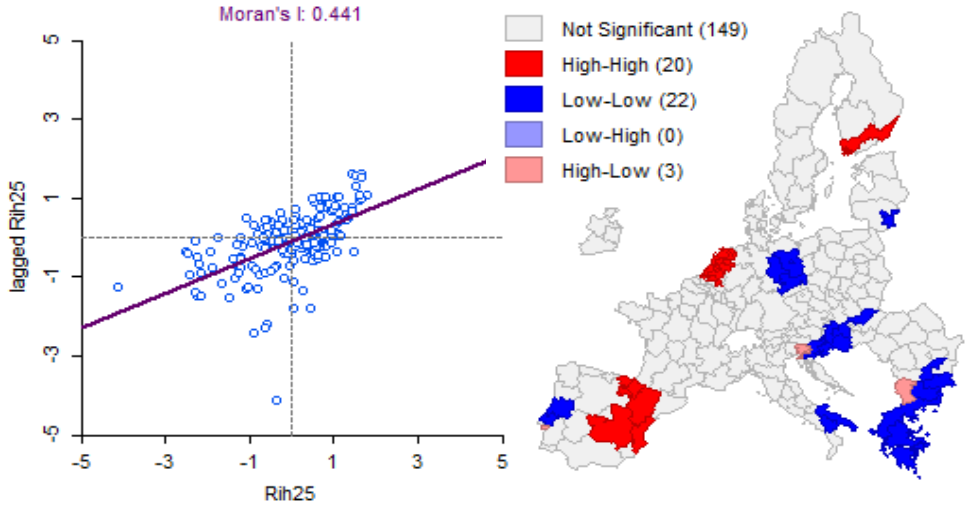


Figure 3 Moran scatterplot (*left*) and LISA cluster map (*right*) of household internet access across EU regions, Source: own calculation in GeoDa

High–High clusters are predominantly concentrated in parts of Western Europe and the Iberian Peninsula, particularly in Spain. Additional highly digitalized regions are also visible around major urban and metropolitan areas. These regions are generally associated with stronger digital infrastructure, higher economic performance, and greater technological readiness. Conversely, *Low–Low* clusters are predominantly located in parts of South-Eastern Europe, especially in Greece and neighbouring regions characterized by below-average levels of household internet access. These spatial concentrations suggest the persistence of geographically conditioned digital disadvantages related to weaker infrastructure, lower digital inclusion, and slower technological diffusion. The LISA map further indicates that digital inequalities within the EU are not distributed randomly but instead form relatively compact territorial clusters.

The analysis also identified a limited number of High–Low spatial outliers, namely the Bulgarian region "Yugozapaden", the Slovenian region "Zahodna Slovenija", and the Portuguese region "Península de Setúbal". These regions are characterized by above-average levels of household internet access despite being surrounded by neighbouring territories with comparatively lower digital accessibility. The identified regions are closely connected to important metropolitan and economically advanced urban areas, particularly Sofia, Ljubljana, and the Lisbon metropolitan area.

The presence of these spatial outliers suggests a concentration of digital infrastructure, economic activity, and technological readiness within major urban regions, while surrounding territories continue to exhibit lower levels of digital accessibility. This pattern may indicate uneven territorial diffusion of digital development and the persistence of metropolitan advantages in household internet access across the EU.

Overall, the empirical findings confirm the existence of significant spatial clustering and territorial polarization in household internet access across EU regions. The results demonstrate that digital inequalities cannot be analysed independently of geographical context, as neighbouring regions tend to share similar levels of digital accessibility and digital development.

[4] The results are available upon request.

5 Conclusion

This paper analysed regional disparities in household internet access across EU regions in 2025 using exploratory and spatial analytical methods. The results revealed significant territorial differences and confirmed the existence of positive spatial autocorrelation among EU regions.

The findings indicate that regions with high levels of household internet access tend to form spatial clusters, particularly in Northern and Western Europe, while lower levels of digital accessibility remain concentrated in selected Southern and Eastern European regions. The identified spatial patterns confirm that digital inequalities within the European Union possess a strong geographical dimension.

The study also demonstrated the usefulness of EDA and ESDA methods for analysing regional digital disparities, as these approaches enable the identification of spatial clustering and territorial heterogeneity that cannot be fully captured by conventional statistical methods alone. By incorporating spatial relationships among regions, the analysis provides a more comprehensive understanding of the territorial structure of digital inequalities and the geographical concentration of digital development within the EU. The existence of spatial clusters suggests that regional digital development is influenced by neighbourhood effects, where the characteristics of surrounding regions significantly affect local outcomes.

The main contribution of the paper lies in providing a spatially explicit analysis of digital inequalities, demonstrating that regional disparities are not randomly distributed but form statistically significant spatial clusters across the EU. From a policy perspective, the identified spatial concentration of digital inequalities suggests that regional digital policies should not be designed uniformly across the EU. Instead, geographically targeted interventions may be necessary in regions characterised by persistently lower internet accessibility and weaker digital infrastructure. In this context, the results may contribute to the evaluation of EU digital cohesion policies and support the design of more effective measures aimed at reducing regional digital inequalities and strengthening inclusive digital development across the EU. Further studies could also apply spatial econometric models and dynamic analyses in order to examine the determinants and evolution of regional digital inequalities over time.

Acknowledgements

This work was supported by the Grant Agency of Slovak Republic – VEGA 1/0435/26 “Investment Effectiveness and Spatial Spillover Effects in the EU Green Transformation: A Spatial Econometric and Machine Learning Approach”.

References

Anselin L (1995) Local indicators of spatial association-LISA. Geographical Analysis 27(2):93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Billon M, Marco R, Lera-Lopez F (2009) Disparities in ICT adoption: A multidimensional approach to study the cross-country digital divide. Telecommunications Policy 33(10–11): 596–610. https://doi.org/10.1016/j.telpol.2009.08.006

Cliff A D, Ord J K (1981) Spatial processes: Models and applications. Pion, London

de Clercq M, D’Haese M, Buysse J (2023) Economic growth and broadband access: The European urban-rural digital divide. Telecommunications Policy 47(6):102579. https://doi.org/10.1016/j.telpol.2023.102579

European Parliament and Council of the European Union (2022) Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030. Official Journal of the European Union L323:4–26. https://eur-lex.europa.eu/eli/dec/2022/2481/oj

Eurostat (2025) Households with access to the internet at home by NUTS 2 regions [Data set]. Eurostat Database. https://ec.europa.eu/eurostat/databrowser/view/isoc_r_iacc_h/default/table?lang=en

Furková A (2016) Spatial pattern of innovative activity in the EU regions: Exploratory spatial data analysis and spatial econometric approach. In: Gomes O (ed) Advances in Applied Business Research: The L.A.B.S. Initiative. Nova Science Publishers, New York, pp 159–186

Gomes A, Dias J G (2025) Digital Divide in the European Union: A Typology of EU Citizens. Soc Indic Res 176: 149–172. https://doi.org/10.1007/s11205-024-03452-2

Kazar G, Kazar A (2026) Path dependence and spatial spillovers in regional digitalization: Evidence from dynamic spatial panel analysis in Europe. Sustainability 18(10): 4839. https://doi.org/10.3390/su18104839

Kopczewska K (2021) Applied Spatial Statistics and Econometrics: Data Analysis in R. Routledge, London

Moran P A P (1950) Notes on continuous stochastic phenomena. Biometrika 37(1–2):17–23

Tislenko M I (2024) Digital divide in the European Union: Assessing spatial disparities and neighborhood effects. Journal of the Geographical Institute "Jovan Cvijić" 74(2):181–194. https://doi.org/10.2298/IJGI2402181T

Tukey J W (1977) Exploratory data analysis. Addison-Wesley, Reading

Vicente M R, López A J (2011) Assessing the regional digital divide across the European Union-27. Telecommunications Policy 35(3):220–237. https://doi.org/10.1016/j.telpol.2010.12.013

Minimal Change of Ambulance Deployment to Reach Given Quality of Service

Jaroslav Janáček[1], Marek Kvet[2]

Abstract. This paper reformulates the original problem of optimal ambulance location with the aim of minimizing the average response time of the associated emergency medical care system. The reformulation consists in taking into account the existing ambulance location, which should be changed as little as possible. In this way, the change in the original fleet location is minimized, assuming that a given reduction in the average response time of the system is achieved. Since the reduction is limited by the optimal average response time obtained by unrestricted reallocation of the ambulance fleet, the prescribed value of the response time reduction is chosen from values between the current and the optimal response times. The computational study indicates the magnitude of the necessary changes in the current ambulance location and the required improvement in the average response time.

Keywords: Location science, mathematical modelling, optimal ambulance deployment, quality of service access

JEL Classification: C44, C61

1 Introduction

The Emergency Medical Service (EMS) is instituted to deliver immediate pre-hospital healthcare in critical circumstances where the lives or health of individuals is at risk. This service is typically offered by private EMS companies authorized by their licenses or permits to operate a medical service center. EMS is a component of the integrated rescue system, which is part of the broader category of public service systems (Brotcorne et al., 2003; Chanta et al., 2014; Doerner et al., 2005; Ingolfsson et al., 2008; Karatas and Yakıcıa E, 2019).

The Integrated Rescue System (IRS) is a group of different rescue services that work together to quickly gather information, activate rescue forces, and make sure they are used and coordinated properly to provide the help that is needed in case of a threat to life, health, property, or the environment. The IRS's principal job is to quickly and professionally help the entity affected by emergency. A number of rescue services offer the help that was mentioned. Fire and rescue officers, EMS providers, control chemical labs, mountain and mine rescue services, and other groups are the main parts of IRS. The Coordination Center or the Emergency Operations Center coordinates all of these issues to give professional, medical, technical, and other help in an emergency. However, in our research we will limit ourselves to ground emergency medical services.

From the perspective of clients' ability to get urgent health care, this article examines the present deployment of service centers in selected regions of Czechia and Slovakia. Evidently, the service's accessibility improves in direct proportion to the proximity of the nearest available facility, from which an ambulance is dispatched to meet the patient's need (Jánošíková, 2007; Jánošíková et al., 2021; Reuter-Oppermann et al., 2017). There is a subset of potentially fatal illnesses that must be addressed in the quest for the best places to put emergency medical services stations. Serious diagnoses like heart attack, stroke, unconsciousness, and many more are included in the first hour quintet (FHQ). Paramedics are insufficient for the adequate rescue and treatment of patients with these diagnoses; a doctor must be present at all times.

To improve service accessibility of the current service center deployment, numerous minor or substantial changes to the network of EMS centers located are required. In this paper, we study the problem of achieving required average response time by minimal changes in current locations of EMS stations. It must be understood that even minor changes incur financial costs, which is why the optimization must be carried out in such a way that it is acceptable to the public and the authorities responsible for the efficacy of EMS system performance.

[1] University of Žilina, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovakia, jaroslav.janacek@fri.uniza.sk
[2] University of Žilina, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovakia, marek.kvet@fri.uniza.sk

To formulate the model, we follow the weighted *p*-median problem, which must be extended and adjusted for real conditions of EMS system. According to previous research (Avella et al., 2007; García et al., 2011; Janáček, 2008), we make use of so-called *radial approach*, which enables us to obtain the exact solution in much smaller computational time. Since the results of this research are assumed to be applied on the EMS system in the Czech Republic or Slovakia (even if the computational study was performed on smaller benchmarks), which is not a small system concerning the number of cities, villages and service centers, the former location-allocation approach would not be able to give the optimal solution due to enormous memory and computational demands. Furthermore, there are several advanced possibilities to accelerate mentioned radial approach to perform even faster keeping acceptable solution accuracy (Janáček and Kvet, 2016; Kvet, 2015). Another model adjustment option may be aimed at incorporating uncertain conditions in the associated transportation network, which may directly affect EMS system performance and service accessibility on time (Czimmermann and Koháni, 2018).

There are two major flaws in the standard mathematical models (Matiaško and Kvet, 2017; Revelle et al., 2008). Real systems' unpredictable behavior is the main issue. The unpredictability of intervention requests means that the closest EMS station may be temporarily unavailable owing to other obligations. In such a case, the second-to-nearest or nearest-available station would handle the continuing emergency call. The suggested modeling approach improves the objective function's realism while somewhat simulating the stochastic nature of actual systems. According to (Jankovič, 2016; Kvet, 2014), this idea has been put out as the notion of *generalized disutility*. The assumption that the service will be provided from the nearest available station, which may not be the geographically closest, raises another important assumption, particularly in densely populated cities. The assumption is based on the idea of allowing multiple facilities to be located in one network node.

Finally, it's important to remember that the results of any optimization will introduce a new network structure design into the current system. Some stations will relocate within the network to areas with more demand or those that would yield a more noticeable boost to the optimized quality metric. The public, however, views change with great sensitivity and frequently do not accept them with understanding. This fact motivated us to address a fundamental scientific question defined as follows: How to achieve the desired value of the optimization criterion using the minimal number of changes compared to the current service stations deployment.

2 Problem Specification by Means of Integer Programming

2.1 Preliminaries

The following denotations are required to be presented in order to fulfill the requirement of providing the readers with the mathematical framework of the associated optimization problem. To maintain the highest level of clarity, we have summarized the symbols used in the following table.

Symbol	Index	Meaning
I		finite set of places, at which an ambulance can be located
I^c		set of places, at which ambulances are currently located
u_i	$i \in I$	upper bound of the number of ambulances, which can be placed at location $i \in I$
f_i	$i \in I$	lower bound of the number of ambulances, which can be placed at location $i \in I$
J		finite set of places of the serviced region, where demands for service are generated
w_j	$j \in J$	frequency of demands for service generated at place $j \in J$
r		maximum considered order of ambulance placement in order of increasing time distance from the demand location
K		range $\{1, .., r\}$
q_k	$k \in K$	probability that the k-th nearest ambulance is the first available for $k \in K$
V		finite set of radii used in the associated radial model of the location problem
e		time difference between a pair of neighboring zones. The zones are uniformly spaced, and we do not consider a radius-dependent approach (like e_v for $v \in V$).
T^d		demanded average response time of the system
p		total number of ambulances at disposal
a_{ijv}	$i \in I,\ j \in J,\ v \in V$	gets its value from the domain $\{0, 1\}$ and the value 1 means that possible ambulance location $i \in I$.is inside the radius $v \in V$ from the place $j \in J$.

Table 1 Used symbols and coefficients

The following Table 2 summarizes the meaning of the decision variables.

Variable	Index	Values	Meaning
y_i	$i \in I$	Z_0^+	number of ambulances, which should be located at the place $i \in I$
x_{jvk}	$j \in J$, $v \in V$, $k \in K$	$\{0, 1\}$	takes the value of 1 if the k-th closest ambulance is located outside the radius $v \in V$ from the demand location $j \in J$. In the opposite case, the variable takes the zero value.

Table 2 Summary of used decision variables

Having introduced all necessary input sets, symbols, coefficients and decision variables, we can continue by formulating the mathematical model.

2.2 Mathematical model

The mathematical model follows from the common weighted p-median problem formulation. It has been extended by the concept of generalized disutility with the possibility of locating more stations in the same location. The complete model takes the following form:

$$Maximize \quad \sum_{i \in I^c} y_i \tag{1}$$

$$Subject\ to \quad \sum_{k \in K} x_{jvk} + \sum_{i \in I} a_{ijv} y_i \geq r \quad for\ j \in J,\ v \in V \tag{2}$$

$$\frac{e \sum_{j \in J} \sum_{k \in K} \sum_{v \in V} w_j q_k x_{jvk}}{\sum_{j \in J} w_j} \leq T^d \tag{3}$$

$$\sum_{i \in I} y_i = p \tag{4}$$

$$y_i \geq f_i \quad for\ i \in I \tag{5}$$

$$y_i \leq u_i \quad for\ i \in I \tag{6}$$

$$y_i \in Z_0^+ \quad for\ i \in I \tag{7}$$

$$x_{jvk} \in \{0, 1\} \quad for\ j \in J,\ v \in V,\ k \in K \tag{8}$$

The objective function (1) maximizes the number of stations located in places occupied currently. This objective also minimizes the number of changes. The series of constraints (2) ensures the relation among the decisions on locating ambulances and the auxiliary decision variables x_{jvk}. In this formulation, the sum of variables x_{jvk} is greater than or equal to the complement of the number of service centers located in the radius v from the customer j to the number r. In other words, the optimization process will cause the sum to take the value of the difference of r and the number of service centers located in the radius v from the customer j if the difference is positive; otherwise, it takes the value of zero. The left-hand side of inequality (3) expresses the average response time of the system, and the constraint assures that the average response time is less than or equal to the demanded value. The mathematical expression describing the average time to reach the request location from the closest available facility is based on the radial formulation, which is described in detail in (Janáček, 2008; Kvet, 2015). The expression (4) limits the number of EMS ambulances located by p. The series of constraints (5) ensures that the minimal number of ambulances will be located at the locations where it is necessary. The series of constraints (6) limits the number of ambulances located at location i. The constraints (7) and (8) are obligatory constraints determining the feasible values of decision variables.

3 Case Study

3.1 Datasets and technical equipment

This study conducted numerical experiments utilizing road network data. The tailored radial model of the Emergency Medical Service system, designed to meet intervention requirements, was employed to enhance the systems functioning in the self-governing region of Žilina (ZA) in Slovakia and the Moravskoslezský kraj (MSK) in the Czech Republic.

The mathematical model (1)-(8) formulated in previous section was solved by the optimization software FICO Xpress 8.3 (64-bit, release 2017) on a common desktop computer equipped with the Intel® Core™ i7 8565U 1.80 GHz processor and 128 GB RAM.

3.2 Analysis of achieved results

Before running the model (1)-(8), it was necessary to analyze the interval, in which the demanded average response time T^d may move. The highest possible value was taken from the current stations' deployment, and it is further denoted by T^c. No optimization should worsen the measurable indicators of the current state. On the contrary, it is done to improve them. The lowest possible average response time T^{opt} achievable with current number p of located EMS stations was computed exactly by solving the following model (9)-(15). Special description and explanation of particular constraints is not necessary, most of them are identical to the previous model (1)-(8).

$$Minimize \quad \frac{e\sum_{j\in J}\sum_{k\in K}\sum_{v\in V} w_j q_k x_{jvk}}{\sum_{j\in J} w_j} \tag{9}$$

$$Subject\ to \quad \sum_{k\in K} x_{jvk} + \sum_{i\in I} a_{ijv} y_i \geq r \quad for\ j\in J,\ v\in V \tag{10}$$

$$\sum_{i\in I} y_i = p \tag{11}$$

$$y_i \geq f_i \quad for\ i\in I \tag{12}$$

$$y_i \leq u_i \quad for\ i\in I \tag{13}$$

$$y_i \in Z_0^+ \quad for\ i\in I \tag{14}$$

$$x_{jvk} \in \{0, 1\} \quad for\ j\in J,\ v\in V,\ k\in K \tag{15}$$

The original number p of ambulances at disposal was 64 for the region MSK and 37 for the region ZA. The values of current and optimal average response times in minutes for MSK are $T^c = 7.076$ and $T^{opt} = 6.430$. The values of current and optimal average response times for ZA are $T^c = 6.819$ and $T^{opt} = 6.00$. The computational time consumed to obtain the optimal solution T^{opt} was 57.4 seconds for MSK and 30.4 seconds for ZA.

If we know the values of T^c and T^{opt}, the demanded average response time T^d may get its value from the interval $[T^{opt}, T^c]$ based on the following expression: $T^d = \alpha T^c + (1-\alpha)T^{opt}$ for $\alpha = 0.0, 0.1, \ldots, 1.0$.

The results for the benchmarks ZA and MSK are reported in Table 3. Symbol *NoOrig* denotes the number of ambulances which stay at their original locations, *CT* stands for computational time in seconds and T^* is the resulting average response time in minutes.

α	Region MSK			Region ZA		
	NoOrig	*CT*	T^*	*NoOrig*	*CT*	T^*
0.0	43	1585.6	6.430	24	1018.7	6.000
0.1	54	2591.3	6.494	27	942.0	6.063
0.2	57	715.4	6.553	29	1067.1	6.135
0.3	60	594.0	6.620	31	1249.3	6.224
0.4	61	472.9	6.649	33	820.8	6.315
0.5	63	646.7	6.748	34	677.7	6.371
0.6	63	576.6	6.748	35	614.5	6.440
0.7	63	479.2	6.748	36	648.5	6.520
0.8	63	520.7	6.748	36	666.9	6.520
0.9	63	482.9	6.748	37	602.0	6.706
1.0	64	310.9	7.049	37	487.1	6.706

Table 3 Results of performed computational experiments

The results achieved have confirmed our expectations. As far as the computational complexity of the problem is concerned, the radial model is solvable quite well within several minutes that is acceptable. The lower average response time we request, the more changes that are necessary to be made and thus, the lower number of stations are kept in the same location as currently. Presented results may serve also for the negotiation between urgent healthcare accessibility and the number of emergency service station transfers within the serviced network. The presented research and the results achieved thus represent a contribution in the field of strategic decision-making.

4 Conclusions

This paper deals with the optimal reconstruction of the existing Emergency Medical Service system while maintaining the size of the existing ambulance fleet. Given that the average response time of the existing system is known, and the optimal average system time can be calculated, a reasonable reconstruction of the system can lead to a new average response time that is between these two values. The problem is formulated so that a given average response time value is achieved by making minimal changes to the current ambulance fleet layout, to avoid additional costs associated with ambulance relocation.

Experiments conducted in two regions with different desired average response times have shown that the optimal average response time can be achieved by changing a maximum of 35 percent of the original ambulance locations. The results also suggest solutions that require minimal changes in ambulance layout (up to 7 percent) and achieve significant improvements in average response time.

The goal of future studies in this area might be to find better modeling tools to maximize the placement of ambulances while maintaining a reasonable average response time.

Acknowledgements

This work was supported by the following Slovak national research grants: VEGA 1/0220/25 “Regional equality in access to public rescue services provided from service centers on the transport network”, VEGA 1/0696/26 “Predictions of delays and vehicle occupancies to inform the planning of public transport”, and VEGA 1/0516/26 “Optimal design of charging infrastructure for electric vehicles and its efficient operation in urban mobility conditions”.

References

Avella P, Sassano A, Vasil'ev I (2007) Computational study of large scale p-median problems. Mathematical Programming 109: 89-114. https://doi.org/10.1007/s10107-005-0700-6

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. European Journal of Operations research 147: 451-463. https://doi.org/10.1016/S0377-2217(02)00364-8

Chanta, S, Mayorga, M E, McLay, L A (2014). Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. Annals of Operations Research 221:133-159

Czimmermann P, Koháni M (2018) Characteristics of changes of transportation performance for pairs of critical edges. Communications 20: 84-87

Doerner K F, Gutjahr W J, Hartl R F, Karall M, Reimann M (2005) Heuristic Solution of an Extended Double-Coverage Ambulance Location Problem for Austria. Central European Journal of Operations Research 13: 325-340

García S, Labbé M, Marín A (2011) Solving large p-median problems with a radius formulation. INFORMS Journal on Computing 23: 546-556. https://doi.org/10.1287/ijoc.1100.0418

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. Health care management science 11: 262-274. https://doi.org/10.1007/s10729-007-9048-1

Janáček J (2008) Approximate Covering Models of Location Problems. In: Lecture Notes in Management Science: Proceedings of the 1st International Conference ICAOR, Yerevan, Armenia, pp 53-61

Janáček J, Kvet M (2016) Sequential approximate approach to the p-median problem. Computers & industrial engineering 94: 83-92. https://doi.org/10.1016/j.cie.2016.02.004

Jankovič P (2016) Calculating Reduction Coefficients for Optimization of Emergency Service System Using Microscopic Simulation Model. In: CINTI 2016. Conference Proceedings, pp 163-167. https://doi.org/10.1109/CINTI.2016.7846397

Jánošíková Ľ (2007) Emergency Medical Service Planning. Communications 9: 64-68

Jánošíková Ľ, Jankovič P, Kvet M, Zajacová F (2021) Coverage versus response time objectives in ambulance location. International Journal of Health Geographics 20: 1-16. https://doi.org/10.1186/s12942-021-00285-x

Karatas M, Yakıcıa E (2019) An analysis of p-median location problem: Effects of backup service level and demand assignment policy. European Journal of Operational Research 272: 207-218

Kvet M (2014) Computational Study of Radial Approach to Public Service System Design with Generalized Utility. In: Digital Technologies 2014. Conference Proceedings, Žilina, Slovakia, pp 198-208

Kvet M (2015) Advanced Radial Approach to Resource Location Problems. In: Studies in Computational Intelligence: Developments and Advances in Intelligent Systems and Applications, pp 29-48. https://doi.org/10.1007/978-3-319-58965-7_3

Matiaško K, Kvet M (2017) Medical data management. Informatics 2017. Conference Proceedings, pp 253-257

Reuter-Oppermann M, van den Berg P L, Vile J L (2017) Logistics for Emergency Medical Service systems, Health Systems 6: 187-208. https://doi.org/10.1057/s41306-017-0023-x

Revelle C S, Eiselt H A, Dakin M S (2008) A bibliography for some fundamental problem categories in discrete location science, European Journal of Operational Research 184: 817–848. https://doi.org/10.1016/j.ejor.2006.12.044

Robust Regression Methods for Composite Indicator Analysis: An Application to TTDI

Jan Kalina[1], Kristýna Dvořáková [2]

Abstract. This paper investigates the performance of a robust regression framework based on recently proposed methods, including least weighted squares (LWS), its regularized extension LWS-lasso, and a robust coefficient of determination. Although these methods have recently been introduced, their statistical properties and practical performance under data contamination have not yet been systematically investigated. We study the behavior and stability of these methods under contamination by outliers and local noise, comparing their performance with competing estimators. The analysis highlights the advantages of robust approaches in regression settings where classical methods may lead to unstable conclusions. As an empirical illustration in the context of composite indicators, we apply the framework to the Human Resources and Labor Market sub-pillar of the Travel and Tourism Development Index for 119 countries.

Keywords: robust linear regression, least weighted squares, robust coefficient of determination, outliers, multicollinearity, human resources

JEL Classification: C13, C21, J24, J11

1 Introduction

Regression analysis is one of the fundamental tools for statistical modeling and empirical research (Čížek, 2016). However, classical estimators can be highly sensitive to outliers, influential observations, and deviations from ideal data assumptions. Robust regression methods address this issue by reducing the influence of atypical observations and providing more stable estimates under contaminated data conditions. An important research question is therefore to understand the behavior of robust estimators and their advantages over classical approaches in practical applications.

Recently, several robust regression methods based on least weighted squares (LWS) have been proposed. These include the LWS-lasso estimator (Kalina, 2025), which combines robustness with L_1-regularization, and the robust coefficient of determination R^2_{LWS} (Kalina, 2025). Although these methods are promising, a systematic assessment of their practical performance, robustness properties, and stability under data contamination remains an open research question. This paper contributes to filling this gap by evaluating their behavior under clean and contaminated conditions and comparing their performance with classical regression approaches.

Composite indicators provide a relevant application setting for robust regression methods, as they often involve multiple dimensions, multicollinearity, and potentially influential observations (Kalina, 2014). As an empirical illustration, we apply the proposed methods to the Human Resources and Labor Market (HRLM) sub-pillar of the Travel and Tourism Development Index (TTDI), compiled by the World Economic Forum. The TTDI was introduced in 2022 as a successor to the Travel and Tourism Competitiveness Index (TTCI), and its dimension weights within several sub-pillars, including HRLM, have not been published or analyzed.

Previous studies have used TTDI data to investigate tourism competitiveness, country performance, and the role of human capital, infrastructure, and sustainability in tourism development (Allahverdi et al., 2025). These applications demonstrate the relevance of TTDI as a composite indicator and motivate its use as a real-world illustration of robust regression methods (Zhang, 2025). The cross-country nature of TTDI data provides a challenging empirical setting, where heterogeneous observations may affect the stability of regression-based estimates (Marti and Puertas, 2025).

This paper is organized as follows. Section 2 presents the statistical methodology, including the LWS estimator, its regularized extension LWS-lasso, and the robust coefficient of determination. Section 3 applies these methods to the HRLM sub-pillar of the TTDI and analyzes the sensitivity of the estimates under potential contamination

[1] The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, Prague 8, Czech Republic & The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, Prague 8, Czech Republic, kalina@cs.cas.cz

[2] The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, Prague 8, Czech Republic, dvorakova@cs.cas.cz

by outliers and local noise. The results demonstrate that LWS-lasso provides the most stable estimates among the considered methods, particularly under contaminated data conditions. Finally, Section 4 summarizes the main findings and their implications.

2 Robust statistical methodology

This section introduces the methodological background of the paper. We first recall robust estimation in linear regression, with a particular focus on the least weighted squares (LWS) estimator and its relation to the least trimmed squares approach. We then briefly present a regularized extension of LWS suitable for high-dimensional settings, referred to as LWS-lasso. Finally, we introduce a robust coefficient of determination based on implicit weighting, which builds upon the LWS framework and serves as a robust alternative to the classical R^2 measure.

We consider the standard linear regression model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n, \tag{1}$$

the least squares estimator represents the best linear unbiased estimator of $\beta = (\beta_0, \ldots, \beta_p)^T \in \mathbb{R}^p$. This is true even without normality of the errors $e_1, \ldots, e_n$. The coefficient of determination R^2 is the most common measure of quality of the linear fit given by the least squares (Greeene, 2018). We will use the notation $X_i = (X_{i1}, \ldots, X_{ip})^T \in \mathbb{R}^p$ for $i = 1, \ldots, n$. We denote the design matrix $X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times p}$ and the response $Y = (Y_1, \ldots, Y_n)^T$. Because the least squares estimator is highly vulnerable to the presence of outliers in the data, various robust alternatives have been proposed.

2.1 The least weighted squares estimator

Let us recall the least weighted squares (LWS) estimator for the linear model (1). Under mild assumptions that the errors are independent and identically distributed, the estimator has appealing properties including high robustness in terms of the breakdown point and also high efficiency; see (Víšek, 2011) and references cited therein. The breakdown point, which represents a measure of robustness of a statistical estimator of an unknown parameter under the presence of outliers in the data, is formally defined as the minimal fraction of data that can drive an estimator beyond all bounds when set to arbitrary values (Jurečková et al., 2019). Under regularity conditions (Čížek, 2016), the LWS estimator is consistent and asymptotically normal, while retaining a high breakdown point and efficiency (Víšek, 2011).

Let us denote the residuals corresponding to a particular $\hat{\beta} \in \mathbb{R}^p$ as $u_i(\hat{\beta}) = Y_i - X_i^T \hat{\beta}$ and order statistics of their squares as

$$u^2_{(1)}(\hat{\beta}) \leq \cdots \leq u^2_{(n)}(\hat{\beta}). \tag{2}$$

Let us denote the ranks of the squares residuals by $R_1(\hat{\beta}), \ldots, R_n(\hat{\beta})$; thus, $R_i(\hat{\beta})$ denotes the rank of $u_i^2(\hat{\beta})$ among $u_1^2(\hat{\beta}), \ldots, u_n^2(\hat{\beta})$. The LWS estimator $\hat{\beta}^{LWS} = (\hat{\beta}_0^{LWS}, \ldots, \hat{\beta}_p^{LWS})^T \in \mathbb{R}^p$ is formally defined by

$$\hat{\beta}^{LWS} = \underset{\tilde{\beta} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \psi\left(\frac{R_i(\tilde{\beta}) - 1/2}{n}\right) u_i^2(\tilde{\beta}), \tag{3}$$

where ψ denotes a given non-increasing weight function $\psi : [0, 1] \to [0, 1]$. Let us write

$$w_i = \psi\left(\frac{i - 1/2}{n}\right), \quad i = 1, \ldots, n. \tag{4}$$

Apparently, the LWS estimator has the form of the (classical) weighted least squares estimator. Importantly, $\hat{\beta}^{LWS}$ is computed with weights that are assigned to individual observations after the *optimal permutation* given by the minimization in (3). This is different from the classical weighted regression (weighted least squares) with given (known) weights for individual observations. Once more, the LWS estimator requires to choose the weight function ψ and the individual weights $w_1, \ldots, w_n$ are assigned in an implicit way after a permutation that is determined intrinsically during the estimation. For the computation of the LWS estimator, a straightforward extension of the FAST-LTS algorithm of (Rousseeuw and Van Driessen, 2006) may be used. The estimator is the basis for proposing its extensions, which will be now presented.

2.2 LWS-lasso

The LWS-lasso estimator is a regularized extension of LWS that combines its robustness with L_1-regularization, enabling sparse estimation while retaining resistance to outliers. Although the original proposal in (Kalina, 2025) was intended for high-dimensional settings with $n < p$, regularization can also be beneficial when $n > p$, helping to stabilize estimates and enhance robustness (Kalina and Rensová, 2015). Empirical results based on simulations and real data analyses indicate that LWS-lasso can outperform regularized regression methods, particularly under contamination by outliers.

The LWS-lasso estimator was implemented by the authors in the R software. Its computation is based on an extension of the FAST-LTS algorithm adapted to the LWS objective function. The algorithm starts from a number of randomly generated initial subsets and iteratively refines each candidate solution by repeatedly updating the subset of observations defining the weights and recomputing the corresponding weighted L_1-regularized regression estimate. The final estimate is selected as the candidate minimizing the LWS objective function over all random starts. The regularization parameter λ is determined by 10-fold cross-validation over a predefined logarithmically spaced grid of candidate values, with the value yielding the smallest cross-validation criterion being selected.

2.3 The LWS-based correlation coefficient

The coefficient of determination R^2 is widely used to assess goodness of fit in linear regression but is sensitive to outliers and heteroskedasticity, which can lead to misleading conclusions. This motivated the development of robust alternatives, such as the LWS-based coefficient of determination R^2_{LWS} (Kalina, 2025), which weights observations according to the LWS procedure to reduce the influence of outliers or high-leverage points. R^2_{LWS} combines robustness with favorable efficiency under the classical model, providing reliable performance for both contaminated and clean data. We also recall its formal definition from (Kalina, 2025).

Definition 1. Let us consider the model (1) and the LWS estimator with a given weight function ψ. Let $\tilde{w}$ denote the optimal permutation of the weights (4) determined by the LWS estimator $\hat{\beta}_{LWS}$. Let us denote

$$u_{LWS} = (u_1^{LWS}, \ldots, u_n^{LWS})^T = Y - X\hat{\beta}_{LWS}. \tag{5}$$

The LWS-based coefficient of determination is defined as

$$R^2_{LWS} = 1 - \frac{\sum_{i=1}^n \tilde{w}_i (u_i^{LWS})^2}{\sum_{i=1}^n \tilde{w}_i (Y_i - \bar{Y}_{LWS})^2}, \tag{6}$$

where $\bar{Y}_{LWS} = \sum_{i=1}^n \tilde{w}_i Y_i$ is the LWS-mean of the response values.

2.4 Regression L_1-estimator

The L_1-estimator, also known as least absolute deviations (LAD) regression, is a classical regression method that is inherently robust to outliers in the response variable (Jurečková et al., 2019). A robust R^2 measure related to the L_1-estimator will be denoted here as L_1-R^2. Let $\hat{Y}_1, \ldots, \hat{Y}_n$ denote the L_1-predictions of $Y_1, \ldots, Y_n$, and let $\tilde{Y}$ denote the median of $Y_1, \ldots, Y_n$. Then, the L_1-R^2 measure is defined as

$$L_1\text{-}R^2 = 1 - \frac{\sum_{i=1}^n |Y_i - \hat{Y}_i|}{\sum_{i=1}^n |Y_i - \tilde{Y}|}. \tag{7}$$

3 Empirical analysis of the HRLM sub-pillar

In this section, we present a detailed empirical analysis of the Human Resources and Labor Market (HRLM) sub-pillar of TTDI. Our focus is on illustrating the performance and robustness of recently proposed statistical estimators, as introduced in Section 2, when applied to real-world socio-economic data. We derive the weights of the three HRLM dimensions and examine their stability under artificially introduced contamination, including outliers and local noise. This study demonstrates the practical advantages of robust regression techniques in preserving reliable estimates under challenging data conditions.

3.1 Data description and modeling framework

The data for the HRLM subindex and its dimensions come from the website (World Economic Forum, 2024) and correspond to $n = 119$ countries worldwide evaluated for the year 2024. This dataset provides a comprehensive

Model	Predictor(s)	R^2	L_1-R^2	R^2_{LWS}
Single-predictor	X_1	0.82	0.80	0.85
Single-predictor	X_2	0.31	0.33	0.31
Single-predictor	X_3	0.79	0.84	0.86
Two-predictor	X_1, X_2	0.89	0.86	0.92
Two-predictor	X_2, X_3	0.90	0.87	0.92
Two-predictor	X_1, X_3	0.94	0.92	0.95

Table 1 A comparison of R^2 measures for models predicting HRLM from its dimensions (Section 3.3).

overview of HRLM indicators globally. According to the TTDI documentation, which is available at the same website (World Economic Forum, 2024) as the TTDI data, the HRLM sub-pillar is constructed as a linear combination of its three dimensions. This motivates our approach to estimate these weights from the available data.

The HRLM subindex is structured into three dimensions (sub-subpillars), which play the role of the predictors in the regression model:

- Qualification of the Labor Force,
- Labor Market Dynamics, and
- Labor Market Resilience and Equality.

These dimensions capture key aspects of workforce quality, skills, and labor market efficiency that are essential for sustainable travel and tourism development.

All the values of the HRLM dimension are positive, and a visual inspection of the data did not reveal any obvious outliers. We use R software for the computations and perform the regresion modeling with the following three estimators:

- Ordinary least squares,
- Median (L_1) regression,
- LWS-lasso with linearly decreasing weights.

3.2 Baseline model for raw data

We begin the analysis with the linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i, \quad i = 1, \ldots, n, \tag{8}$$

where HRLM is the response and the three dimensions of HRLM are considered together with an intercept in the model. Thus, we have $p = 4$. Using the ordinary least squares (OLS), we obtain $R^2 = 1.00$ and all estimated weights for the three HRLM dimensions are equal to $1/3$. This result is identical when LWS-lasso and L_1 estimators are used, highlighting the stability and consistency of these methods on uncontaminated data. Moreover, all three coefficients are highly significant when using OLS. Also, the values of L_1-R^2 and R^2_{LWS} are equal to 1.00.

The data allow us to verify the construction of the HRLM sub-pillar, which is explicitly defined as an equally weighted average of its three constituent dimensions according to the appendix of the TTDI documentation (World Economic Forum, 2024). In this sense, the linear model serves as a consistency check, confirming that the observed dimension scores aggregate to the subindex through equal weights. Such a setting provides a useful validation of the internal structure of the data. The resulting equal weights facilitate interpretation, as the composite HRLM index can be understood as an equally weighted combination of its constituent dimensions, ensuring transparency in the aggregation scheme.

3.3 A more detailed analysis

The correlation matrix of the three HRLM dimensions reveals notable differences in their pairwise associations. Dimensions X_1 and X_3 are relatively strongly correlated with Pearson correlation coefficient $r = 0.72$, indicating substantial overlap in the information they capture. In contrast, the correlations between X_1 and X_2 ($r = 0.34$) and between X_2 and X_3 ($r = 0.27$) are moderate to weak, suggesting that X_2 is less related to the other two dimensions and may contribute less as a predictor in linear models. Overall, while some dimensions are closely related, others provide complementary information for constructing the HRLM sub-pillar.

Shift	OLS	L_1	LWS-lasso
10	0.87	0.01	0.01
20	1.73	0.01	0.01
30	2.60	0.01	0.01
40	3.47	0.01	0.01
50	4.34	0.01	0.01

Table 2 Total absolute deviation of HRLM weights under outlier contamination of Section 3.4.

The condition number κ of the full design matrix X is 32.2, exceeding the commonly used threshold of 30, which indicates severe multicollinearity. It is worth noting that the presence of multicollinearity cannot be reliably detected from individual t-tests of the regression coefficients. Despite some coefficients appearing statistically significant, the high correlations among predictors (X_1, X_2, X_3) and the condition number of 32.2 indicate severe multicollinearity, which inflates standard errors and complicates interpretation of individual effects.

Table 1 presents R^2 and its robust versions for all possible single- and two-predictor linear models of the HRLM dimensions. The results indicate that X_1 and X_3 are strong predictors, while X_2 is relatively weak. The high R^2 values for the two-predictor models, particularly X_1 and X_3, suggest substantial multicollinearity among the dimensions. Values of robust R^2 measures are very similar to those of R^2.

The robust coefficient of determination R^2_{LWS} closely matches the classical R^2 across all submodels, demonstrating that it preserves the explanatory power of the predictors on uncontaminated data. This behavior confirms that the HRLM data contain no significant outliers.

3.4 Analysis under artificial contamination by outliers

We can expect that the advantage of robust R^2 compared to the classical R^2 will be revealed especially on contaminated data. To investigate this, we perform the same analysis on the data with artificially introduced outliers.

To assess the robustness of the HRLM dimension weights, we conducted a repeated contamination study. In 1,000 simulation runs, five randomly selected countries had their HRLM subindex values perturbed by varying shifts, generating extreme outliers. This setup simulates atypical or extreme country profiles and allows us to systematically evaluate the impact of extreme observations on different regression estimators. Particularly, we compare the performance of three regression estimators. For each run, we computed the regression coefficients corresponding to the three HRLM dimensions. The results across all runs were averaged, providing a quantitative assessment of estimator stability under repeated contamination.

Table 2 summarizes the effect of artificial outliers on the estimated HRLM dimension weights for different regression estimators. Each row corresponds to a specific magnitude of the outlier shift, while the values in the coumns columns represent the average total (absolute) deviation from the reference weights computed on the uncontaminated data across multiple simulation runs. The total deviation is computed as

$$\text{Total deviation} = \sum_{j=1}^{4} \left| \hat{\beta}_j^{\text{contam}} - \hat{\beta}_j^{\text{raw}} \right|, \tag{9}$$

where $\hat{\beta}_j^{\text{raw}}$ and $\hat{\beta}_j^{\text{contam}}$ denote the estimated weight of the j-th HRLM dimension for the original and outlier-contaminated data, respectively.

The table highlights the extreme sensitivity of classical OLS to even a few outliers, whereas robust methods maintain stability, showing only minor deviations from the original weights. It should be noted that small total deviation values for robust estimators, such as 0.01, do not indicate an error; rather, they result from the nearly perfect fit of the regression model to the uncontaminated data. Moreover, while the classical R^2 can fluctuate widely across individual simulation runs, even approaching zero, the robust versions remain consistently close to 1, highlighting the stability of robust estimators. On the whole, these results demonstrate the advantage of robust estimators, with LWS-lasso providing the most stable HRLM weight estimates under repeated outlier contamination.

3.5 Analysis under artificial contamination by local noise

To further evaluate the robustness of the HRLM dimension weights, we conducted a local contamination study across multiple noise levels. In this experiment, small Gaussian noise with varying values of the standard deviation

σ^*	**OLS**	L_1	**LWS-lasso**
0.01	0.004	0.004	0.003
0.05	0.020	0.028	0.014
0.10	0.042	0.051	0.030

Table 3 Total deviation of HRLM weights under local contamination of Section 3.5.

($\sigma^* = 0.01, 0.05, 0.10$) was added independently to each of the three HRLM dimension scores for all countries here, simulating minor perturbations in the predictors. We do not modify HRLM as the response variable.

For each noise level, regression coefficients were recomputed using the same estimators as before. For each combination of noise level and estimator, we calculated the total absolute deviation of the estimated weights from the reference values obtained on the original, uncontaminated data, defined as (9), where $\hat{\beta}_j^{\text{contam}}$ are obtained in the model contaminated by local noise.

The results in Table 3 indicate that LWS-lasso remains the most stable estimator across all noise magnitudes, OLS exhibits moderate sensitivity, and L_1 regression is the most affected by local perturbations. A similar pattern is observed for R^2 measures, with the classical R^2 and its robust counterparts showing corresponding stability and sensitivity across the different noise levels. Overall, this highlights that LWS-lasso provides superior protection for HRLM weight estimation against both minor perturbations of the input data.

4 Conclusion

This study was motivated by the recent introduction of robust regression methods whose behavior and advantages over classical approaches have not yet been systematically investigated. We evaluated the performance of these methods, with particular emphasis on LWS-lasso, under clean and contaminated data conditions. The results demonstrate that the robust estimators behave consistently in the absence of contamination: their estimates coincide with OLS, confirming that robustness does not compromise the behavior of classical regression in ideal settings.

When the original data are artificially contaminated by outliers or local noise, LWS-lasso provides more stable estimates than classical estimators, demonstrating the advantage of combining robustness with regularization. This property is particularly relevant in regression settings affected by influential observations, multicollinearity, or near-degenerate structures (Kalina et al., 2019). These findings highlight the potential of robust regression methods for the analysis of composite indicators, especially when the underlying aggregation structure is deterministic or nearly degenerate.

As an empirical illustration, we applied the proposed framework to the Human Resources and Labor Market sub-pillar of the Travel and Tourism Development Index (TTDI) for 119 countries. The analysis confirmed that the three dimensions of this sub-pillar receive equal weights, in agreement with the TTDI construction methodology. These equal weights should, however, be interpreted as an aggregation choice rather than as evidence that the underlying dimensions have identical economic importance.

Acknowledgments

This work was supported by the Czech Science Foundation grant 24-10078S ("New nonparametric tools for econometric data analysis") of the Czech Science Foundation.

References

Allahverdi M, Akandere G, Varol F (2025) Evaluation of the competitiveness and performance of destinations through clustering method within the scope of the Travel and Tourism Development Index. Journal of Travel Research 27: e70122

Čížek P (2016) Generalized method of trimmed moments. Journal of Statistical Planning and Inference 171: 63–78.

Greeene W H (2018) Econometric analysis. 8th ed. Pearson Education Limited, Harlow

Jurečková J, Picek J, Schindler M (2019) Robust statistical methods with R. 2nd ed. CRC, Boca Raton

Kalina J (2014) On robust information extraction from high-dimensional data. Serbian Journal of Management 9: 131–144

Kalina J (2025) Robust dimensionality reduction: How to increase efficiency of machine learning computations. In Ben Youssef, B., Ben Ismail, M.M., eds.: Integrating machine learning into HPC-based simulations and analytics. IGI Global, Hershey, pp 37–60

Kalina J, Rensová D (2015) How to reduce dimensionality of data: Robustness point of view. Serbian Journal of Management 10: 131–140

Kalina J, Vašaničová P, Litavcová E (2019) Regression quantiles under heteroscedasticity and multicollinearity: Analysis of travel and tourism competitiveness. Ekonomický časopis/Journal of Economics 67: 69–85

Marti L, Puertas R (2025) Relationship between tourism development and the socioeconomic progress of tourist destinations. Competitiveness Review 35: 175–189

Rousseeuw P J, Van Driessen K (2006) Computing LTS regression for large data sets. Data Mining and Knowledge Discovery 12: 29–45

World Economic Forum (2024). Travel & Tourism Development Index 2024. Available at: https://www.weforum.org/publications/travel-tourism-development-index-2024/

Víšek J Á (2011) Consistency of the least weighted squares under heteroscedasticity. Kybernetika 47: 179–206

Zhang J (2025) Patterns of innovation-driven tourism competitiveness: Insights from 270 Chinese cities. Tourism Management 107: 105063

Predicting Prices and Discovering Structure in Magic: The Gathering Card Data

Nikola Kaspříková[1]

Abstract. The secondary market for Magic: The Gathering cards represents a mature example of a collectibles market where prices are shaped by a combination of intrinsic card properties, gameplay relevance, and external demand signals. This paper presents an analysis of Magic: The Gathering card data with two complementary objectives: (i) predicting card prices on the secondary market using machine learning models, and (ii) identifying latent structural groupings of cards through clustering techniques.
Using publicly available card metadata from the Scryfall bulk dataset augmented with secondary market price information, we construct a feature–rich dataset capturing card rarity, mana cost, color identity, card type, and textual characteristics derived from rules text. For price prediction, we build linear regularized regression model. To explore the underlying structure of the card space, we apply a price-driven segmentation based on a regression tree model.

Keywords: collectible card games, secondary market pricing, regression tree, market segmentation

JEL Classification: C38, Z11

1 Introduction

The increasing popularity of the trading card games has led to increase in the trading cards market, which is attracting players, collectors as well as investors seeking alternative investment assets. Magic: The Gathering is a collectible trading card game published by Wizards of the Coast since 1993. Players build decks from individual cards with different game functions, rarities, colors, and formats of legality. The secondary market for Magic: The Gathering cards represents a mature example of a collectibles market where prices are shaped by a combination of intrinsic card properties, gameplay relevance, and external demand signals. For a general economic background in card games collecting with focus on Magic: The Gathering, see (Abrahamse et al, 2026). The price prediction models in (Sakaji et al, 2020) are based on predictors including the features obtained from card text, whereas other papers refer to artificial neural network models which may capture card image representation, see (Bertram et al, 2024 or Zilio et al, 2018). This paper presents an analysis of Magic: The Gathering card data with two complementary objectives: (i) predicting card prices on the secondary market using machine learning models, and (ii) segmentation of cards, i.e. identifying latent structural groupings of cards.

The structure of this paper is as follows: the sample selected for the analysis is described first, followed by models description and the discussion of results, followed by some concluding remarks.

This study restricts the sample to cards from expansion sets released in 2024. This choice is intended to reduce temporal and product–level heterogeneity in the dataset. Cards released within a single year were designed and introduced under relatively similar market conditions, product structures, and gameplay environments.

In addition, focusing on expansion sets excludes special products such as promotional cards, Commander decks, Secret Lair releases, and other collector–oriented printings. These products often differ in distribution model, target audience, scarcity, and pricing dynamics, which could introduce additional noise into the analysis. Expansion sets therefore provide a more homogeneous sample for examining the relationship between card characteristics and secondary–market prices.

Finally, since the analysis is conducted in 2026, cards released in 2024 are no longer newly released products. Their prices have had time to partially stabilize after initial release-driven speculation, while still representing recent card design and contemporary player demand. The restriction to 2024 expansion sets therefore offers a balance between recency, comparability, and market maturity.

This paper applies a regression tree as a price–driven segmentation method. While unsupervised clustering groups cards according to overall feature similarity, a regression tree partitions the dataset according to differences in the

[1] Prague University of Economics and Business, Department of Mathematics, nám. W. Churchilla 1938/4, Praha, Czech Republic, nikola.kasprikova@vse.cz

target variable, namely the secondary–market price. The model recursively splits the data using card–level features such as rarity, mana value, card type, format legality, printing characteristics, and text–derived mechanics. The terminal nodes of the tree are interpreted as market segments containing cards with similar predicted price levels. This approach is useful because it provides an interpretable structure: it shows not only which cards are grouped together, but also which features and thresholds explain differences in price. The regression tree is therefore used as an exploratory tool for identifying meaningful price segments among cards from 2024 expansion sets.

The features to be used in the analysis were derived from card metadata, including rarity, mana value, color identity, card type, legality across formats, and selected textual indicators extracted from oracle text on the card which provides a standardized description of the card's gameplay function. The dataset for the analysis has been obtained from Scryfall Default cards dataset retrieved on 30th April 2026 (Scryfall 2026) and the data analysis is performed in R software (R Foundation for Statistical Computing, 2026).

2 Price prediction and regression–tree–based market segmentation

The selected sample represents price and other card features of 1657 cards. The mean value of price in USD is 2.12, median 0.26, minimum 0.02, maximum 78.87, and 1st and 3rd quartiles are 0.12 and 0.83. The distribution of raw price in USD is skewed, so the dependent variable for the regression models is the log–transformed Scryfall card price in USD, defined as $log(1 + price)$. The original raw price values in USD will then be used for profiling the resulting segments for easier interpretation.

The explanatory variables were divided into gameplay, text, color, and printing-related characteristics. Gameplay variables include card type indicators and selected mechanics extracted from oracle text. Text complexity is represented by text length, word count, and keyword count. Mana and color structure are captured by the total number of mana symbols, color–specific mana symbols, color count, and color identity count. Printing–level characteristics include reprint status, foil and nonfoil availability, promo status, full-art treatment, digital status, border color, frame, and rarity. This combination allows the regression tree to segment cards according to both their gameplay function and their market-relevant printing characteristics. The following list shows the complete set of variables which have been offered as potential predictors for learning the models:

- Text–Derived Gameplay Mechanics: indicator whether the card text contains a mana-related effect; indicator whether the card creates or refers to tokens.
- Card Type Indicator: card is a creature; card is an instant; card is a sorcery; card is an artifact; card is an enchantment; card is a planeswalker; card is a land; card is a battle; card has the legendary supertype; card is an equipment; card is an aura; card is a saga.
- Printing and Product Characteristics: card is a reprint; a foil version exists; a nonfoil version exists; the card is a promotional printing; the card has full-art treatment; the card is digital-only or associated with a digital release.
- Mana Cost and Color Structure: total number of mana symbols in the mana cost; number of white mana symbols; number of blue mana symbols; number of black mana symbols; number of red mana symbols; number of green mana symbols; number of colors of the card; number of colors in the card's color identity;
- Text and Keyword Complexity length of the oracle text in characters; number of words in the oracle text; number of keyword abilities assigned to the card.
- Rarity and Visual Characteristics: rarity category of the card (common, uncommon, rare, or mythic rare); border color of the card; frame style or frame version of the card.

2.1 LASSO regression model

We build a supervised LASSO regression model for price prediction, as implemented in the glmnet package (see (Friedman et al, 2010)) first. A LASSO regression model was used to select a parsimonious subset of predictors for log–transformed card prices. The model was estimated using 10–fold cross–validation. The final model was selected using the one–standard–error rule, which favors a simpler model whose cross–validated error remains close to the minimum–error solution.

The resulting model for $\lambda = 0.0486$ has RMSE=0.58 and MAE=0.369 in cross–validation. This is not bad considering the (untransformed) mean price 2.12, but clearly the model still leaves out some price drivers. The coefficients in the final model are shown in Table 1. The resulting predictors set is small enough to allow easy interpretation, even if there have been originally offered more than 40 predictors to choose from. According to our model, higher price can be expected for land cards, rares or mythics, borderless and full–art, which seems reasonable.

	Coefficient
Intercept	0.3975
is land	0.1852
full–art	0.1781
mana symbols total	-0.0215
color count	-0.0855
rarity=common	-0.1439
rarity=mythic	1.2256
rarity=rare	0.3972
border color=borderless	0.1947

Table 1 Coefficients in LASSO regression model for log–transformed card price

2.2 Regression tree for segmentation

This study uses a regression tree (as implemented in rpart (Therneau and Atkinson, 2026) R package) to perform price–driven segmentation of Magic: The Gathering cards. Unlike classical unsupervised clustering methods, which group cards based on overall similarity across selected features, a regression tree segments the dataset specifically with respect to the target variable, in this case the secondary–market price. The objective is therefore not only to identify cards that are similar in their characteristics, but to find combinations of card attributes that are associated with systematically different price levels.

After fitting the regression tree, each card was assigned to a terminal node of the tree. These terminal nodes were interpreted as price–driven market segments, because the tree partitions the dataset by minimizing within–segment variation in log–transformed card prices. Segment profiles were then constructed by summarizing card–level characteristics within each terminal node, including average price, rarity composition, mana value, card type indicators, format legality, and selected text–derived mechanics. This allows the regression tree to be used not only as a predictive model, but also as an interpretable segmentation tool.

Since nearly all the cards in the selected sample are still legal for standard as well as other formats, the features referring to legality have been excluded form further segment profiling.

The resulting regression tree model, which has been obtained after a cost–complexity pruning based on cross–validation, has four leaf nodes, i.e. four segments. Such number of segments is easily managed and interpreted. The segment sizes are 896, 452, 258 and 51. See the final regression tree model in Figure 1. The final model uses card rarity value and total mana symbols as predictors. The segment drivers correspond quite well with the predictors in the LASSO model mentioned above, considering that the land cards are cards with zero mana symbols. The darker nodes in the plot mean higher (log–transformed) prices. The smallest segment of size 51 cards contains the cards with the highest predicted price value. This "Top value" segment consists of cards of rarity rare (perhaps surprisingly not mythic rare) and without mana symbols. The largest, "Mass" segment includes the cards of rarities common or uncommon and has the lowest predicted price. The second largest segment representing approximately 27% of the selected sample has sort of medium predicted price between the predicted price of the Mass segment and the segment of mythic rare cards, which have the second largest predicted price.

It can be observed in Figure 2 showing the boxplots for prices by segment, that the segments are distinguishing the cards with respect to price well. See also Table 2 which gives basic descriptive statistics of price in USD by segment.

2.3 Segment Profiles

Getting segment profiles is an important step in cluster analysis in case that there is a demand for interpretable segments. Even card characteristics which have not been included in the final regression tree predictors should be used for segment profiling. Table 3 gives the rarity profiles of the segments. Segment 3 are mythic rare cards, Segment 2 and Segment 4 are rare cards and Segment 1 are common and uncommon cards. Table 4 gives the profiles of the segments for other interesting features which may be helpful for interpreting the segments. All cards in the smallest segment (Segment 4) are Land cards. Segments 1 and 4 are cards with shorter text, Segment 4 has the highest share of graveyard. Segments 2 and 3 may be considered similar, the remarkable difference is in rarity - mythic rare in Segment 3 versus rare in Segment 2 and in slightly lower mana value and color count and higher share of artifacts for Segment 3.

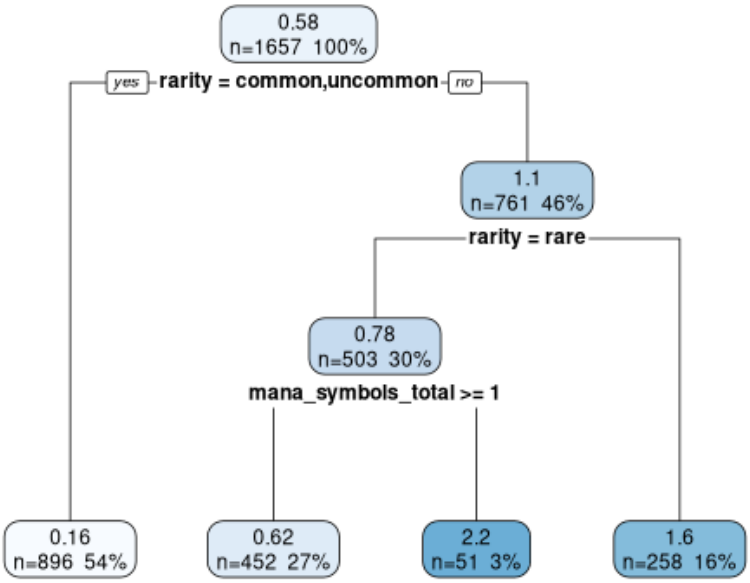


Figure 1 Regression Tree for Price-Driven Segmentation

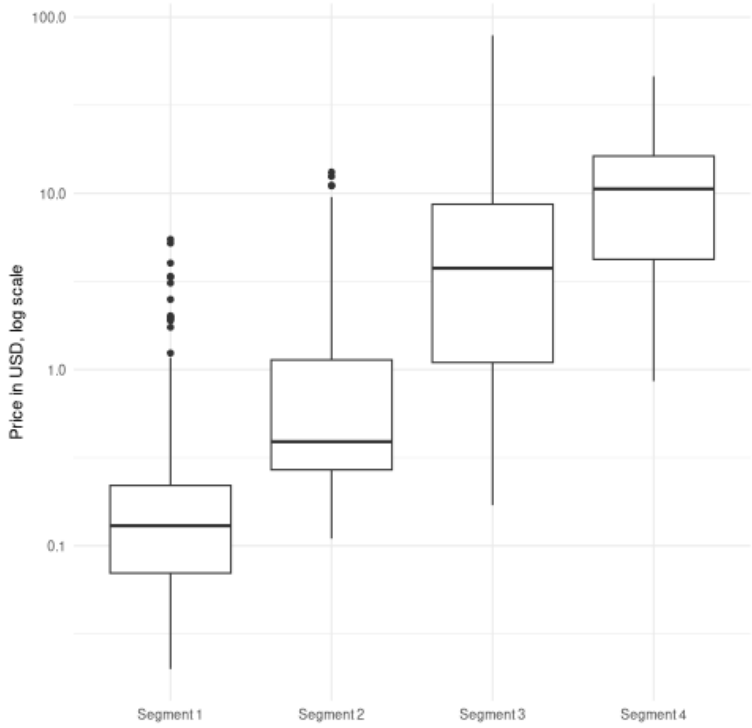


Figure 2 Card Price by Segment

Segment	Mean	Median	Min	Max
Segment 1	0.20	0.13	0.02	5.5
Segment 2	1.35	0.39	0.11	13.2
Segment 3	8.13	3.76	0.17	78.9
Segment 4	12.20	10.60	0.86	46.1

Table 2 Prices in USD by segment

segment	rarity	count	share
Segment 1	common	440	0.49
Segment 1	uncommon	456	0.51
Segment 2	rare	452	1
Segment 3	mythic	258	1
Segment 4	rare	51	1

Table 3 Rarity by segment

Segment	Segment 1	Segment 2	Segment 3	Segment 4
Size	896	452	258	51
Mean mana value	2.58	3.18	4.03	0
Mean color count	0.97	1.37	1.21	0
Mean text length	162	236	239	130
Share: creature	0.49	0.62	0.55	0
Share: instant	0.16	0.05	0.02	0
Share: sorcery	0.10	0.13	0.06	0
Share: artifact	0.08	0.08	0.18	0
Share: enchantment	0.11	0.19	0.16	0.04
Share: land	0.13	0	0.02	1
Share: legendary	0.02	0.30	0.39	0.10
Share: draw	0.15	0.25	0.21	0.04
Share: exile	0.11	0.21	0.24	0
Share: token	0.14	0.21	0.30	0.04
Share: graveyard	0.19	0.24	0.15	0.39

Table 4 Segment profiles

3 Conclusion

While prior research on trading card markets has mainly examined price prediction and forecasting, this study applies a regression–tree–based approach for price–driven segmentation. The method partitions cards into interpretable market segments according to the features that best explain variation in secondary–market prices. This provides a complementary perspective to purely predictive models by emphasizing segment interpretation rather than only predictive accuracy and answers the question Which interpretable card segments are associated with different price levels? The price–driven regression tree–based clustering analysis reveals interpretable and meaningful card groupings that correspond to gameplay roles and price segments, highlighting structural regularities in the market. The segmentation rules obtained from our model are based just on a small number of characteristics and are easy to implement and interpret. The segments have distinct interpretable characteristics with respect to other features i.e. not included in the rules for segment assignment too. Overall, the study illustrates how modern data analysis techniques can be applied to trading card game datasets to gain insights into pricing mechanisms and market structure, and may provide framework applicable to other collectible markets.

Acknowledgements

The paper has been produced using the institutional support for long term development at Prague University of Economics and Business.

References

Abrahamse S, Berghman M, Schaap J (2026) Investing in meaning: justifying card game collecting in an age of frictionless consumption. Journal of Cultural Economy 19(1):135–150. https://doi.org/10.1080/17530350.2025.2559315, https://doi.org/10.1080/17530350.2025.2559315

Bertram T, Fürnkranz J, Müller M (2024) Learning with generalised card representations for "magic: The gathering". pp 1–8, https://doi.org/10.1109/CoG60054.2024.10645602

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1):1–22. https://doi.org/10.18637/jss.v033.i01

R Foundation for Statistical Computing (2026) R: A language and environment for statistical computing. https://www.R-project.org

Sakaji H, Kobayashi A, Kohana M, et al (2020) Card price prediction of trading cards using machine learning methods. In: Barolli L, Nishino H, Enokido T, et al (eds) Advances in Networked-based Information Systems. Springer International Publishing, Cham, pp 705–714

Scryfall (2026) Scryfall bulk data. https://scryfall.com/docs/api/bulk-data

Therneau T, Atkinson B (2026) rpart: Recursive Partitioning and Regression Trees. https://doi.org/10.32614/CRAN.package.rpart, https://CRAN.R-project.org/package=rpart, r package version 4.1.27

Zilio F, Prates M, Lamb L (2018) Neural networks models for analyzing magic: The gathering cards. In: Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, p 227–239, https://doi.org/10.1007/978-3-030-04179-3_20, https://doi.org/10.1007/978-3-030-04179-3_20

The economic impact of natural disasters on NUTS2 regions in EU

Daniil Khokhrunov[1]

Abstract. We explore the association between acute physical risk and economic growth in NUTS2 EU regions. Information on natural disasters is obtained from the EM-DAT dataset. The regressor of main interest is the number of climate-related disasters in the region during the year. Regional economic data are sourced from the Eurostat database. The baseline fixed-effects model showed a weakly significant positive effect of natural disasters on GDP per capita growth in NUTS2 regions. However, the results are not robust against cross-sectional dependence and serial correlation. To address these problems, we use common correlated effects estimators and find that, after accounting for them, the effects of natural disasters on NUTS2 EU regions are not statistically significant.

Keywords: Climate risk, Economic growth, Natural disaster, Panel data

JEL Classification: Q54

1 Introduction

Climate risk is becoming an increasingly important topic, particularly following the Paris Agreement. The literature typically distinguishes between two main types of climate risk:

- Transition risks are the risks arising from the transition to the green economy. Those may include new taxes, changes in preferences, innovations, and so on.
- Physical risks arise from nature itself. Those are risks associated with changes in weather and climate (Pitman et al, 2022)
  - Chronic risks refer to long-term changes in climate, which can pose significant challenges if not addressed in time. Those include, for instance, rising sea level and rising temperature.
  - Acute risks arise from extreme natural events or from increased frequency and severity of such events.

This study focuses on the impact of acute physical risk on the economy. We will explore the consequences of climate-related extreme events for regional economic growth in EU countries.

Although the economic consequences of extreme events are widely covered in the literature, there is no consensus regarding the effect of natural disasters on economic outcomes. Some authors find a negative effect (Hsiang and Jina, 2014). For example, (Shabnam, 2014) writes: "Natural disasters often undo many years of physical as well as human capital accumulation and can lead to slowing down the speed of convergence towards a steady state economy within the country context". Other authors find either no effect (Cavallo et al, 2010) or even a positive effect (Roth Tran, 2020) of climate-related disasters. Such different conclusions may be due to different methodologies, data (regional within a country or across a whole region), a defined time horizon (short-term or long-term), etc. For example, (Usman et al, 2025) examines the impact of heatwaves, floods, and droughts on the regional economic output of EU Member States. On the other hand, (Skidmore and Toya, 2002) examines the effects of natural disasters worldwide, using historical information from (Davis, 1992). It is not strange that such different settings would produce different conclusions.

Similar studies to ours (Di Marcoberardino and Cucculelli, 2024; Usman et al, 2025) find a negative effect of natural disasters on NUTS3 regions in the EU. Di Marcoberardino also pays close attention to the impact of regional governance on the handling of the consequences of natural disasters. Usman et al. consider only selected types of climate events, but also examine the impact of natural disasters on components of economic output.

This study contributes to the literature by examining the impact of natural disasters at the NUTS-2 regional level across EU countries, thus providing evidence at a more aggregated level than similar studies focusing on EU regions and a less aggregated level than those that study the impact at the cross-country level. We suggest using the correlated common effects estimator as a cross-sectional dependence robust approach for estimating the effect of natural disasters on economic growth.

[1] Prague University of Economics and Business, Department of Econometrics, nám. Winstona Churchilla 1938/4, 120 00 Praha 3-Žižkov, khod00@vse.cz

2 Data

To empirically estimate the effects of natural disasters on the EU's macroeconomic environment, data from multiple sources must be combined.

The most popular database of natural disasters is the Emergency Events Database (EM-DAT). The database contains information from different sources and is available from 1900. Pre-2000 data are particularly prone to reporting biases; therefore, we use only information from 2000 onward. For a disaster to be included in the database, one of the four criteria must be satisfied:

- 10 fatalities;
- 100 affected people;
- a declaration of state of emergency;
- a call for international assistance.

The database is sometimes criticised. One concern is that the events are self-reported. Hence, for political or other reasons, some events may not be included in the database. The second concern is that the estimate of damages is affected by the economic state of the region. Nevertheless, the database remains the most popular source of natural disaster events in the literature. Moreover, the chosen region of EU countries has the advantage that its countries are not significantly different in economic terms. In addition, the variety of disaster types is rather limited due to the close geographic location. Since the main focus of this study is climate-related disasters, we include only those disasters considered climatological, hydrological, or meteorological.

Since the EM-DAT database contains information at the country level, we need to assign events to NUTS2 regions. Following the methodology of (Di Marcoberardino and Cucculelli, 2024), only events that can unambiguously be attributed to a specific NUTS2 region are considered. We prefer this strict approach, since wrongly-assigned events may bias our estimates. Firstly, observations with known coordinates are assigned by their location. Then we match EM-DAT location names with NUTS2 and NUTS3 region names, and if the names are fully matched, the locations are assigned to the corresponding NUTS region. Finally, we use the GADM (Database of Global Administrative Areas) information presented in the table. Events that could only be matched to a country or NUTS1 level were excluded. We determine the location of every administrative region if information is available. This left us with 286 total events.

Disaster intensity is often used as a model input variable. The costs of the event (Hallegatte et al, 2007) or the magnitude (Hsiang and Jina, 2014) are popular measures. As an alternative, the affected population or the number of deaths is sometimes used (Becker and Mauro, 2006). However, as discussed in the literature, the economic impact of natural disasters in the EM-DAT database is subject to self-reporting bias, leading to inconsistent estimates and greater economic losses in more developed countries. The number of affected or dead persons is not available for most disasters in our data. This leads us to use the number of disasters in the region per year as a measure of disaster intensity. We acknowledge that the annual count of disasters is an imperfect proxy for economic exposure. Ideally, we would use estimated damages or the affected population. However, given the limitations discussed above, we prefer to use a more reliable option.

The number of disasters does not exhibit a time trend (Figure 1). However, the distribution of disasters across regions is uneven. With a mean of 1.018 disasters, most NUTS2 regions did not report any climate-related natural disasters during the observation period. However, some regions reported a large number of disasters. The highest number is for Languedoc-Roussillon, a former administrative region in southern France (11), Jadranska Hrvatska - Croatian coastal region (9), and Małopolskie, a southern Polish region (9).

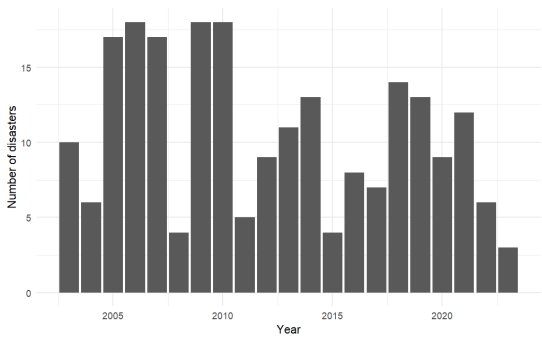


Figure 1 Distribution of disasters over time, 2003-2023

The economic information for the EU regions is obtained from the Eurostat database. We collect GDP and population data, as this is information of primary interest. For the model to be unbiased, we also need to include control variables. Rather limited information is available on the regional level. In addition, many data sources contain significant information gaps. For these reasons, we chose to use three control variables in this research:

- Employment in thousand hours worked. This measure is the most complete among other labour indicators.
- Gross fixed capital formation for the total economy, measured with all the activities included.
- Population density, calculated using available information about population.

We use data from 2003 to 2023. No complete information is available for 2024 yet, and for years before 2003, many GDP values are missing for NUTS2 regions. Clear data is necessary for analysis. However, even in recent time periods, some information is missing. Since the population usually exhibits a trend, we complete the population values by extrapolating the observations using the country-level growth rate. For other variables, missing observations are dropped. Regions with many consecutive missing values are dropped because the information is unreliable. Gross fixed capital formation is highly correlated with employment. Keeping both variables creates multicollinearity (VIF > 14). We choose to keep employment due to better coverage and stability. Hence, only two control variables remain in the model.

3 Methodology

As a baseline specification, we estimate a fixed effects panel model to examine the association between natural disasters and economic growth, following (Islam, 1995). This is one of the most common approaches to studying the impact of natural disasters on the economy (e.g., Felbermayr and Gröschl, 2014; Noy, 2009). To ensure the model is unbiased, we include additional control variables.

$$\Delta \ln y_{it} = \gamma \ln y_{it-1} + \beta d_{it} + \boldsymbol{\rho}' \boldsymbol{x}_{it} + a_i + a_t + \epsilon_{it} \tag{1}$$

Where a_i and a_t are region- and time-specific effects. d_{it} is the number of natural disasters in the region i in the year t. $\boldsymbol{x}_{it}$ is a vector of control variables and y_{it} is the GDP per capita. The primary interest lies in estimating the coefficient β, which indicates the impact of the number of natural disasters on the region's economic output.

The inclusion of a lagged dependent variable violates the strict exogeneity assumption required for a fixed effects (FE) estimator. As a result, the FE estimator is biased and inconsistent (Nickell bias) in panels with a limited time dimension. Therefore, the models' (1) estimates serve primarily as a baseline specification rather than a source of causal inference.

In addition, natural disasters, even when they are attributed to a specific region, may still have spillover effects on nearby regions (Lima and Barbosa, 2019), leading to cross-sectional correlation. We test for cross-sectional dependence using Pesaran's CD test, which does not rely on strict exogeneity and thus remains valid even in specifications with a lagged dependent variable.

To control for the effects described above, we propose a model (2), which uses the common correlated effects (CCE) estimator, proposed by (Pesaran, 2006). The key idea of the CCE approach is to approximate the unobserved common factors $\boldsymbol{F}_t$ (such as macroeconomic conditions or climate patterns affecting all EU regions simultaneously) by including cross-sectional averages of the dependent and independent variables as additional regressors. This augmentation effectively removes the bias that arises when these common factors correlate with our disaster variable, a threat that standard fixed effects cannot handle.

$$\Delta \ln y_{it} = \gamma \ln y_{it-1} + \beta d_{it} + \boldsymbol{\rho}' \boldsymbol{x}_{it} + a_i + \boldsymbol{\lambda}_i' \boldsymbol{F}_t + u_{it} \tag{2}$$

Where $\boldsymbol{F}_t$ is a vector of unknown common factors and $\boldsymbol{\lambda}_i$ is a vector of factor loadings. The model (2) is an augmented version of the model (1), where $\epsilon_{it} = \boldsymbol{\lambda}_i' \boldsymbol{F}_t + u_{it}$.

The model (2) should mitigate bias arising from unobserved common factors that induce cross-sectional dependence. As shown in (Gaibulloev et al, 2014), the failure to account for such a dependence may inflate the bias and lead to inconsistent estimates. We implement two variants: the CCE Pooled (CCEP) estimator, which assumes homogeneous slopes across regions, and the CCE Mean Group (CCEMG) estimator, which allows the disaster impact to vary across regions—a more flexible and realistic assumption given the heterogeneous nature of regional economies. If the CCEMG estimate differs substantially from the CCEP estimate, this suggests that the effect of natural disasters is not uniform across the EU, making the mean-group approach more reliable.

A rather important note is that, for cross-sectional averages to be a good proxy, we need a balanced panel. This requirement arises because the CCE estimator constructs cross-sectional averages at each point in time t using the same set of regions. Missing observations would change the composition of these averages over time, introducing inconsistency. For that reason, in the estimation of the model (2), we discard from the data all regions where, at least in one time period, a variable was missing.

4 Results

The results of the model (1) are presented in Table 1. The estimates for all control variables are statistically significant (the model specification includes y_{it-1}, GDP per capita). The value for the variable Disasters, which is of primary interest, is estimated to be close to zero and weakly significant. Moreover, the estimated coefficient is positive, which is unexpected for the short-term effect.

Variable	Estimate	Std.Error	t-value	Pr(>|t|)	Significance
$\ln y_{it-1}$	-0.1030	0.0313	-3.2939	0.0009	***
$Disasters$	0.0051	0.0025	1.9975	0.0458	*
$\ln(Density)$	-0.3180	0.0862	-3.6907	0.0002	***
$\ln(Employment)$	0.2150	0.0461	4.6634	0.0000	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1 Results of the fixed effect model

However, it was noted that including a lagged dependent variable as a regressor invokes the Nickell Bias of the model. The bias may be even more noticeable, since $N > T$ in our sample. The Nickell bias in this case affects coefficient bias (especially under cross-sectional dependence), and it may also lead to distorted standard errors and over-rejection of null hypotheses. This implies that even weakly significant disaster effects may be overestimated.

Pesaran's CD test indicated cross-sectional dependence (p-value < 0.001). Driscoll–Kraay standard errors are used to obtain inference robust to serial and cross-sectional dependence. We report them in the second column of the Table 1.

Even the weakly positive effect of natural disasters is consistent with reconstruction-driven short-run growth effects documented in the literature. Some authors also found a positive effect of natural disasters. There may be several reasons for this result. The positive effect on the region's economic output may be due to additional investments following the disaster in the form of financial assistance. In addition, the renovation of some capital may occur when old and damaged capital is replaced. Next, the effects of natural disasters are stronger on developing than on advanced countries (Fomby et al, 2013), as is the case of European countries and their regions.

Group	Mean growth	Mean disasters	Mean density
Belgium	0.0322	0.0545	5.99
Rest of the data	0.0356	0.0483	4.86

Table 2 Comparison of statistics for Belgium and the rest of the sample

The tests confirmed our assumptions of cross-sectional dependence. We present the results for both CCE models. Both models were estimated on the truncated version of the data, in which 11 Belgian regions were excluded because their lagged GDP per capita was missing for the first year. There are no substantial structural differences between Belgium and the rest of the sample (Table 2).

Variable	Estimate	Std.Error	z-value	Pr(>|z|)	Significance
$\ln y_{it-1}$	-0.4287	0.0499	-8.5841	0.0000	***
$Disasters$	0.0047	0.0025	1.8983	0.0576	.
$\ln(Density)$	-0.3517	0.0839	-4.1910	0.0000	***
$\ln(Employment)$	0.3429	0.0655	5.2325	0.0000	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3 Results of the CCEP model

Variable	Estimate	Std.Error	z-value	Pr(>|z|)	Significance
$\ln y_{it-1}$	-0.7134	0.0238	-30.0352	0.0000	***
$Disasters$	0.0018	0.0013	1.3862	0.1657	
$\ln(Density)$	-0.3132	0.1697	-1.8456	0.0650	.
$\ln(Employment)$	0.6302	0.0450	14.0185	0.0000	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4 Results of the CCEMG model

As shown in Tables 3 and 4, the effect of natural disasters on economic growth became even more negligible after controlling for cross-sectional dependence. Hence, the initial conclusion of a positive effect of natural disasters is not robust against cross-sectional dependence and Nickell bias.

When we examine only the CCE model results, we see differences between the pooling and mean-group approaches. The disaster's effect in CCEMG Table 4 is weaker and is not statistically significant. The difference between CCEP and CCEMG reflects slope heterogeneity. The weaker and insignificant effect in CCEMG suggests that disaster impacts differ significantly across regions. The loss of statistical significance after accounting for cross-sectional dependence suggests that the baseline fixed effects results were likely driven by omitted common shocks. Given the limited significance of previous results and robust models, we conclude that there is no statistically significant association between climate-related disasters and the economic growth of NUTS2 regions in the EU.

A critical caveat is that our conclusions are conditional on the available data, the chosen disaster frequency measure, and the modelling strategy. Future research with richer data on disaster intensity at the NUTS2 level could provide different results.

5 Conclusions

In this study, we explored the association between natural disasters and the economic growth of the NUTS2 EU regions. The disaster data were obtained from the popular EM-DAT database. We included only climate-related disasters in the EU countries. We included only those disasters which could be attributed to the exact NUTS2 region. The economic information about the regions was taken from the Eurostat database. We chose the information that is available and complete for the period 2003-2023. The number of disasters during the year was used as the primary variable of interest.

As a baseline, the fixed panel approach was used. The model suggests a weakly positive effect of the climate-related disasters on the economic growth of NUTS2 regions. However, the estimates are biased because of the presence of a lagged dependent variable and cross-sectional dependence. As a result, we reported the original results with Driscoll-Kraay standard errors and calculated two robust models.

As a robust approach, the common correlated effects estimator was chosen. This model remains consistent under cross-sectional dependence. The CCE models showed even weaker significance of the impact of the disasters on the economy. The CCEMG model showed that the effect of natural disasters is not statistically significant at any level. Furthermore, the results confirmed that the effect of natural disasters is heterogeneous across NUTS2 regions in the EU.

Among the reasons for these results may be that European countries and regions are economically developed. Hence, there may be more facilities to mitigate the negative impact. Compared with studies of NUTS3 regions, when the effect was negative, this may be due to greater aggregation. When the NUTS3 region is affected by natural disasters, it suffers significant losses relative to its size. However, at the NUTS2 level, the relative size of the disaster would be smaller because the region itself is larger, and hence the effect is smaller. Moreover, the NUTS2 region has more resources to deal with the consequences of natural disasters. If the disaster is large enough to affect whole NUTS2 the effect may be big enough for the whole country and other regions take part in the restoration process.

There are many paths of further research in the field. It will be useful to examine how disasters of varying intensities affect economic growth. It is likely that only big hazards influence the region's output. In addition, our results showed cross-sectional dependence of the effect, hence a spatial model may be used for further research. In addition, many studies have indicated heterogeneity in the impacts of different disaster types. The main problem in research on regional impacts is data availability. Much less information is available at the regional level than at the country level, both in terms of economics and on disasters.

Acknowledgements

The research project was supported by Grant No. F4/18/2024 of the Internal Grant Agency, Faculty of Informatics and Statistics, Prague University of Economics and Business.

References

Becker T, Mauro P (2006) Output drops and the shocks that matter. Tech. Rep. 926227, International Monetary Fund, https://papers.ssrn.com/abstract=926227

Cavallo EA, Noy I, Pantano J, et al (2010) Catastrophic natural disasters and economic growth. IDB Publications https://doi.org/10.18235/0010949

Davis LLA (1992) Natural disasters: from the Black Plague to the eruption of Mt. Pinatubo. New York : Facts on File, http://archive.org/details/naturaldisasters0000davi

Di Marcoberardino D, Cucculelli M (2024) "natural" disasters and regional governance: Evidence from european NUTS-3 regions. Papers in Regional Science 103(1):100003. https://doi.org/10.1016/j.pirs.2024.100003

Felbermayr G, Gröschl J (2014) Naturally negative: The growth effects of natural disasters. Journal of Development Economics 111:92–106. https://doi.org/10.1016/j.jdeveco.2014.07.004

Fomby T, Ikeda Y, Loayza NV (2013) The growth aftermath of natural disasters. Journal of Applied Econometrics 28(3):412–434. https://doi.org/https://doi.org/10.1002/jae.1273

Gaibulloev K, Sandler T, Sul D (2014) Dynamic panel analysis under cross-sectional dependence. Political Analysis 22(2):258–273. https://doi.org/doi:10.1093/pan/mpt029

Hallegatte S, Hourcade JC, Dumas P (2007) Why economic dynamics matter in assessing climate change damages: Illustration on extreme events. Ecological Economics 62(2):330–340. https://doi.org/10.1016/j.ecolecon.2006.06.006

Hsiang SM, Jina AS (2014) The causal effect of environmental catastrophe on long-run economic growth: Evidence from 6,700 cyclones. NBER Working Papers 20352, National Bureau of Economic Research, Inc, https://doi.org/None, https://ideas.repec.org/p/nbr/nberwo/20352.html

Islam N (1995) Growth empirics: A panel data approach. The Quarterly Journal of Economics 110(4):1127–1170. https://doi.org/10.2307/2946651

Lima RCdA, Barbosa AVB (2019) Natural disasters, economic growth and spatial spillovers: Evidence from a flash flood in brazil. Papers in Regional Science 98(2):905–925. https://doi.org/10.1111/pirs.12380

Noy I (2009) The macroeconomic consequences of disasters. Journal of Development Economics 88(2):221–231. https://doi.org/10.1016/j.jdeveco.2008.02.005

Pesaran MH (2006) Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74(4):967–1012. https://doi.org/10.1111/j.1468-0262.2006.00692.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2006.00692.x

Pitman A, Fiedler T, Ranger N, et al (2022) Acute climate risks in the financial system: examining the utility of climate model projections. Environmental Research: Climate 1. https://doi.org/10.1088/2752-5295/ac856f

Roth Tran DJWBrigitte (2020) The local economic impact of natural disasters. Federal Reserve Bank of San Francisco, Working Paper Series pp 1.000–61. https://doi.org/10.24148/wp2020-34

Shabnam N (2014) Natural disasters and economic growth: A review. International Journal of Disaster Risk Science 5(2):157–163. https://doi.org/10.1007/s13753-014-0022-5

Skidmore M, Toya H (2002) Do natural disasters promote long-run growth? Economic Inquiry 40(4):664–687. https://doi.org/10.1093/ei/40.4.664

Usman S, González-Torres Fernández G, Parker M (2025) Going nuts: The regional impact of extreme climate events over the medium term. European Economic Review 178:105081. https://doi.org/https://doi.org/10.1016/j.euroecorev.2025.105081

Missing Data and Inference in Survey-Weighted Ordered Logistic Models

Ján Kobida[1], Ľubica Bartová[2]

Abstract. This paper examines how alternative missing-data treatments influence statistical inference in survey-weighted ordered logistic regression models. Using pooled data from the Health Behaviour in School-aged Children (HBSC) study across 22 European Union Member States, the analysis compares complete-case estimation with two multiple imputation approaches: Multiple Imputation by Chained Equations (MICE) and a machine-learning-based framework combining Light Gradient Boosting Machine with predictive mean matching (LGBM-PMM) implemented via the miceforest package. The comparison focuses on coefficient stability, inferential consistency, estimation precision and measures of missing-data uncertainty. All 18 substantive coefficients retained their signs across the three specifications, while one coefficient that was not significant in the complete-case model became statistically significant under both imputation approaches. LGBM-PMM produced higher standard errors on average and substantially greater missing-data uncertainty, reflected in a higher average RVI and a larger maximum FMI, than conventional MICE. The findings indicate directional stability but not complete inferential equivalence across missing-data treatments and support the comparison of alternative imputation approaches as a sensitivity analysis in incomplete complex survey data.

Keywords: missing data, multiple imputation, machine-learning imputation, complex survey data, survey-weighted ordered logistic regression

JEL Classification: C25, C81, C83

1 Introduction

Missing data are common in large social, health and economic surveys because of item nonresponse, measurement difficulties or respondent attrition. Ignoring them may reduce statistical efficiency and bias inference when missingness is related to observed characteristics (Little and Rubin, 2019). Multiple imputation generates plausible values conditional on observed information and combines estimates across completed datasets using Rubin's rules (Rubin, 1987). Multiple Imputation by Chained Equations (MICE) is widely used because it accommodates variables measured on different scales (Azur et al., 2011; Carpenter and Kenward, 2013; van Buuren, 2018).

Traditional MICE implementations rely mainly on parametric conditional models, which may be less effective when relationships are nonlinear, interactions are complex or patterns are heterogeneous (Shah et al., 2014). Tree-based methods can model such structures without their explicit specification. Light Gradient Boosting Machine (LightGBM) is an efficient gradient-boosting method based on decision-tree ensembles (Ke et al., 2017). Tree-based imputation and conventional MICE may differ not only in predictive performance but also in their consequences for coefficient estimation and inference (Schwerter et al., 2024).

These issues are particularly relevant for complex survey data, where estimation must account for survey weights, clustering and stratification (Heeringa et al., 2017). However, evidence comparing imputation engines in survey-weighted ordinal regression models remains limited. The principal concern is whether coefficient magnitude, direction and statistical significance remain stable across missing-data treatments.

Building on our earlier comparison of complete-case estimation and conventional MICE using the same HBSC data (Kobida and Bartová, 2026), this study extends the comparison to LGBM-PMM and evaluates coefficient stability, statistical significance, estimation precision and missing-data uncertainty within the same survey-weighted ordered logistic model. Complete-case estimation serves as a practical benchmark, and the comparison is interpreted as a sensitivity analysis rather than as evidence that any method is unbiased or uniquely robust.

[1] Slovak University of Agriculture in Nitra, Faculty of Economics and Management, Nitra, Slovakia, xkobida@uniag.sk
[2] Slovak University of Agriculture in Nitra, Faculty of Economics and Management, Nitra, Slovakia, lubica.bartova@uniag.sk

2 Data and Methods

2.1 Data and variables

The empirical application examines associations between adolescents' body-weight status and selected demographic, socioeconomic and health-behaviour characteristics. It is used to assess whether coefficient estimates, statistical significance, estimation precision and measures of missing-data uncertainty remain stable under alternative missing-data treatments.

The analysis combines repeated cross-sectional observations from 22 European Union Member States across four survey rounds, denoted by their end years as 2006, 2010, 2014 and 2018. The sample includes old EU Member States (BE, DK, FI, FR, DE, GR, IE, LU, NL, PT, ES and SE) and new EU Member States (BG, CZ, EE, HU, LV, LT, PL, RO, SK and SI). The pooled dataset consists of 439,466 individual-level observations from adolescents aged 11, 13 and 15 years. Of these, 336,217 observations are included in the complete-case analysis. Restricted-use international microdata for the selected survey rounds were provided by the HBSC Data Management Centre (n.d.). The analysis used the harmonized sampling weights and cross-national survey-design identifiers supplied in the international HBSC files. These variables were retained when the selected country–survey-round samples were pooled, and the sampling weights were used without additional rescaling at the estimation stage.

The dependent variable is an ordinal indicator of body-weight status derived from self-reported height and weight. Respondents are classified as underweight or normal weight, overweight, or obese using age- and sex-specific BMI-for-age thresholds from the WHO growth reference for school-aged children and adolescents (de Onis et al., 2007). Explanatory variables include survey year, age group, sex, socioeconomic status, parental employment characteristics, breakfast consumption frequency and physical activity. Socioeconomic status is represented by a standardized index constructed as the first principal component of the number of cars and computers in the household and the number of family holidays during the previous year. Father's employment status is measured by a six-category indicator, with care/home denoting fathers who care for others or are full-time at home. Breakfast consumption is measured by the number of weekdays on which the respondent eats breakfast, and physical activity by the number of days per week with at least 60 minutes of activity.

Because the HBSC survey employs a complex sampling design, all regression models account for the sampling weights provided by the HBSC, school-level clustering and stratification by subregion, with singleton strata handled using the centred adjustment, to obtain design-consistent inference (Dey et al., 2025; Heeringa et al., 2017).

Missing values affect the dependent variable and six explanatory variables included in the imputation models. Body-weight status has the highest missingness rate (14.59%), while missingness among the explanatory variables ranges from 1.44% for weekday breakfast consumption to 4.76% for father's employment status (Table 1).

Variable	Missing, n (%)	Variable	Missing, n (%)
Body-weight status	64,131 (14.59)	Cars in household	16,200 (3.69)
Father's employment status	20,915 (4.76)	Computers in household	16,128 (3.67)
Family holidays	16,544 (3.76)	Physical activity	9,889 (2.25)
Weekday breakfast consumption	6,350 (1.44)		

Table 1 Missingness in variables included in the imputation models

2.2 Missing data treatments

The analysis compares complete-case analysis (CC), Multiple Imputation by Chained Equations (MICE) and LightGBM-based MICE with predictive mean matching (LGBM-PMM). CC excludes observations with a missing value in any variable entering the regression model and yields a sample of 336,217 observations. It is treated as a practical benchmark for sensitivity analysis rather than as an unbiased reference, since its validity depends on the missing-data mechanism (Little and Rubin, 2019).

For each incomplete variable Y_j, chained-equations imputation draws missing values iteratively from

$$Y_{j,\text{mis}}^{(m,t)} \sim p_j\left(Y_j \mid \mathbf{Y}_{-j}^{(m,t)}, \mathbf{Z}; \boldsymbol{\theta}_j^{(m,t)}\right) \tag{1}$$

where $\mathbf{Y}_{-j}$ contains the remaining incomplete variables, $\mathbf{Z}$ denotes fully observed predictors, $p_j(\cdot)$ is the conditional imputation model for variable Y_j, and $\boldsymbol{\theta}_j^{(m,t)}$is its parameter vector. The index j identifies the incomplete variable, $m = 1, \dots, M$ indexes completed datasets and t indexes iterations.

Seven variables are imputed: body-weight status, household cars, computers, family holidays, father's employment status, weekday breakfast consumption and physical activity. Survey year, sex and age group are included as fully observed predictors. Conventional MICE uses ordered logistic models for the ordinal variables and a multinomial logistic model for father's employment status. It generates $M = 20$ completed datasets after 10 burn-in iterations, using random seed 42 and augmentation to address potential perfect-prediction problems. Given the largest estimated FMI of 0.739, 20 imputations provide an approximate relative efficiency of 96.4% and were considered adequate for the comparative purpose of the study.

Both imputation methods used the same substantive predictor set to maintain a common specification. Survey weights, school identifiers and strata were retained in all completed datasets and used in the survey-weighted regression models, but they were not included as predictors in either imputation method. After imputation, the socioeconomic-status index is reconstructed separately in each completed dataset as the standardized first principal component of household cars, computers and family holidays.

In LGBM-PMM, the parametric conditional models are replaced by LightGBM ensembles implemented within the miceforest framework (Ke et al., 2017; Wilson et al., 2022). For variable Y_j, the prediction function is

$$\hat{\mu}_{ij} = F_j(\mathbf{x}_i) = \sum_{b=1}^{L} \eta\, h_{jb}(\mathbf{x}_i), \qquad i = 1, \dots, n, \tag{2}$$

where i indexes respondents, n is the sample size, $\hat{\mu}_{ij}$ is the predicted value of Y_j for respondent i, and $F_j(\mathrm{x}_i)$ is the corresponding LightGBM prediction function. $\mathbf{x}_i$ is the predictor vector for respondent i, h_{jb} denotes the b-th decision tree used to impute variable Y_j, L is the number of trees and η is the learning rate.

Predictive mean matching replaces direct model predictions with observed donor values. For a missing value Y_{ij}, the donor set consists of the d observed cases with predictions closest to $\hat{\mu}_{ij}$:

$$\mathcal{D}_{ij} = \underset{\substack{\mathcal{D}\subseteq\mathcal{O}_j \\ |\mathcal{D}|=d}}{\arg\min} \sum_{\ell\in\mathcal{D}} |\, \hat{\mu}_{ij} - \hat{\mu}_{\ell j} \,|, \qquad Y_{ij,\mathrm{mis}}^{(m)} = Y_{Kj,\mathrm{obs}}, \qquad K \sim \mathrm{Uniform}(\mathcal{D}_{ij}) \tag{3}$$

where $\mathcal{O}_j$ is the set of respondents with an observed value of variable Y_j, $\mathcal{D}_{ij}$ is the donor set for respondent i, and d is the donor-pool size. The index ℓ identifies a potential donor, while K denotes the donor randomly selected from $\mathcal{D}_{ij}$.

The LGBM-PMM procedure uses the same 439,466 observations, seven incomplete variables and predictor set as conventional MICE. It generates $M = 20$ completed datasets using five chained-equations iterations, $L = 50$ trees, a maximum tree depth of four, a fixed random seed of 12,345 and a donor pool of $d = 5$ observations. These settings were chosen before estimating the regression models to keep the imputation model sufficiently flexible without making it unnecessarily complex. All remaining LightGBM hyperparameters were left at their package defaults. The diagnostic checks did not reveal systematic trends across iterations or marked differences between the observed and imputed distributions.

The ordered logistic model is estimated separately in every completed dataset. Let $\widehat{\boldsymbol{\beta}}^{(m)}$ and $\widehat{\mathbf{U}}^{(m)}$ denote the coefficient vector and its covariance matrix in dataset m. Rubin's pooled estimator and total covariance matrix are

$$\bar{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\boldsymbol{\beta}}^{(m)}, \qquad \mathbf{T} = \bar{\mathbf{U}} + \left(1 + \frac{1}{M}\right)\mathbf{B} \tag{4}$$

where $\bar{\boldsymbol{U}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\boldsymbol{U}}^{(m)}$ is the within-imputation covariance matrix, $\mathbf{B}$ is the covariance matrix of the coefficient estimates across the M completed datasets, and T is the total covariance matrix (Rubin, 1987).

2.3 Evaluation of inferential stability

The same survey-weighted ordered logistic regression model is estimated under each missing-data treatment:

$$\Pr\,(Y_i \le j \mid \mathbf{x}_i) = \Lambda\left(\kappa_j - \mathbf{x}_i'\boldsymbol{\beta}\right), \qquad i = 1, \dots, n, \qquad j = 1, \dots, J-1 \tag{5}$$

where Y_i denotes the ordinal body-weight status of respondent i, $\mathbf{x}_i$ is the vector of explanatory variables, $\boldsymbol{\beta}$ is the coefficient vector and κ_j denotes the category thresholds. The function $\Lambda(z) = \exp(z)/[1 + \exp(z)]$ is the logistic cumulative distribution function, J is the number of outcome categories, and the left-hand side represents the probability that respondent i belongs to category j or below. Survey weights, school-level clustering and stratification by subregion are incorporated in the estimation.

The comparison is interpreted as a sensitivity analysis, with complete-case estimates serving as a benchmark rather than as true parameter values. Inferential stability is evaluated using the direction and magnitude of the substantive coefficients, changes in statistical significance at the 5% level, standard-error ratios and, for the imputed models, the relative variance increase (RVI) and fraction of missing information (FMI). The absolute relative coefficient change and standard-error ratio are defined as

$$\Delta_{k,a}^{\text{rel}} = 100\,\frac{|\,\hat{\beta}_{k,a} - \hat{\beta}_{k,CC}\,|}{|\,\hat{\beta}_{k,CC}\,|}, \quad R_{k,a}^{SE} = \frac{SE\,(\hat{\beta}_{k,a})}{SE\,(\hat{\beta}_{k,CC})}, \qquad a \in \{MICE, LGBM\text{-}PMM\} \tag{6}$$

where k indexes the substantive regression coefficients and a indexes the imputation method. The measure $\Delta_{k,a}^{\text{rel}}$ is the absolute percentage difference between an imputed-model coefficient and the corresponding CC estimate, while $R_{k,a}^{\text{SE}}$ compares their standard errors.

3 Results

Alternative missing-data treatments produced directionally consistent, but not fully equivalent, inferential results. All 18 substantive coefficients retained their signs under both MICE and LGBM-PMM relative to complete-case estimation, indicating directional stability across specifications (Table 2).

The magnitude of estimated coefficients was generally only modestly affected by imputation. Relative to complete-case estimation, conventional MICE produced a mean absolute relative coefficient change of 4.5%, while the corresponding value for LGBM-PMM reached 6.4%. The maximum changes were 37.5% under MICE and 47.8% under LGBM-PMM and, in both cases, concerned the coefficient for the paternal care/home category (Table 2). Because the complete-case estimate for this category was small, these percentages represented absolute coefficient differences of only 0.011 under MICE and 0.014 under LGBM-PMM.

Metric	MICE	LGBM-PMM
Coefficients retaining their sign vs CC, n	18	18
Mean absolute relative coefficient change vs CC (%)	4.50	6.40
Maximum absolute relative coefficient change vs CC (%)	37.50	47.76
Significance-status changes vs CC, n	1	1
Mean SE ratio relative to CC	0.954	1.180
Average RVI	0.2475	1.0335
Largest FMI	0.3512	0.7393

Table 2 Summary of inferential sensitivity across missing-data treatments

Note: Coefficient-change and standard-error measures are calculated relative to complete-case estimates. RVI and FMI are reported only for the multiple-imputation models.

Changes in inferential conclusions were limited. The coefficient for the paternal care/home category was small and positive in all three models: CC $\hat{\beta} = 0.029$ (95% CI: −0.008 to 0.066), MICE $\hat{\beta} = 0.040$ (95% CI: 0.003 to 0.077), and LGBM-PMM $\hat{\beta} = 0.043$ (95% CI: 0.008 to 0.078). The corresponding proportional-odds ratios ranged only from 1.03 to 1.04. The change at the 5% significance threshold therefore reflected small differences around the zero boundary rather than a substantively different association. All remaining coefficients retained their significance status.

The summary measures in Table 2 are based on all 18 substantive coefficients. The selected estimates in **Table 3** illustrate that coefficient sensitivity differed across predictors and imputation approaches. The coefficient for survey year 2014 changed more under MICE, whereas those for survey year 2010, father looking for work and breakfast consumption on three weekdays changed more under LGBM-PMM. In contrast, the socioeconomic-status coefficient remained comparatively stable across all three specifications.

The largest differences between the imputation approaches were observed in measures of missing-data uncertainty. The average RVI increased from 0.2475 under MICE to 1.0335 under LGBM-PMM, while the largest FMI increased from 0.3512 to 0.7393 (Table 2). These results indicate that LGBM-PMM attributed a substantially greater share of the total estimation variance to missing-data uncertainty.

Differences in estimation precision were also evident. The mean standard-error ratio reached 0.954 under conventional MICE and 1.180 under LGBM-PMM, indicating slightly lower standard errors under MICE but, on average, 18% higher standard errors under LGBM-PMM than under complete-case estimation (Table 2). Overall, coefficient directions were stable, but coefficient magnitude, statistical significance for one predictor, estimation precision and missing-data uncertainty remained sensitive to the imputation approach.

Variable	CC β (SE)	MICE β (SE)	LGBM-PMM β (SE)
Survey year: 2010	0.150*** (0.017)	0.149*** (0.016)	0.135*** (0.018)
Survey year: 2014	0.210*** (0.017)	0.232*** (0.015)	0.224*** (0.021)
Father: care/home	0.029 (0.019)	0.040* (0.019)	0.043* (0.018)
Father: looking for work	0.107*** (0.025)	0.114*** (0.024)	0.121*** (0.027)
Breakfast: three weekdays	−0.121*** (0.020)	−0.123*** (0.020)	−0.132*** (0.019)
Socioeconomic status	−0.100*** (0.005)	−0.098*** (0.005)	−0.097*** (0.006)

Table 3 Selected survey-weighted ordered logit estimates by missing-data treatment

Notes: Standard errors are reported in parentheses. $*p<0.05$, $**p<0.01$, $***p<0.001$. Reference categories are survey year 2006, father employed and no weekday breakfast. CC uses 336,217 observations; MICE and LGBM-PMM use 439,466 observations and 20 completed datasets.

4 Discussion and Conclusions

The results demonstrate that alternative imputation engines can affect the magnitude of estimated coefficients and, to a lesser extent, the statistical significance of individual predictors. All 18 substantive coefficients retained the same signs across complete-case estimation, conventional MICE and LGBM-PMM, indicating strong directional stability. However, the small positive coefficient for the paternal care/home category crossed the 5% significance threshold under both imputation approaches while remaining similar in magnitude across all three models; coefficient changes also differed across predictors. The results therefore support broadly consistent substantive interpretation, but not complete inferential equivalence across missing-data treatments.

The main distinction between the two imputation approaches concerned estimation precision and missing-data uncertainty. Conventional MICE produced standard errors close to those from complete-case estimation, whereas LGBM-PMM produced standard errors that were, on average, approximately 18% higher than those from complete-case estimation. LGBM-PMM also produced a substantially higher average RVI and a larger maximum FMI. Thus, the greater flexibility of the tree-based imputation model did not translate into lower inferential uncertainty in this application. This is consistent with the broader literature showing that machine-learning imputation may reproduce complex relationships while still yielding different inferential properties from conventional parametric imputation (Shah et al., 2014; Schwerter et al., 2024).

The findings also show why complete-case estimates should be interpreted only as a practical benchmark. Differences relative to complete-case estimation cannot be regarded as estimates of bias because the true parameter values are unknown and complete-case results may themselves be affected by selective nonresponse. The comparison should therefore be viewed as a sensitivity analysis of coefficient magnitude, statistical significance, precision and missing-data uncertainty rather than as evidence that one approach is universally superior.

Several limitations remain. The analysis is based on one empirical dataset and does not use a simulation design in which the true parameters and missing-data mechanism are controlled. The validity of both imputation approaches depends on the adequacy of the predictor set and the conditional missing-at-random assumption, which cannot be tested directly. The proportional-odds assumption was not assessed separately, as the ordered-logit specification was used as a common inferential framework for comparing the missing-data treatments. The pooled specification does not explicitly model country-level heterogeneity; therefore, the findings concern inferential sensitivity within the selected model rather than country-specific relationships. The findings may also differ under other levels or

patterns of missingness, variable distributions and survey designs. Consequently, the results should not be generalized as a universal ranking of MICE and LGBM-PMM.

Overall, the direction of the estimated associations was stable across missing-data treatments, but coefficient magnitude, the statistical significance of one small coefficient, estimation precision and missing-data uncertainty were sensitive to the imputation engine. LGBM-PMM did not provide a clear inferential advantage over conventional MICE in this application. For empirical analyses of incomplete complex survey data, the practical implication is to evaluate alternative imputation approaches through sensitivity analysis rather than selecting a machine-learning method solely on the basis of its predictive flexibility.

Acknowledgements

This research was supported by the APVV-24-0412 project: Sustainable and Healthy Food Consumption: Effects of Socio-economic Factors and Policies on Eating Behaviour, Obesity, and Food System.

References

Azur MJ, Stuart E A, Frangakis C, Leaf P J (2011) Multiple imputation by chained equations: What is it and how does it work? International Journal of Methods in Psychiatric Research 20(1):40–49. https://doi.org/10.1002/mpr.329

Carpenter J R, Kenward M G (2013) Multiple Imputation and its Application. Wiley, Chichester. https://doi.org/10.1002/9781119942283

de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J (2007) Development of a WHO growth reference for school-aged children and adolescents. Bulletin of the World Health Organization 85(9):660–667. https://doi.org/10.2471/BLT.07.043497

Dey D, Haque M S, Islam M M, Aishi U I, Shammy S S, Mayen M S A, Noor S T A, Uddin M J (2025) The proper application of logistic regression model in complex survey data: A systematic review. BMC Medical Research Methodology 25:15. https://doi.org/10.1186/s12874-024-02454-5

HBSC Data Management Centre (n.d.) Health Behaviour in School-aged Children (HBSC) international data files: 2005/06, 2009/10, 2013/14 and 2017/18 survey rounds [restricted-use data set]. Department of Health Promotion and Development, University of Bergen, Bergen. https://www.hbsc.org/

Heeringa S G, West B T, Berglund P A (2017) Applied Survey Data Analysis, 2nd ed. Chapman & Hall/CRC, Boca Raton. https://doi.org/10.1201/9781315153278

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: A highly efficient gradient boosting decision tree. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30. Curran Associates, Red Hook, pp 3146–3154

Kobida J, Bartová Ľ (2026) Multiple imputation and inference in complex survey data: Evidence from HBSC. In: Quantitative Methods in Economics: Multiple Criteria Decision Making XXIII. Vydavateľstvo EKONÓM, Bratislava, pp 77–83.

Little R J A, Rubin D B (2019) Statistical Analysis with Missing Data, 3rd ed. Wiley, Hoboken. https://doi.org/10.1002/9781119482260

Rubin D B (1987) Multiple Imputation for Nonresponse in Surveys. Wiley, New York. https://doi.org/10.1002/9780470316696

Shah A D, Bartlett J W, Carpenter J, Nicholas O, Hemingway H (2014) Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. American Journal of Epidemiology 179(6):764–774. https://doi.org/10.1093/aje/kwt312

Schwerter J, Gurtskaia K, Romero A, Zeyer-Gliozzo B, Pauly M (2024) Evaluating tree-based imputation methods as an alternative to MICE PMM for drawing inference in empirical studies. arXiv preprint arXiv:2401.09602. https://doi.org/10.48550/arXiv.2401.09602

van Buuren S (2018) Flexible Imputation of Missing Data, 2nd ed. Chapman & Hall/CRC, Boca Raton. https://doi.org/10.1201/9780429492259

Wilson S V, Cebere B, Myatt J, Wilson S (2022) AnotherSamWilson/miceforest: Release for Zenodo DOI. Zenodo. https://doi.org/10.5281/zenodo.7428632

Integrated job shop scheduling and bin packing problem

František Koblasa[1], Miroslav Vavroušek[2]

Abstract. Production scheduling in customised manufacturing represents a critical challenge within manufacturing systems management. Among various operational models, the Job Shop Scheduling Problem (JSSP) is widely utilised due to its ability to handle heterogeneous job routing. While JSSP typically involves single-piece or small-lot production, it frequently incorporates batch-processing stages, such as raw material cutting, chemical treatment, or packing. Existing scheduling models often simplify batching into a one-dimensional capacity constraint; consequently, the multi-dimensional complexities of spatial product placement within a batch are neglected.
This paper addresses the integration of an orthogonal free-space 2D Bin-Packing Problem (2D-BPRF) into the JSSP framework. It proposes an extension of standard models for both domains, enabling the JSSP to formally accommodate the simultaneous processing of multiple jobs within a single batch based on spatial feasibility.
Furthermore, a multi-stage constructive heuristic is proposed to generate feasible schedules by integrating sequencing, batching, and spatial allocation logic. The performance of priority-based dispatching rules is evaluated against an Evolutionary Algorithm to assess trade-offs between computational efficiency and solution quality, as measured by job tardiness. To validate the methodology, new benchmark instances were synthesised from JSSP and 2D-BPRF data, addressing both real-world constraints and theoretical optimisation needs.

Keywords: Job shop scheduling, 2D Bin packing, Evolutionary Algorithm, Heuristics

JEL Classification: C63, M11

1 Introduction

Production scheduling in customised manufacturing environments requires meeting customer-defined due dates while strictly adhering to technological constraints. The Job Shop Scheduling Problem (JSSP) provides a standard framework for managing job sequences and enables the modelling and optimisation of customer temporal requirements. However, JSSP lacks integration of spatial constraints for specific manufacturing stages, typically the material division at the beginning, chemical or heat treatment in the middle and packing operations in the end. Those are 2D batches, constrained by the size of the semi-finished product, the machine workspace, or the package. Additionally, the processing time cannot be modelled linearly, as in organisational lot-sizing, where the duration is typically a linear function of the job count.

This problem can be addressed by integrating bin packing (BPP) as a batching procedure. In real-world operations, batch formation is strictly limited by geometric feasibility and technological precedence, as only operations with completed preceding tasks can be considered for the current batch. This creates a fundamental decision-making dilemma regarding whether to initiate a batch with low spatial utilisation to meet a deadline or to delay processing in anticipation of subsequent operations to improve material yield.

Batching and scheduling have attracted the attention of both scientists and practitioners for a long time (Wahl et al. 2025). Generally, two main types of batches (serial and parallel) are studied. S-batching, in which batch duration is determined by cumulative processing times, is commonly studied in lot-sizing and scheduling contexts to mitigate the impact of sequence-dependent setup times (Rocholl and Mönch 2023). P-batches, where the processing time is determined by the specific batch composition (typically the longest processing time among the included jobs), usually focus on a single machine or a single stage of parallel machines (Fowler and Mönch 2022). Only very few studies (Ma et al. 2026) focus on multistage problems, such as the JSSP or the Flow Shop Scheduling Problem (FSSP). Moreover, JSSP (Knopp et al. 2017) and FSSP (Chung et al. 2017) are mentioned almost exclusively in the context of the semiconductor industry. That is due to its complexity, given by the number

[1,2]Technical university of Liberec, Department of manufacturing systems and automation, Studentská 2, Liberec 1, Czech Republic
[1] frantisek.koblasa@tul.cz
[2] miroslav.vavrousek@tul.cz

of jobs and their stages. However, all of those multistage scheduling problems focus on simplified batching and treat wafer capacity as a one-dimensional, slot-based constraint. Batching in two dimensions is very rare and limited to single (Mao et al. 2024) or parallel machines (Hu et al. 2024). The need to address both 2D-BPP and multistage scheduling is highlighted in a recent article (Xiong et al. 2025), which focuses on the flow shop problem with a 2D-BPP constraint in civil engineering. The problem is decomposed into a master problem for tardiness-based scheduling and a sub-problem for geometric validation, which employs a best-fit heuristic optimised via local search.

In summary, while there is development of 2D-constrained scheduling in flow shop environments, similar applications within Job Shop Scheduling are notably missing from current research.

This paper integrates BPP constraints into the JSSP decision-making logic, using an integrated schedule generation scheme to minimise tardiness while treating the spatial problem as a hard constraint. To address this problem, the paper is organised as follows. Section 2 proposes a modified model of 2D-BPP and JSSP, which serves as the basis for the constructive algorithm in Section 3. A constructive algorithm is then tested on benchmark problems (Section 4), evaluated based on the objective function and computational efficiency (Section 5), and concluded.

2 Model of the Integrated JSSP and 2D-BPRF

Both the JSSP (Lenstra and Rinnooy Kan 1979) and the 2D-BPP (Albers et al. 2025) are categorised as NP-hard. The proposed integrated model addresses a complex hierarchy of independent decision-making subproblems, including job sequencing, batch composition, and the 2D Bin-Packing problem for Rectangular items in Free space (2D-BPRF) with 90-degree orthogonal rotation (Cui et al. 2015).

The JSSP-2D-BPRF considers the processing of n jobs on m machines. In contrast to the classical JSSP, where machine capacity is restricted to a single unit, this study considers a batch-processing environment. In this model, machines are constrained by a defined working area rather than a simple one-dimensional capacity or s-batching. This approach allows the simultaneous processing of multiple operations within a single p-batch (processing time is the maximum of all jobs in the batch), provided that the jobs' physical dimensions satisfy the machine's two-dimensional workspace constraints.

Each scheduled job $J_i \in \{J_1,\ldots,J_n\}$consists of a strict technological sequence (routing) of operations O_{ij}, where i denotes the job index and j denotes the position in the sequence. Each operation O_{ij} is assigned to a specific machine $M_k \in \{M_1,\ldots,M_m\}$ and requires a fixed processing time p_{ij}. Unlike the classical JSSP, machines in this model are characterised by a defined working area with dimensions W_k and H_k, which allows for batch formation.

A batch B_{bk} represents the b-th group of operations processed simultaneously on machine M_k. Within each batch, each operation O_{ij} is associated with a physical component of width w_{ij} and height h_{ij}. The system must determine the start time S_{ij} and completion time C_{ij} for each operation, ensuring that all items in batch B_{kb} fit within the machine dimensions W_k and H_k without overlapping. All n jobs and m machines are available at time 0, and pre-emption is not allowed. Unlike the majority of the research analysed in the introduction, all jobs are assumed to have identical material properties, eliminating the need to group by job families. The following model serves as a mathematical description of the feasibility conditions within the constructive algorithm, rather than a complete optimisation model.

The goal of the problem is to minimise the delay of all jobs (total Tardiness) (1).

$$\min T_{total} = \sum_{i=1}^{n} max\{0, C_{i,n_i} - d_i\} \tag{1}$$

where n_i is the index of the last operation of job J_i, $C_{i,ni}$ is the completion time of J_i, which, in the case of p-batch, is given by the completion of the batch following the principle of (8), and d_i is the customer's due date. Constraint (2) ensures that the operations O_{ij} of a job follow a strict technological sequence and do not overlap in time.

$$S_{ij+1} \geq S_{ij} + p_{ij} \qquad \forall i, j \tag{2}$$

The assignment constraint (3) is governed by a binary variable X_{ijk}, ensuring that each operation O_{ij} is assigned to exactly one compatible machine $M_{k,}$ strictly defined by the technological routing of the JSSP problem.

$$\sum_{k=1}^{M} X_{ijk} = 1 \qquad \forall i, j \tag{3}$$

In the classical JSSP, the machine capacity is limited to processing only one operation at a time (4) (Guo et al. 2006; Xiong et al. 2022).

$$\left(S_{ij} \geq S_{uv} + p_{uv}\right) \vee \left(S_{uv} \geq S_{ij} + p_{ij}\right) \quad \forall i, j, u, v \tag{4}$$

To enable simultaneous processing, we introduce a batch index $b \in B$. The assignment is governed by a binary variable z_{ijbk}. Each operation is uniquely assigned to a specific batch on a specific machine

$$\sum_{k=1}^{m} \sum_{b=1}^{n_k} z_{ijbk} = x_{ijk} \quad \forall i, j, k \tag{5}$$

where n_b is the number of batches.

The proposed model replaces the classical disjunctive constraints by synchronising the start times of all operations assigned to the same batch B_{kb}. Equation (6) ensures that every operation within a batch commences at the exact same moment, which is governed by the latest arrival time among all constituent tasks.

$$S_{ij} = \sum_{k=1}^{m} \sum_{b=1}^{n_b} z_{ijbk}\left(max\left\{\left(max\left\{C_{u,v-1} z_{uvbk}\right\}, C_k\right)\right\}\right) \qquad \forall i, j, b, k \tag{6}$$

In this formulation, the first binary variable z_{ijbk} acts as a selector that assigns the specific batch's start time to operation O_{ij}. Simultaneously, the second variable z_{uvbk} within the maximum function identifies all other operations belonging to the same batch b on machine k, ensuring the batch waits for the completion of the preceding technological step $C_{u,v,-1}$ of the last arriving component and machine availability C_k. The processing time of a batch p_{bk} is determined by (7) the characteristics of parallel batching (p-batching) by the maximum of processing times of batched operations.

$$P_{bk} = max(p_{uv}\, z_{uvbk}) \quad \forall b, k \tag{7}$$

In the case that the machine is batch processing (4) is substituted with (8)

$$S_{b+1,k} \geq S_{bk} + P_{bk}, \forall b \in \{1, \dots, n_b - 1\}, k \;\wedge\; C_{ij} = S_{bk} + P_{bk}, \;\; \forall i, j, b, k: (i, j) \in B_k \tag{8}$$

Once the temporal boundaries of a batch are established, the physical arrangement of operations within the machine's workspace must be defined. To maximise utilisation of the available area, the model 2BPRF allows each component to rotate 90 degrees. This is managed by a binary variable $r_{ij} \in \{0,1\}$, where $r_{ij} = 1$ signifies a rotated state. The changed width $\widetilde{w}_{ij}$ and height $\tilde{h}_{ij}$ of operation O_{ij} are subsequently defined in (9)(10)

$$\widetilde{w}_{ij} = w_{ij}(1 - r_{ij}) + h_{ij} r_{ij} \;\; \forall i, j \tag{9}$$

$$\tilde{h}_{ij} = h_{ij}(1 - r_{ij}) + w_{ij} r_{ij} \;\; \forall i, j \tag{10}$$

To ensure the batch composition is feasible, the cumulative area of all operations assigned to a batch must not exceed the physical surface area of the machine's workspace (11).

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \widetilde{w}_{ij} \tilde{h}_{ij} z_{ijbk} \leq W_k H_k \;\; \forall b, k \tag{11}$$

Each assigned operation must remain within the physical limits of the machine's space (11)(12)(13) and they cannot overlap (14)

$$0 \leq x_{ij} \leq W_k - \widetilde{w}_{ij} \;\; \forall i, j \;\; \in B_k \tag{12}$$

$$0 \leq y_{ij} \leq H_k - \tilde{h}_{ij} \;\; \forall i, j \;\; \in B_k \tag{13}$$

$$\begin{gathered}\left(x_{ij} + \widetilde{w}_{ij} \leq x_{uv}\right) \vee \left(x_{uv} + \widetilde{w}_{uv} \leq x_{ij}\right) \vee \left(y_{ij} + \tilde{h}_{ij} \leq y_{uv}\right) \vee \left(y_{uv} + \tilde{h}_{uv} \leq y_{ij}\right) \\ \forall (i, j), (u, v) \;\; \in B_k, (i, j) \neq (u, v)\end{gathered} \tag{14}$$

where x_{ij} and y_{ij} are the coordinates of O_{ij} operation in the workspace.

3 Constructive algorithm for JSSP-BPRF

The proposed constructive algorithm (CA) is an integrated schedule-generation scheme that simultaneously manages decisions in job sequencing, batch composition, and spatial allocation. CA iteratively constructs a feasible schedule by synchronising temporal decisions with geometric constraints. The procedure follows a dynamic loop that progresses as operations become available according to the technological precedence defined in (2).

Algorithm 1: Integrated JSSP-2D-BPRF

Input: Jobs J_i (d_i, w_i, h_i, p_{ij}) with routing R, Machines M_k (W_k, H_k)
Output: Feasible schedule (S_{ij}, C_{ij}, B_{bk}, x_{ij},y_{ij}, r_{ij}) minimising total tardiness T_{total} (1)

1. $t \leftarrow 0$; $U \leftarrow$ set of all unscheduled operations O_{ij}
2. **while** $U \neq \emptyset$
3. $A \leftarrow$ identify available operations satisfying R by (2)
4. **foreach** $M_k \in M$ do
5. $O_k \leftarrow$ sort operations in A assigned to M_k by **priority set $\boldsymbol{\pi}$**
6. **If** M_k is a batch machine, then
7. $B_{bk} \leftarrow$ initialise empty batch with W_k, H_k area
8. **for** $O_{ij} \in O_k$ do
9. **If** O_{ij} fits $B_{bk,}$ (9)-(14) and does not pass the insertion threshold ($\boldsymbol{B_{bk}}$, $\boldsymbol{\rho}$)
10. *Add* O_{ij} to $B_{bk,}$; update available space
11. **Else** select/add the following batch $B_{b+1,k}$
12. **If** there are no operations for the machine batch available
13. Set starting S_{ij} according (6), processing P_{bk} by (7) and $C_{bk} = S_{bk} + P_{bk}$ ending time
14. **Else** (M_k is standard machine)
15. Set starting $S_{ij} = max(C_{i,j-1}, C_k)$; and ending $C_{ij} = S_{ij} + p_{ij}$ time.
16. $t_{min} \leftarrow min(\{C_{ij} \mid O_{ij} \in A \wedge \mathrm{M}(O_{ij}) = M_k\} \vee \{C_{bk} \mid B_{bk}\}$
17. $M^* \leftarrow$ on which t_{min} is achieved
18. Create conflict set G $\leftarrow \{O_{ij} \in A \mid S_{ij} < t_{min}$ on $M^*\} \cup \{B_{bk} \mid S_{bk} < t_{min}$ on $M^*\}$
19. e* $\leftarrow$ (G, $\boldsymbol{\mu}$) (select candidate from conflict set by rule $\boldsymbol{\mu}$)
20. $U \leftarrow U \setminus (\{O_{ij} \mid O_{ij} = e^*\} \vee \{O_{ij} \mid O_{ij} \in B_{bk} \wedge B_{bk} = e^*\})$

Besides the bin-packing algorithm, which follows previous research (Koblasa et al. 2015), the scheduling process is governed by three major decisions. Sequencing, at batch filling (line 5 rules $\boldsymbol{\pi}$), operation/batch selection (line 19 rules $\boldsymbol{\mu}$), and finally the threshold (line 19 rule $\boldsymbol{\rho}$), which decides if the item fits a batch sparsely (P) and based on the process's suitability to place it in the batch. All of these decisions can be solved using a combination of dispatching rules and a metaheuristic, such as an Evolutionary Algorithm (EA). In the case of sequencing (π, μ), besides metaheuristics, the following rules were considered: First In First Out (FIFO), Shortest and Longest Processing time (SPT, LPT), Most Work Remaining (MWKR), Critical Ratio (CR) and Total Space (TS). Batch selection depends on the threshold defined as an absolute or relative change of the indicator between the first item placed in the batch and the newly placed items. The absolute threshold is used for the arrival time (FIFO) and the remaining process time (MWKR), while the change in CR and processing time (PT) is measured relatively.

4 Benchmark problems generation

The classical problem of introducing new models is to generate suitable theoretical benchmarks to test the ability of the designed algorithms to reach optimal or local-optimal solutions. Two benchmarks presented in this paper are inspired by the famous JSSP and 2D BPP problems. Developed a benchmark generator is taking in account:

- Machine properties: number of machines m; percentage of batch machines $m_{b\%}$; workspace size (W,H).
- Job properties as: number of jobs n; size (w,h); due date d defined by sum of all processing times at job multiplied by tightness factor t_f=<1,10>; number of operations j. Size (w, h) is defined by 4 groups based on machine spatial properties. Thin and long types: Type 1 (0,01-0,5W, 0.66-1 H) and rotated Type 2 (0.66-1 W;0,01-0,5H), big Type 3 (0,5-1W;0,5-1H) and small Type 4 (0,01-0.05W;0,01-0.05H).
- Operation property: minimal and maximal processing time ($p_{ij;min}$;$p_{ij;max}$); probability that the operation will require a batch machine is given by its location within the routing by three positions at the start at the middle, and at the end ($sO_{batch\%}$;$mO_{batch\%}$;$eO_{batch\%}$)

There are generated 2 benchmark problems. JSP-2D-BPRF_10_5 has n=10; m=8; j=5; $m_{b\%}$ = 0.38; W,H = 100; t_f=1,3; Type1-4 evenly distributed; p=<10;20>; ($sO_{batch\%}$;$mO_{batch\%}$;$eO_{batch\%}$) = (0.8;0.6;0.8). JSP-2D-BPRF_50_10 has n=50; m=j=10; $m_{b\%}$ = 0.6; W,H = 100; t_f=3; Type1-4 are evenly distributed; p=<1;20>; ($sO_{batch\%}$;$mO_{batch\%}$;$eO_{batch\%}$) = (0.8;0.25;0.8). Both benchmarks have a default version, as previously described, and a full version in which all machines can process p-batches.

5 Tests and results

All four benchmarks (Koblasa and Vavroušek 2026) are solved by **Algorithm 1** with ***π/ρ/μ*** dispatching rules and by 10 runs of a simple random-key EA (uniform crossover with probability 85%, roulette wheel selection and elitist elimination, population size and number of generations = 100, the experiment was prematurely ended when there was no improvement within 20 generations or population converged) is described in detail at (Koblasa and Vavrousek 2021). Thresholds for ***ρ*** rule are set at 0.8 percentiles of each: the total processing time of each job in MWKR, the theoretical average utilisation of machines for CR, the standard deviation of all operation durations for PT and the simple operation duration in the case of FIFO. FIFO was surprisingly the rule that performed best on average across all settings (see **Table 1** results - bold), and it was used, same as CR, for ***ρ/μ*** (queuing sequence, batch placement) when combined with EA managing the decision at ***π*** (operation/batch selection).

Tardiness - opt. time	**JSP-2D-BPRF_10_5**				**JSP-2D-BPRF_50_10**			
π/ρ/μ	**Default**	t **[min]**	**Full**	t **[min]**	**Default**	t **[min]**	**Full**	t **[min]**
MWKR/P/MWKR	257	0.048	348	0.038	11 893	0.215	3 142	0.272
CR/P/CR	407	0.052	237	0.032	6 596	0.274	2 181	0.276
SPT/P/SPT	**254**	0.051	321	0.034	9 984	0.313	4 422	0.314
LPT/P/LPT	337	0.049	209	0.037	10 511	0.27	1 540	0.251
FIFO/P/SPT	350	0.048	220	0.038	6 332	0.187	1 738	0.237
TS/MWKR/MWKR	367	0.047	368	0.037	14 697	0.288	5 579	0.328
TS/CR/CR	542	0.034	394	0.035	6 911	0.179	3 499	0.445
TS/PT/SPT	426	0.033	565	0.035	9 633	0.311	6 035	0.516
TS/PT/LPT	312	0.03	370	0.034	12 689	0.23	2 741	0.378
TS/FIFO/FIFO	291	0.036	240	0.037	5 836	0.232	1 466	0.273
MWKR/PT/MWKR	409	0.04	459	0.036	13 160	0.315	4 733	0.266
CR/CR/CR	445	0.04	249	0.038	6 703	0.244	3 165	0.302
SPT/PT/SPT	384	0.038	404	0.038	9 971	0.279	6 316	0.364
LPT/PT/LPT	364	0.032	280	0.034	10 379	0.226	1 904	0.219
FIFO/FIFO/FIFO	268	0.039	**203**	0.035	**5 862**	0.172	**1 130**	0.214
EA/CR/CR average	223.4	12.5	139.3	13.7	5 620.5	383.8	1 875.1	557.9
EA/FIFO/FIFO average	**119.9**	9.8	**73**	13.8	**5 065.4**	370.3	**464.8**	779.6

Table 1 Tardiness results of Benchmark tests

It was expected that strategies that include CR and consider the due date and remaining time would perform better than FIFO. Analysing the range of results from dispatching rules, ***π*** had the least impact on results compared to ***ρ/μ***. EA achieved significantly better results, especially when combined with FIFO rather than CR. That can be caused by FIFO outperforming CR in general, but most likely by FIFO following the EA's natural sequence. EA FIFO variant reduced Tardiness on average to approximately 62 % (range 40-79 %) of the CA value for the 10_5 problem, and to 81 % (range 59-92 %) for the 50_10 problem. That comes at a cost as CA gets results in seconds, while EA takes minutes, even hours. However, the simple EA at the time of experimentation was unoptimized.

6 Conclusion

The paper presented a general approach for integrating the common scheduling problem of JSSP with the special case of orthogonal free 2D-BPP to model the spatial constraints of manufacturing devices. CA and EA were tested on four benchmark problems, showing that applying FIFO at both can surprisingly outperform Tardiness-oriented

CR. That is, of course, limited to the benchmarks and algorithms presented. To reach a more robust conclusion, it would be necessary to test it on numerous benchmark problems and to evaluate different combinations of dispatching rules. Most importantly, the development of more advanced EAs for ρ/μ decisions would be desirable, as would enriching integration with other problems from scheduling and packing.

Acknowledgements

This publication was written at the Technical University of Liberec, Faculty of Mechanical Engineering, with the support of the Institutional Endowment for the Long-Term Conceptual Development of Research Institutes, as provided by the Ministry of Education, Youth and Sports of the Czech Republic in the year 2026. The research reported in this paper was supported by institutional support for nonspecific university research.

References

Albers S, Gálvez W, Özdemir Ö B (2025) On the 2D Demand Bin Packing Problem: Hardness and Approximation Algorithms. Procedia Comput Sci 273:301–308. https://doi.org/10.1016/j.procs.2025.10.312

Chung T-P, Sun H, Liao C-J (2017) Two new approaches for a two-stage hybrid flowshop problem with a single batch processing machine under waiting time constraint. Comput Ind Eng 113:859–870. https://doi.org/10.1016/j.cie.2016.11.031

Cui Y-P, Cui Y, Tang T (2015) Sequential heuristic for the two-dimensional bin-packing problem. Eur J Oper Res 240:43–53. https://doi.org/10.1016/j.ejor.2014.06.032

Fowler J W, Mönch L (2022) A survey of scheduling with parallel batch (p-batch) processing. Eur J Oper Res 298:1–24. https://doi.org/10.1016/j.ejor.2021.06.012

Guo Z X, Wong W K, Leung S Y S, et al (2006) Mathematical model and genetic optimization for the job shop scheduling problem in a mixed- and multi-product assembly environment: A case study based on the apparel industry. Comput Ind Eng 50:202–219. https://doi.org/10.1016/j.cie.2006.03.003

Hu K, Che Y, Ng T S, Deng J (2024) Unrelated parallel batch processing machine scheduling with time requirements and two-dimensional packing constraints. Comput Oper Res 162:106474. https://doi.org/10.1016/j.cor.2023.106474

Knopp S, Dauzère-Pérès S, Yugma C (2017) A batch-oblivious approach for Complex Job-Shop scheduling problems. Eur J Oper Res 263:50–61. https://doi.org/10.1016/j.ejor.2017.04.050

Koblasa F, Vavroušek M (2026) JSP-2D-BPRF benchmarks. https://github.com/VavrousekTUL/jsp-2d-bprf-benchmarks

Koblasa F, Vavrousek M (2021) Flexible Job Shop Schedule generation in Evolution Algorithm with Differential Evolution hybridisation. In: Hlavaty R (ed) 39th International Conference on Mathematical Methods in Economics (mme 2021). Czech Univ Life Sciences Prague, Prague 6, pp 249–254

Koblasa F, Vavrousek M, Manlig F (2015) Two-dimensional Bin Packing Problem in batch scheduling. In: Martincik D, Ercingova J, Janecek P (eds) 33rd International Conference on Mathematical Methods in Economics (MME). University of West Bohemia, Plzeň, pp 354–359

Lenstra J K, Rinnooy Kan A H G (1979) Computational Complexity of Discrete Optimization Problems. Ann Discrete Math 4:121–140. https://doi.org/10.1016/S0167-5060(08)70821-5

Ma W, Jiang G, Sun N, et al (2026) Modeling and algorithm for job shop scheduling with batch operations in semiconductor fabs. Comput Oper Res 185:107287. https://doi.org/10.1016/j.cor.2025.107287

Mao Z, Fu E, Huang D, et al (2024) Combinatorial Benders decomposition for single machine scheduling in additive manufacturing with two-dimensional packing constraints. Eur J Oper Res 317:890–905. https://doi.org/10.1016/j.ejor.2024.05.001

Rocholl J, Mönch L (2023) Scheduling Jobs in Flexible Flow Shops with s-batching Machines Using Metaheuristics. In: Metaheuristics. Springer International Publishing, Cham, pp 560–568

Wahl S, Gahm C, Tuma A (2025) Serial- and hierarchical-batch scheduling: a systematic review and future research directions. Int J Prod Res 63:4238–4268. https://doi.org/10.1080/00207543.2024.2432473

Xiong F, Chen X, Liu H, et al (2025) The flow shop batch scheduling problem in a prefabricated component manufacturing system with two-dimensional bin packing constraints. Int J Prod Res 63:7196–7237. https://doi.org/10.1080/00207543.2025.2496970

Xiong H, Shi S, Ren D, Hu J (2022) A survey of job shop scheduling problem: The types and models. Comput Oper Res 142:105731. https://doi.org/10.1016/j.cor.2022.105731

Spatial Dynamic Model on Hierarchical Panel Data: Basic Log-likelihood Properties

Martin Konopásek[1]

Abstract. This paper develops a hierarchical panel data model with spatial dependence in both the dependent variable and the random component. The specification allows spatial dependence to vary across first-level cross-sectional units and incorporates spatial and spatio-temporal dynamics, as well as time-invariant fixed effects. The corresponding log-likelihood function is formulated, and its key analytical properties are derived, including the score vector and Hessian matrix. The asymptotic distribution of the score is then established, together with the Fisher information matrix. These results provide the theoretical basis for ongoing work on bias correction for the associated quasi-maximum likelihood estimators.

Keywords: spatial and temporal dynamics, spatial heterogeneity, hierarchical panel model, maximum likelihood estimation, fixed effects

JEL Classification: C13, C23, C51

1 Introduction

Spatial dynamic panel models provide a flexible framework for analysing outcomes that exhibit temporal persistence, contemporaneous spatial spillovers, and spatio-temporal feedback effects. In the presence of fixed effects, however, the within transformation used to eliminate unobserved time-invariant heterogeneity generates an incidental parameters bias in quasi-maximum likelihood (QML) estimation. For the standard spatial dynamic panel model with spatial autoregressive (SAR) dependence in the dependent variable, Yu et al 2008 derive the leading $O(T^{-1})$ bias term and propose an analytical bias correction.

The objective of this paper is to derive the likelihood-based analytical ingredients required for extending this bias-correction framework to a hierarchical spatial dynamic panel model. The proposed setting contains two cross-sectional levels, for example regions nested within countries, and is therefore suitable for applications in which spatial interaction is naturally defined within higher-level units and may differ across them. Relative to the standard spatial dynamic panel specification, the model allows the SAR coefficient in the dependent variable, the SAR coefficient in the disturbances, and the innovation variance to be higher-level-unit specific.

The main contribution of the paper is the derivation of the concentrated QML log-likelihood after eliminating the fixed effects, together with the associated score expansion, expected Hessian, and score covariance matrix. The normalized score is decomposed into a leading stochastic component and a deterministic time-demeaning component, which identifies the source of the incidental parameters bias. The resulting analytical expressions characterize the information structure of the hierarchical model, including the role of the relative cross-sectional sizes of the higher-level units. These results lead to the asymptotic distribution of the QML estimator and provide the theoretical basis for constructing an analytical bias correction for hierarchical spatial dynamic panels with heterogeneous spatial dependence.

The paper is intentionally focused on the theoretical likelihood properties of the proposed model. In particular, the derived score, information matrix, and leading bias term form the necessary intermediate step for the subsequent development of a feasible bias-corrected QML estimator. Monte Carlo evaluation and empirical implementation are therefore left for follow-up research.

2 The Model Specification

The hierarchical model is specified as follows:

$$\begin{aligned} \boldsymbol{y}_{ct} &= \lambda_c \boldsymbol{W}_c \boldsymbol{y}_{ct} + \left(\tau \boldsymbol{I}_{N_c} + \delta \boldsymbol{W}_c\right) \boldsymbol{y}_{c,t-1} + \boldsymbol{X}_{ct}\boldsymbol{\beta} + \boldsymbol{\eta}_c + \boldsymbol{\varepsilon}_{ct}, \\ \boldsymbol{\varepsilon}_{ct} &= \rho_c \boldsymbol{W}_c \boldsymbol{\varepsilon}_{ct} + \boldsymbol{\nu}_{ct}, \qquad t = 1, \ldots, T, \quad c = 1, \ldots, C. \end{aligned} \tag{1}$$

[1] Prague University of Economics and Business, Department of Econometrics, Winston Churchill Square 1938/4, 130 67 Praha – Žižkov, konm04@vse.cz

Here T denotes the number of time observations, C denotes the number of higher-level cross-sectional units, e.g., countries, and $\boldsymbol{y}_{ct} = [y_{r_c t}, \ldots, y_{(r_c+N_c-1)t}]^\top$ is the $N_c \times 1$ vector of observations on the dependent variable for country c at time t, where N_c is the number of lower-level cross-sectional units, e.g., regions, and $r_c = 1 + \sum_{q=1}^{c-1} N_q$. Further, $\boldsymbol{W}_c$ is a country-specific nonstochastic spatial weights matrix, $\boldsymbol{I}_{N_c}$ is the $N_c \times N_c$ identity matrix, $\boldsymbol{X}_{ct}$ is an $N_c \times k$ matrix of exogenous regressors, $\boldsymbol{\eta}_c$ is an $N_c \times 1$ vector of country-specific time-invariant effects with zero mean and finite variance $\sigma^2_{\eta_c}$, $\boldsymbol{\varepsilon}_{ct} = [\varepsilon_{r_c t}, \ldots, \varepsilon_{(r_c+N_c-1)t}]^\top$ is an $N_c \times 1$ vector of spatially dependent unobserved components, and $\boldsymbol{\nu}_{ct} = [\nu_{r_c t}, \ldots, \nu_{(r_c+N_c-1)t}]^\top$ is an $N_c \times 1$ vector of i.i.d. disturbances with zero mean and finite country-specific variance σ^2_c.

Further, define $\boldsymbol{\xi} = [\tau, \delta, \boldsymbol{\beta}^\top]^\top$ as the $(k+2) \times 1$ vector of response parameters, $\boldsymbol{\varsigma}_c = [\boldsymbol{\xi}^\top, \lambda_c, \rho_c]^\top$, $\boldsymbol{\theta} = [\boldsymbol{\xi}^\top, \boldsymbol{\lambda}^\top, \boldsymbol{\rho}^\top, (\boldsymbol{\sigma}^2)^\top]^\top$, and $\boldsymbol{\theta}_0 = [\boldsymbol{\xi}_0^\top, \boldsymbol{\lambda}_0^\top, \boldsymbol{\rho}_0^\top, (\boldsymbol{\sigma}_0^2)^\top]^\top$ as the $(k+2+3C) \times 1$ parameter vector and its true value, respectively. For notational convenience, let $\boldsymbol{A}_c(\lambda_c) = \boldsymbol{I}_{N_c} - \lambda_c \boldsymbol{W}_c$, $\boldsymbol{B}_c(\rho_c) = \boldsymbol{I}_{N_c} - \rho_c \boldsymbol{W}_c$, $\boldsymbol{A}_c = \boldsymbol{I}_{N_c} - \lambda_{0,c} \boldsymbol{W}_c$, $\boldsymbol{B}_c = \boldsymbol{I}_{N_c} - \rho_{0,c} \boldsymbol{W}_c$, and $\boldsymbol{S}_c = \boldsymbol{A}_c^{-1}(\tau_0 \boldsymbol{I}_{N_c} + \delta_0 \boldsymbol{W}_c)$. Also define the block-diagonal spatial weights matrix $\boldsymbol{W}_N = \mathrm{diag}(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_C)$, and define $\boldsymbol{A}_N$, $\boldsymbol{B}_N$, and $\boldsymbol{S}_N$ analogously. Using this notation, let

$$\boldsymbol{\mu} = \sum_{h=0}^{\infty} \boldsymbol{S}_N^h \boldsymbol{A}_N^{-1} \boldsymbol{\eta}, \qquad \mathcal{X}_t = \sum_{h=0}^{\infty} \boldsymbol{S}_N^h \boldsymbol{A}_N^{-1} \boldsymbol{X}_{t-h}, \qquad \boldsymbol{\nu}_t = \sum_{h=0}^{\infty} \boldsymbol{S}_N^h \boldsymbol{A}_N^{-1} \boldsymbol{B}_N^{-1} \boldsymbol{\nu}_{t-h}. \tag{2}$$

Then (1) can be rewritten compactly as

$$\boldsymbol{y}_t = \mathcal{X}_t \boldsymbol{\beta} + \boldsymbol{\mu} + \boldsymbol{\nu}_t. \tag{3}$$

For the subsequent analysis, following assumptions are imposed:

Assumption 1. For each $c = 1, \ldots, C$, the disturbances $\{\nu_{cit}\}$ are i.i.d. over $i = r_c, \ldots, r_c + N_c - 1$ and $t = 1, \ldots, T$, with zero mean, variance σ^2_c, and $E|\nu_{cit}|^{4+\epsilon} < \infty$ for some $\epsilon > 0$.

Assumption 2. For each $c = 1, \ldots, C$, the spatial weights matrix $\boldsymbol{W}_c$ is obtained by row-normalizing a connectivity matrix $\boldsymbol{C}_c$, i.e., $w_{cij} = c_{cij} / \sum_j c_{cij}$, where

$$c_{cij} = \begin{cases} 0, & i = j, \\ 0, & \text{if the pre-specified spatial connectivity criterion is not met}, \\ 1, & \text{if the pre-specified spatial connectivity criterion is met}. \end{cases} \tag{4}$$

The spatial connectivity criterion may be based, for example, on geographical distance or common-border contiguity between spatial units i and j. The row and column sums of $\boldsymbol{C}_c$ are uniformly bounded in N_c, i.e.,

$$\sum_{i=r_c}^{r_c+N_c-1} |c_{cij}| \le B < \infty \quad \forall j, \qquad \sum_{j=r_c}^{r_c+N_c-1} |c_{cij}| \le B < \infty \quad \forall i. \tag{5}$$

Assumption 3. The elements of $\boldsymbol{X}^*$ have finite first two moments, i.e., $E[(x^*_{p,it})^2] < \infty$ for $i = 1, \ldots, N$, $t = 1, \ldots, T$, and $p = 1, \ldots, k$. Moreover, $\lim_{T\to\infty} (NT)^{-1} \sum_{t=1}^{T} (\boldsymbol{X}_t^*)^\top \boldsymbol{X}_t^*$ exists, is nonsingular, and has rank k.

Assumption 4. The innovations are mean independent of the regressors: $E(\boldsymbol{\nu} \mid \boldsymbol{X}) = \boldsymbol{0}_{NT\times 1}$.

Assumption 5. For each $c = 1, \ldots, C$, the matrices $\boldsymbol{A}_c(\lambda_c)$ and $\boldsymbol{B}_c(\rho_c)$ are invertible for all $\lambda_c, \rho_c \in \Lambda_c$, where Λ_c is compact and $\boldsymbol{W}_c$ satisfies Assumption 2. Moreover, $\boldsymbol{A}_c^{-1}(\lambda_c)$ and $\boldsymbol{B}_c^{-1}(\rho_c)$ are uniformly bounded over $\lambda_c, \rho_c \in \Lambda_c$.

Assumption 6. The matrix series $\sum_{h=0}^{\infty} \mathrm{abs}(\boldsymbol{S}^h)$ is uniformly bounded.

Assumption 7. $N = \sum_{c=1}^{C} N_c$ is a non-decreasing function of T, and $T \to \infty$.

3 The Log-Likelihood Function

The log-likelihood (LL) function is defined using distributional assumptions on the disturbances $\boldsymbol{\nu}$, as if they were normally distributed. Using the Jacobian of transformation; see Anselin 1988, and the first-order condition for $\boldsymbol{\eta}$; see Yu et al 2008, the concentrated LL function, where the fixed effects $\boldsymbol{\eta}$ are concentrated out, is given by

$$\begin{aligned} LL(\boldsymbol{\theta} \mid \boldsymbol{\eta}) = & -\frac{NT}{2} \log(2\pi) + T \sum_{c=1}^{C} \log |\boldsymbol{A}_c(\lambda_c)| + T \sum_{c=1}^{C} \log |\boldsymbol{B}_c(\rho_c)| \\ & - \sum_{c=1}^{C} \frac{N_c T}{2} \log(\sigma^2_c) - \sum_{c=1}^{C} \sum_{t=1}^{T} \frac{1}{2\sigma^2_c} \left(\boldsymbol{\nu}^*_{ct}(\boldsymbol{\varsigma}_c)\right)^\top \boldsymbol{\nu}^*_{ct}(\boldsymbol{\varsigma}_c). \end{aligned} \tag{6}$$

Here, $\boldsymbol{\nu}^*_{ct}(\boldsymbol{\varsigma}_c) = \boldsymbol{\nu}_{ct} - \bar{\boldsymbol{\nu}}_c = \boldsymbol{B}_c(\rho_c)\{\boldsymbol{A}_c(\lambda_c)\boldsymbol{y}^*_{ct} - (\tau \boldsymbol{I}_{N_c} + \delta \boldsymbol{W}_c)\boldsymbol{y}^*_{c,t-1} - \boldsymbol{X}^*_{ct}\boldsymbol{\beta}\}$, $\boldsymbol{y}^*_{ct} = \boldsymbol{y}_{ct} - \bar{\boldsymbol{y}}_c$, $\boldsymbol{y}^*_{c,t-1} = \boldsymbol{y}_{c,t-1} - \bar{\boldsymbol{y}}_{c,-1}$, and $\boldsymbol{X}^*_{ct} = \boldsymbol{X}_{ct} - \bar{\boldsymbol{X}}_c$, where $\bar{\boldsymbol{\nu}}_c = T^{-1}\sum_{t=1}^T \boldsymbol{\nu}_{ct}$, $\bar{\boldsymbol{y}}_c = T^{-1}\sum_{t=1}^T \boldsymbol{y}_{ct}$, $\bar{\boldsymbol{y}}_{c,-1} = T^{-1}\sum_{t=1}^T \boldsymbol{y}_{c,t-1}$, and $\bar{\boldsymbol{X}}_c = T^{-1}\sum_{t=1}^T \boldsymbol{X}_{ct}$. Maximizing (6) gives the QML estimator of $\boldsymbol{\theta}_0$ as $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} LL(\boldsymbol{\theta} \mid \boldsymbol{\eta})$.

Gradient. Following Yu et al 2008, define

$$\begin{aligned}\tilde{\boldsymbol{Z}}_t &= [\boldsymbol{y}^*_{t-1},\ \boldsymbol{W}_N\boldsymbol{y}^*_{t-1},\ \boldsymbol{X}^*_t] \\ &= \boldsymbol{Z}^*_t - (\bar{\boldsymbol{\nu}}_{T-1},\ \boldsymbol{W}_N\bar{\boldsymbol{\nu}}_{T-1},\ \boldsymbol{0}_{N\times k})\end{aligned} \tag{7}$$

where the right-hand-side variables are split into endogenous and weakly exogenous components, with the latter being uncorrelated with $\boldsymbol{\nu}_t$. The weakly exogenous component of $\boldsymbol{Z}_t$ is

$$\boldsymbol{Z}^*_t = (\tilde{\boldsymbol{X}}_{t-1}\boldsymbol{\beta} + \boldsymbol{\nu}_{t-1},\ \boldsymbol{W}_N\tilde{\boldsymbol{X}}_{t-1}\boldsymbol{\beta} + \boldsymbol{W}_N\boldsymbol{\nu}_{t-1},\ \boldsymbol{X}^*_t). \tag{8}$$

Let $\boldsymbol{G}_c = \boldsymbol{W}_c\boldsymbol{A}_c^{-1}$, $\boldsymbol{H}_c = \boldsymbol{W}_c\boldsymbol{B}_c^{-1}$, $\ddot{\boldsymbol{W}}_c = \boldsymbol{B}_c\boldsymbol{W}_c\boldsymbol{B}_c^{-1}$, $\ddot{\boldsymbol{G}}_c = \ddot{\boldsymbol{W}}_c(\boldsymbol{I}_{N_c} - \lambda_{0,c}\ddot{\boldsymbol{W}}_c)^{-1}$, $\ddot{\boldsymbol{Z}}_t = \boldsymbol{B}_N\widetilde{\boldsymbol{Z}}_t$, $\ddot{\boldsymbol{Z}}^*_t = \boldsymbol{B}_N\boldsymbol{Z}^*_t$, and $\ddot{\boldsymbol{X}}_t = \boldsymbol{B}_N\widetilde{\boldsymbol{Z}}_t$. Then

$$\frac{1}{\sqrt{NT}}\frac{\partial LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} = \frac{1}{\sqrt{NT}}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} - \Delta_{\theta_0}, \tag{9}$$

where

$$\frac{1}{\sqrt{NT}}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} = \begin{bmatrix} \sum_{c=1}^C\sum_{t=1}^T \frac{1}{\sigma^2_{0,c}\sqrt{NT}}\left(\ddot{\boldsymbol{Z}}^*_{ct}\right)^\intercal \boldsymbol{\nu}_{ct} \\ \frac{1}{\sigma^2_1\sqrt{NT}}\sum_{t=1}^T\Big(\left(\ddot{\boldsymbol{G}}_1\ddot{\boldsymbol{Z}}^*_{1t}\xi_0\right)^\intercal\boldsymbol{\nu}_{1t} + \\ +\boldsymbol{\nu}^\intercal_{1t}\ddot{\boldsymbol{G}}\boldsymbol{\nu}_{1t} - \sigma^2_{0,1}\mathrm{tr}(\ddot{\boldsymbol{G}}_1)\Big) \\ \vdots \\ \frac{1}{\sigma^2_{0,C}\sqrt{NT}}\sum_{t=1}^T\Big(\left(\ddot{\boldsymbol{G}}_C\ddot{\boldsymbol{Z}}^*_{Ct}\xi_0\right)^\intercal\boldsymbol{\nu}_{Ct} + \\ +\boldsymbol{\nu}^\intercal_{Ct}\ddot{\boldsymbol{G}}\boldsymbol{\nu}_{Ct} - \sigma^2_{0,C}\mathrm{tr}(\ddot{\boldsymbol{G}}_C)\Big) \\ \frac{1}{\sigma^2_{0,1}\sqrt{NT}}\sum_{t=1}^T\left(\boldsymbol{\nu}^\intercal_{1t}\boldsymbol{H}_1\boldsymbol{\nu}^*_{1t} - \sigma^2_{0,1}\mathrm{tr}(\boldsymbol{H}_1)\right) \\ \vdots \\ \frac{1}{\sigma^2_{0,C}\sqrt{NT}}\sum_{t=1}^T\left(\boldsymbol{\nu}^\intercal_{Ct}\boldsymbol{H}_C\boldsymbol{\nu}_{Ct} - \sigma^2_{0,C}\mathrm{tr}(\boldsymbol{H}_C)\right) \\ \frac{1}{2\sigma^4_{0,1}\sqrt{NT}}\sum_{t=1}^T\left(\boldsymbol{\nu}^\intercal_{1t}\boldsymbol{\nu}_{1t} - N_1\sigma^2_{0,1}\right) \\ \vdots \\ \frac{1}{2\sigma^4_{0,C}\sqrt{NT}}\sum_{t=1}^T\left(\boldsymbol{\nu}^\intercal_{Ct}\boldsymbol{\nu}_{Ct} - N_C\sigma^2_{0,C}\right) \end{bmatrix},\ \Delta_{\theta_0} = \begin{bmatrix} \sum_{c=1}^C\frac{\sqrt{T}}{\sigma^2_{0,c}\sqrt{N}}\left(\boldsymbol{B}_c\left(\bar{\boldsymbol{\nu}}_{c,T-1},\ \boldsymbol{W}_c\bar{\boldsymbol{\nu}}_{c,T-1},\ \boldsymbol{0}_{N_c\times k}\right)\right)^\intercal\bar{\boldsymbol{\nu}}_{c,T} \\ \frac{\sqrt{T}}{\sigma^2_{0,1}\sqrt{N}}\Big(\left(\boldsymbol{B}_1\boldsymbol{G}_1\left(\bar{\boldsymbol{\nu}}_{1,T-1},\ \boldsymbol{W}_1\bar{\boldsymbol{\nu}}_{1,T-1},\ \boldsymbol{0}_{N\times k}\right)\xi\right)^\intercal\bar{\boldsymbol{\nu}}_{1,T} + \\ +\bar{\boldsymbol{\nu}}^\intercal_{1,T}(\boldsymbol{B}_1^{-1})^\intercal\boldsymbol{G}_1^\intercal\boldsymbol{B}_1^\intercal\bar{\boldsymbol{\nu}}_{1,T}\Big) \\ \vdots \\ \frac{\sqrt{T}}{\sigma^2_{0,C}\sqrt{N}}\Big(\left(\boldsymbol{B}_C\boldsymbol{G}_C\left(\bar{\boldsymbol{\nu}}_{C,T-1},\ \boldsymbol{W}_C\bar{\boldsymbol{\nu}}_{C,T-1},\ \boldsymbol{0}_{N\times k}\right)\xi\right)^\intercal\bar{\boldsymbol{\nu}}_{C,T} + \\ +\bar{\boldsymbol{\nu}}^\intercal_{C,T}(\boldsymbol{B}_C^{-1})^\intercal\boldsymbol{G}_C^\intercal\boldsymbol{B}_C^\intercal\bar{\boldsymbol{\nu}}_{C,T}\Big) \\ \frac{\sqrt{T}}{\sigma^2_{0,1}\sqrt{N}}\bar{\boldsymbol{\nu}}^\intercal_{1,T}\boldsymbol{H}_1\bar{\boldsymbol{\nu}}_{1,T} \\ \vdots \\ \frac{\sqrt{T}}{\sigma^2_{0,C}\sqrt{N}}\bar{\boldsymbol{\nu}}^\intercal_{C,T}\boldsymbol{H}_C\bar{\boldsymbol{\nu}}_{C,T} \\ \frac{\sqrt{T}}{2\sigma^4_{0,1}\sqrt{N}}\bar{\boldsymbol{\nu}}^\intercal_{1,T}\bar{\boldsymbol{\nu}}_{1,T} \\ \vdots \\ \frac{\sqrt{T}}{2\sigma^4_{0,C}\sqrt{N}}\bar{\boldsymbol{\nu}}^\intercal_{C,T}\bar{\boldsymbol{\nu}}_{C,T} \end{bmatrix}.$$

From $\Delta_{\theta_0} = \left(\Delta_{\theta_0} - E(\Delta_{\theta_0})\right) + E(\Delta_{\theta_0})$ we can use Lemma 11 from Yu et al, 2008 to obtain $\Delta_{\theta_0} = \sqrt{\frac{N}{T}}\varphi + O\left(\sqrt{\frac{N}{T^3}}\right) + O_p\left(\frac{1}{\sqrt{T}}\right)$ where $\varphi = O(1)$ is defined as

$$\varphi = \begin{bmatrix} \frac{1}{N}\mathrm{tr}\left(\left(\sum_{h=0}^\infty \boldsymbol{S}^h_N\right)\boldsymbol{A}^{-1}_N\right) \\ \frac{1}{N}\mathrm{tr}\left(\boldsymbol{W}_N\left(\sum_{h=0}^\infty \boldsymbol{S}^h_N\right)\boldsymbol{A}^{-1}_N\right) \\ \boldsymbol{0}_{k\times 1} \\ \frac{1}{N}\tau_0\mathrm{tr}\left(\boldsymbol{G}_1\left(\sum_{h=0}^\infty \boldsymbol{S}^h_1\right)\boldsymbol{A}^{-1}_1\right) + \frac{1}{N}\delta_0\mathrm{tr}\left(\boldsymbol{G}_1\boldsymbol{W}_1\left(\sum_{h=0}^\infty \boldsymbol{S}^h_1\right)\boldsymbol{A}^{-1}_1\right) + \frac{1}{N}\mathrm{tr}\left(\boldsymbol{G}^\intercal_1\right) \\ \vdots \\ \frac{1}{N}\tau_0\mathrm{tr}\left(\boldsymbol{G}_C\left(\sum_{h=0}^\infty \boldsymbol{S}^h_C\right)\boldsymbol{A}^{-1}_C\right) + \frac{1}{N}\delta_0\mathrm{tr}\left(\boldsymbol{G}_C\boldsymbol{W}_C\left(\sum_{h=0}^\infty \boldsymbol{S}^h_C\right)\boldsymbol{A}^{-1}_C\right) + \frac{1}{N}\mathrm{tr}\left(\boldsymbol{G}^\intercal_C\right) \\ \frac{1}{N}\mathrm{tr}(\boldsymbol{H}_1) \\ \vdots \\ \frac{1}{N}\mathrm{tr}(\boldsymbol{H}_C) \\ \frac{N_1}{2\sigma^2_1 N} \\ \vdots \\ \frac{N_C}{2\sigma^2_C N} \end{bmatrix}. \tag{10}$$

Under Assumptions 1–6, the first two elements of φ are $O(1)$, while the remaining elements are no larger than $O(N_r/N)$, where N_r denotes the country-specific cross-sectional dimension with the highest convergence rate.

Hessian and the information matrix. Define the matrix $\boldsymbol{\Sigma}_{\theta_0} = -E\left(\frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\intercal}\right)$, with blocks denoted by $\mathcal{H}_{\theta_i\theta_j} = -E\left(\frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_i\,\partial\boldsymbol{\theta}_j^\intercal}\right)$, and define $\boldsymbol{\Gamma}_{\theta_0} = E\left(\frac{1}{NT}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^\intercal}\right)$, with blocks $\mathcal{Q}_{\theta_i\theta_j} = E\left(\frac{1}{NT}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_i}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_j^\intercal}\right)$, where $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \subset \boldsymbol{\theta}$. First, consider the block associated with the common response parameter vector $\boldsymbol{\xi}$:

$$\begin{aligned}
\mathcal{H}_{\xi\xi} &= -E\left(\frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^\intercal}\right) = E\left(\sum_{c=1}^{C}\sum_{t=1}^{T}\frac{1}{\sigma_{0,c}^2 NT}\left(\boldsymbol{B}_c\tilde{\boldsymbol{Z}}_{ct}\right)^\intercal\boldsymbol{B}_c\tilde{\boldsymbol{Z}}_{ct}\right) = \frac{1}{\sigma_{0,c}^2 NT}\sum_{c=1}^{C}\sum_{t=1}^{T}E\left(\left(\ddot{\boldsymbol{Z}}_{ct}\right)^\intercal\ddot{\boldsymbol{Z}}_{ct}\right) = O(1)\\
&= \frac{1}{\sigma_{0,c}^2 NT}\sum_{c=1}^{C}\sum_{t=1}^{T}E\left(\left(\ddot{\boldsymbol{Z}}_{ct}^*\right)^\intercal\ddot{\boldsymbol{Z}}_{ct}^*\right) - \frac{1}{\sigma_{0,c}^2 NT}E\left(\sum_{t=1}^{T}(\boldsymbol{B}_N\bar{\boldsymbol{v}}_{T-1},\ \boldsymbol{B}_N\boldsymbol{W}_N\bar{\boldsymbol{v}}_{T-1},\ \boldsymbol{0}_{N\times k})^\intercal(\boldsymbol{B}_N\bar{\boldsymbol{v}}_{T-1},\ \boldsymbol{B}_N\boldsymbol{W}_N\bar{\boldsymbol{v}}_{T-1},\ \boldsymbol{0}_{N\times k})\right)\\
&= \mathcal{Q}_{\xi\xi} - O\left(\frac{1}{T}\right).
\end{aligned}$$

This follows from the identity

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\boldsymbol{v}_{t-1}^*,\ \boldsymbol{W}_N\boldsymbol{v}_{t-1}^*,\ \boldsymbol{0}_{N\times k}\right)^\intercal\boldsymbol{B}_N^\intercal\boldsymbol{B}_N\left(\boldsymbol{v}_{t-1}^*,\ \boldsymbol{W}_N\boldsymbol{v}_{t-1}^*,\ \boldsymbol{0}_{N\times k}\right) = \sum_{t=1}^{T}\boldsymbol{v}_{t-1}^\intercal\boldsymbol{B}_N^\intercal\boldsymbol{B}_N\boldsymbol{v}_{t-1} + 2\sum_{t=1}^{T}\boldsymbol{v}_{t-1}^\intercal\boldsymbol{B}_N^\intercal\boldsymbol{B}_N\boldsymbol{W}_N\boldsymbol{v}_{t-1} + \sum_{t=1}^{T}\boldsymbol{v}_{t-1}^\intercal\boldsymbol{W}_N^\intercal\boldsymbol{B}_N^\intercal\boldsymbol{B}_N\boldsymbol{W}_N\boldsymbol{v}_{t-1} -\\
&\quad - T\bar{\boldsymbol{v}}_{T-1}^\intercal\boldsymbol{B}_N^\intercal\boldsymbol{B}_N\bar{\boldsymbol{v}}_{T-1} - 2T\bar{\boldsymbol{v}}_{T-1}^\intercal\boldsymbol{B}_N^\intercal\boldsymbol{B}_N\boldsymbol{W}_N\bar{\boldsymbol{v}}_{T-1} - T\bar{\boldsymbol{v}}_{T-1}^\intercal\boldsymbol{W}_N^\intercal\boldsymbol{B}_N^\intercal\boldsymbol{B}_N\boldsymbol{W}_N\bar{\boldsymbol{v}}_{T-1}.
\end{aligned}$$

By Lemma 7 in Yu et al, 2008, the expected value of the terms involving $\bar{\boldsymbol{v}}_{T-1}$ is of order $O(N)$, whereas the expected value of the remaining terms is of order $O(NT)$. Next, consider the blocks associated with $\boldsymbol{\xi}$ and λ_c, for $c = 1, \ldots, C$:

$$\begin{aligned}
\mathcal{H}_{\xi\lambda_c} &= E\left(\frac{1}{\sigma_{0,c}^2 NT}\sum_{t=1}^{T}\left(\boldsymbol{B}_c\boldsymbol{W}_c\boldsymbol{y}_{ct}^*\right)^\intercal\boldsymbol{B}_c\tilde{\boldsymbol{Z}}_{ct}\right) = \frac{1}{\sigma_{0,c}^2 NT}E\left(\sum_{t=1}^{T}\left(\ddot{\boldsymbol{G}}_c\ddot{\boldsymbol{Z}}_{ct}\boldsymbol{\xi}_0\right)^\intercal\ddot{\boldsymbol{Z}}_{ct}\right) -\\
&\quad - \frac{1}{NT}\left[\operatorname{tr}\left(\boldsymbol{W}_c^\intercal\boldsymbol{B}_c^\intercal\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\boldsymbol{A}_c^{-1}\right),\ \operatorname{tr}\left(\boldsymbol{W}_c^\intercal\boldsymbol{B}_c^\intercal\boldsymbol{G}_c\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\right),\ \boldsymbol{0}_{1\times k}\right] - O\left(\frac{N_c}{NT^2}\right) = O\left(\frac{N_c}{N}\right) - O\left(\frac{N_c}{NT}\right) - O\left(\frac{N_c}{NT^2}\right)\\
&= \frac{1}{\sigma_{0,c}^2 NT}E\left(\sum_{t=1}^{T}\left(\ddot{\boldsymbol{G}}_c\ddot{\boldsymbol{Z}}_{ct}^*\boldsymbol{\xi}_0\right)^\intercal\ddot{\boldsymbol{Z}}_{ct}^*\right) - O\left(\frac{N_c}{NT}\right) = \mathcal{Q}_{\xi\lambda_c} - O\left(\frac{N_c}{NT}\right), \qquad \text{for } c = 1, \ldots, C,
\end{aligned}$$

where Lemmas 7 and 11 of Yu et al, 2008 are used. Applying the same arguments, the remaining blocks of $\boldsymbol{\Sigma}_{\theta_0}$ are obtained as follows:

$$\begin{aligned}
\mathcal{H}_{\xi\rho_c} &= -\frac{1}{NT}\left[\operatorname{tr}\left(\left(\boldsymbol{W}_c + \boldsymbol{H}_c^\intercal\right)\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\boldsymbol{A}_c^{-1}\right),\ \operatorname{tr}\left(\left(\boldsymbol{W}_c^2 + \boldsymbol{H}_c^\intercal\boldsymbol{W}_c\right)\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\boldsymbol{A}_c^{-1}\right),\ \boldsymbol{0}_{1\times k}\right] - O\left(\frac{N_c}{NT^2}\right)\\
&= -O\left(\frac{N_c}{NT}\right) - O\left(\frac{N_c}{NT^2}\right) = \mathbf{Q}_{\xi\rho_c} - O\left(\frac{N_c}{NT}\right), \qquad \text{for } c = 1, \ldots, C,
\end{aligned}$$

$$\mathcal{H}_{\xi\sigma_c^2} = -\frac{1}{NT}\left[\operatorname{tr}\left(\sum_{h=0}^{\infty}\boldsymbol{S}^h\boldsymbol{A}^{-1}\right),\ \operatorname{tr}\left(\boldsymbol{W}\sum_{h=0}^{\infty}\boldsymbol{S}^h\boldsymbol{A}^{-1}\right),\ \boldsymbol{0}_{1\times k}\right] - O\left(\frac{N_c}{NT^2}\right) = -O\left(\frac{N_c}{NT}\right) - O\left(\frac{N_c}{NT^2}\right) = \mathcal{Q}_{\xi\sigma_c^2} - O\left(\frac{N_c}{NT}\right), \qquad \text{for } c = 1, \ldots, C,$$

$$\mathcal{H}_{\lambda_c\lambda_q} = \begin{cases}
\dfrac{1}{\sigma_{0,c}^2 NT}E\left(\displaystyle\sum_{t=1}^{T}\left(\ddot{\boldsymbol{G}}_c\ddot{\boldsymbol{Z}}_{ct}\boldsymbol{\xi}_0\right)^\intercal\ddot{\boldsymbol{G}}_c\ddot{\boldsymbol{Z}}_{ct}\boldsymbol{\xi}_0\right) + \dfrac{T-1}{NT}\operatorname{tr}\left(\ddot{\boldsymbol{G}}_c^\intercal\ddot{\boldsymbol{G}}_c\right) - \dfrac{2}{NT}\operatorname{tr}\left(\tau_0\ddot{\boldsymbol{G}}_c{}^\intercal\ddot{\boldsymbol{G}}_c\displaystyle\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\boldsymbol{A}_c^{-1} + \delta_0\ddot{\boldsymbol{G}}_c{}^\intercal\ddot{\boldsymbol{G}}_c\boldsymbol{W}_c\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\boldsymbol{A}_c^{-1}\right) -\\
\quad -O\left(\dfrac{N_c}{NT^2}\right) + \dfrac{1}{N}\operatorname{tr}\left(\boldsymbol{G}_c^2\right) = O\left(\dfrac{N_c}{N}\right) - O\left(\dfrac{N_c}{NT}\right) - O\left(\dfrac{N_c}{NT^2}\right) = \mathcal{Q}_{\lambda_c\lambda_q} - \dfrac{\mu_{0,c}^4 - 3\sigma_{0,c}^4}{N\sigma_{0,c}^4}\displaystyle\sum_{i=1}^{N_c}\ddot{g}_{c,ii}^2 - O\left(\dfrac{N_c}{NT}\right), \qquad \text{for } c = q,\\
0 = \mathcal{Q}_{\lambda_c\lambda_q}, \qquad \text{for } c \neq q,
\end{cases}$$

$$\mathcal{H}_{\lambda_c\rho_q} = \begin{cases}
\dfrac{T-1}{NT}\operatorname{tr}\left(\boldsymbol{H}_c^\intercal\ddot{\boldsymbol{G}}_c + \boldsymbol{H}_c\ddot{\boldsymbol{G}}_c\right) - \dfrac{1}{NT}\operatorname{tr}\left(\tau_0\left(\boldsymbol{H}_c^\intercal\ddot{\boldsymbol{G}}_c\boldsymbol{B}_c + \boldsymbol{W}_c^2\boldsymbol{A}_c^{-1}\right)\displaystyle\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\boldsymbol{B}_c^{-1}\boldsymbol{A}_c^{-1} + \delta_0\left(\boldsymbol{H}_c^\intercal\ddot{\boldsymbol{G}}_c\boldsymbol{B}_c\boldsymbol{W}_c + \boldsymbol{W}_c^2\boldsymbol{A}_c^{-1}\boldsymbol{W}_c\right)\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\boldsymbol{B}_c^{-1}\boldsymbol{A}_c^{-1}\right) -\\
\quad -O\left(\dfrac{N_c}{NT^2}\right) = O\left(\dfrac{N_c}{N}\right) - O\left(\dfrac{N_c}{NT}\right) - O\left(\dfrac{N_c}{NT^2}\right) = \mathcal{Q}_{\lambda_c\rho_q} - \dfrac{\mu_{0,c}^4 - 3\sigma_{0,c}^4}{N\sigma_{0,c}^4}\displaystyle\sum_{i=1}^{N_c}\ddot{g}_{c,ii}h_{c,ii} - O\left(\dfrac{N_c}{NT}\right), \qquad \text{for } c = q,\\
0 = \mathcal{Q}_{\lambda_c\rho_q}, \qquad \text{for } c \neq q,
\end{cases}$$

$$\mathcal{H}_{\lambda_c\sigma_q^2} = \begin{cases}
\dfrac{T-1}{\sigma_{0,c}^2 NT}\operatorname{tr}\left(\ddot{\boldsymbol{G}}_c\right) - \dfrac{1}{\sigma_{0,c}^2 NT}\operatorname{tr}\left(\left(\tau_0\boldsymbol{B}_c\boldsymbol{W}_c + \delta_0\boldsymbol{B}_c\boldsymbol{W}_c^2\right)\displaystyle\sum_{h=0}^{\infty}\boldsymbol{S}_c^h\boldsymbol{B}_c^{-1}\boldsymbol{A}_c^{-1}\right) - O\left(\dfrac{N_c}{NT^2}\right) = O\left(\dfrac{N_c}{N}\right) - O\left(\dfrac{N_c}{NT}\right) - O\left(\dfrac{N_c}{NT^2}\right)\\
\quad = \mathcal{Q}_{\lambda_c\sigma_q^2} - \dfrac{\mu_{0,c}^4 - 3\sigma_{0,c}^4}{\sigma_{0,c}^4}\dfrac{\operatorname{tr}\left(\ddot{\boldsymbol{G}}_c\right)}{2N\sigma_{0,c}^2} - O\left(\dfrac{N_c}{NT}\right), \qquad \text{for } c = q,\\
0 = \mathcal{Q}_{\lambda_c\sigma_q^2}, \qquad \text{for } c \neq q,
\end{cases}$$

$$\mathcal{H}_{\rho_c\rho_q} = \begin{cases} \frac{1}{N}\mathrm{tr}\left(\boldsymbol{H}_c^\intercal \boldsymbol{H}_c + \boldsymbol{H}_c^2\right) - \frac{1}{NT}\mathrm{tr}\left(\boldsymbol{H}_c^\intercal \boldsymbol{H}_c\right) = O\left(\frac{N_c}{N}\right) - O\left(\frac{N_c}{NT}\right) = Q_{\rho_c\rho_q} - \frac{\mu_{0,c}^4 - 3\sigma_{0,c}^4}{N\sigma_{0,c}^4}\sum_{i=1}^{N_c} h_{c,ii}^2 - O\left(\frac{N_c}{NT}\right), & \text{for } c = q, \\ 0 = Q_{\rho_c\rho_q}, \quad \text{for } c \neq q, \end{cases}$$

$$\mathcal{H}_{\rho_c\sigma_q^2} = \begin{cases} \frac{T-1}{\sigma_{0,c}^2 NT}\mathrm{tr}\left(\boldsymbol{H}_c\right) = O\left(\frac{N_c}{N}\right) - O\left(\frac{N_c}{NT}\right) = Q_{\rho_c\sigma_q^2} - \frac{\mu_{0,c}^4 - 3\sigma_{0,c}^4}{\sigma_{0,c}^4}\frac{\mathrm{tr}\left(\boldsymbol{H}_c\right)}{2N\sigma_{0,c}^2} - O\left(\frac{N_c}{NT}\right), & \text{for } c = q, \\ 0 = Q_{\rho_c\sigma_q^2}, \quad \text{for } c \neq q, \end{cases}$$

$$\mathcal{H}_{\sigma_c^2\sigma_q^2} = \begin{cases} \frac{(T-1)N_c}{2\sigma_{0,c}^4 NT} = O\left(\frac{N_c}{N}\right) - O\left(\frac{N_c}{NT}\right) = Q_{\sigma_c^2\sigma_q^2} - \frac{\mu_{0,c}^4 - 3\sigma_{0,c}^4}{\sigma_{0,c}^4}\frac{N_c}{4N\sigma_{0,c}^4} - O\left(\frac{N_c}{NT}\right), & \text{for } c = q, \\ 0 = Q_{\sigma_c^2\sigma_q^2}, \quad \text{for } c \neq q. \end{cases}$$

The matrix $\boldsymbol{\Sigma}_{\theta_0}$ is obtained by stacking the individual blocks as follows:

$$\boldsymbol{\Sigma}_{\theta_0} = \begin{bmatrix} \mathcal{H}_{\xi\xi} & * & * & * \\ \mathcal{H}_{\xi\lambda} & \mathcal{H}_{\lambda\lambda} & * & * \\ \mathcal{H}_{\xi\rho} & \mathcal{H}_{\lambda\rho} & \mathcal{H}_{\rho\rho} & * \\ \mathcal{H}_{\xi\sigma^2} & \mathcal{H}_{\lambda\sigma^2} & \mathcal{H}_{\rho\sigma^2} & \mathcal{H}_{\sigma^2\sigma^2} \end{bmatrix} = \boldsymbol{\Gamma}_{\theta_0} - \boldsymbol{\Omega}_{\theta_0} - O\left(\frac{1}{T}\right), \tag{11}$$

where the lower $3C \times 3C$ diagonal block is composed of $C \times C$ diagonal matrices. The elements of $\boldsymbol{\Omega}_{\theta_0}$ associated with this block consist of products of $O\left(\frac{N_c}{N}\right)$ terms and the factor $\frac{\mu_{0,c}^4 - 3\sigma_{0,c}^4}{\sigma_{0,c}^4}$, while the blocks involving $\boldsymbol{\xi}$ are zero. Thus, when the country-specific disturbances $\{v_{cit}\}$ are nonnormal, the elements of $\boldsymbol{\Omega}_{\theta_0}$ are $O(1)$ if N_c grows at the same rate as N, and $O\left(\frac{1}{N}\right)$ if N_c is fixed. Under normality of $\{v_{cit}\}$, the corresponding blocks of $\boldsymbol{\Omega}_{\theta_0}$ are zero. Moreover, since N_c cannot grow faster than N, it follows that

$$\lim_{T\to\infty} \boldsymbol{\Sigma}_{\theta_0} = \lim_{T\to\infty} \boldsymbol{\Gamma}_{\theta_0} - \boldsymbol{\Omega}_{\theta_0}. \tag{12}$$

Also, define $\boldsymbol{\theta}_c = [\lambda_c,\ \rho_c,\ \sigma_c^2]^\intercal$ and $\boldsymbol{\theta}_q = [\lambda_q,\ \rho_q,\ \sigma_q^2]^\intercal$ for $c, q = 1, \ldots, C$. The elements of $\boldsymbol{\Sigma}_{\theta_0}$ can then be restacked so that the upper diagonal block is $\mathcal{H}_{\xi\xi} = O(1)$, while the lower diagonal block is block diagonal, with 3×3 diagonal blocks $\mathcal{H}_{\theta_c\theta_c} = O\left(\frac{N_c}{N}\right)$ and off-diagonal blocks $\mathcal{H}_{\theta_c\theta_q} = \mathbf{0}_{3\times 3}$ for $c \neq q$. The first row of off-diagonal blocks consists of the $k \times 3$ matrices $\mathcal{H}_{\xi\theta_c} = \left[O\left(\frac{N_c}{N}\right), O\left(\frac{N_c}{NT}\right), O\left(\frac{N_c}{NT}\right)\right]$, and the first column contains their transposes. This structure implies that $\boldsymbol{\Sigma}_{\theta_0}$ becomes singular when $\frac{N_s}{N} \to 0$ as $N, N_s \to \infty$, where N_s denotes the country-specific cross-sectional dimension with the slowest rate of convergence. The following simplifying assumption is therefore imposed.[1]

Assumption 8. $\frac{N}{N_c} \to k_c \leq b < \infty, \forall\ c \in \{1, \ldots, C\}$ and C is fixed.

Under Assumption 8, $O\left(\frac{N_s}{N}\right) = O(1)$. Moreover, since $\lim_{T\to\infty} \mathcal{H}_{\xi\rho} = \mathbf{0}_{(k+2)\times C}$ and $\lim_{T\to\infty} \mathcal{H}_{\xi\sigma^2} = \mathbf{0}_{(k+2)\times C}$, it follows from the proof of Theorem 2 in Lee and Yu 2010 that $\lim_{T\to\infty} \boldsymbol{\Sigma}_{\theta_0}$ is nonsingular provided that the following assumption holds.

Assumption 9. $\lim_{T\to\infty} \mathcal{K}_{\xi\lambda}$ is nonsingular.

Here, $\mathcal{K}_{\xi\lambda} = \frac{1}{NT}E\left(\sum_{t=1}^T \left[\ddot{\boldsymbol{Z}}_t,\ \boldsymbol{E}_1\ddot{\boldsymbol{G}}\ddot{\boldsymbol{Z}}_t\boldsymbol{\xi}_0, \ldots,\ \boldsymbol{E}_C\ddot{\boldsymbol{G}}\ddot{\boldsymbol{Z}}_t\boldsymbol{\xi}_0\right]^\intercal \left[\ddot{\boldsymbol{Z}}_t,\ \boldsymbol{E}_1\ddot{\boldsymbol{G}}\ddot{\boldsymbol{Z}}_t\boldsymbol{\xi}_0, \ldots,\ \boldsymbol{E}_C\ddot{\boldsymbol{G}}\ddot{\boldsymbol{Z}}_t\boldsymbol{\xi}_0\right]\right)$, where $\boldsymbol{E}_c$ is an $N \times N$ matrix whose diagonal elements corresponding to the cth country are equal to one, while all remaining elements are zero. From the block structure of $\boldsymbol{\Sigma}_{\theta_0}$, using the inverse formula for a block matrix, it follows that the blocks of $\boldsymbol{\Sigma}_{\theta_0}^{-1}$ corresponding to the individual parameters are at most of order $O\left(\frac{N}{N_s}\right)$. More specifically, letting the block of $\boldsymbol{\Sigma}_{\theta_0}^{-1}$ corresponding to parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ be denoted by $\left(\boldsymbol{\Sigma}_{\theta_0}^{-1}\right)_{\theta_i\theta_j}$, we have $\left(\boldsymbol{\Sigma}_{\theta_0}^{-1}\right)_{\xi\xi} = O(1)$, $\left(\boldsymbol{\Sigma}_{\theta_0}^{-1}\right)_{\xi\theta_c} = O(1)$, $\left(\boldsymbol{\Sigma}_{\theta_0}^{-1}\right)_{\theta_c\theta_q} = O(1)$ for $c \neq q$, and $\left(\boldsymbol{\Sigma}_{\theta_0}^{-1}\right)_{\theta_c\theta_c} = O\left(\frac{N}{N_c}\right)$.

Distribution of $\frac{1}{\sqrt{NT}}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}$. For the elements of $\frac{1}{\sqrt{NT}}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}$, the central limit theorem for martingale arrays can be applied. Together with (12), this gives

$$\frac{1}{\sqrt{NT}}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_{(k+2+3C)\times 1},\ \lim_{T\to\infty} \boldsymbol{\Sigma}_{\theta_0} + \boldsymbol{\Omega}_{\theta_0}\right), \tag{13}$$

[1] Different convergence scenarios with $\frac{N_c}{N} \to 0$ are possible and would generally require a different scaling of $\boldsymbol{\Sigma}_{\theta_0}$; this case is left for future research.

where Lemma 13 in Yu et al 2008 applies under Assumptions 1–5.

Distribution of $\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. A Taylor expansion gives

$$\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left(-\frac{1}{NT}\frac{\partial^2 LL(\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal}\right)^{-1}\left(\frac{1}{\sqrt{NT}}\frac{\partial LL^*(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} - \Delta_{\theta_0}\right), \tag{14}$$

where $\bar{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. The Hessian term in (14) can be decomposed as

$$-\frac{1}{NT}\frac{\partial^2 LL(\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} = \left(-\frac{1}{NT}\frac{\partial^2 LL(\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} + \frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal}\right) + \left(-\frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} - \boldsymbol{\Sigma}_{\theta_0}\right) + \boldsymbol{\Sigma}_{\theta_0}.$$

By Lemma 15 in Yu et al, 2008, the elements of $\left(-\frac{1}{NT}\frac{\partial^2 LL(\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} + \frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal}\right)$ are of the order $||\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||$ multiplied by an $O_p(1)$ term for the $(\xi\xi)$ block, by $O_p\left(\frac{N_c}{N}\right)$ terms for the $(\theta_c\theta_c)$ blocks, by $\left[O_p\left(\frac{N_c}{N}\right), O_p\left(\frac{N_c}{NT}\right), O_p\left(\frac{N_c}{NT}\right)\right]$ terms for the $(\xi\theta_c)$ blocks, and by zero terms for the $(\theta_c\theta_q, c \neq q)$ blocks. The elements of $\left(-\frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} - \boldsymbol{\Sigma}_{\theta_0}\right)$ are $O_p\left(\frac{1}{\sqrt{NT}}\right)$ for the $(\xi\xi)$ block, zero for the $(\theta_c\theta_q, c \neq q)$ blocks, and $O_p\left(\frac{\sqrt{N_c}}{N\sqrt{T}}\right)$ for the remaining entries.

Consequently, $-\frac{1}{NT}\frac{\partial^2 LL(\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} = \boldsymbol{\Sigma}_{\theta_0} + ||\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|| \cdot O_p(1) + O_p\left(\frac{1}{\sqrt{NT}}\right)$. The order of $\left(-\frac{1}{NT}\frac{\partial^2 LL(\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal}\right)^{-1}$ is therefore the same as the order of $\boldsymbol{\Sigma}_{\theta_0}^{-1} + R_{NT}$, where R_{NT} is a stochastic term negligible relative to $\boldsymbol{\Sigma}_{\theta_0}^{-1}$, since $\left\|\boldsymbol{\Sigma}_{\theta_0}^{-1}\left(\left(-\frac{1}{NT}\frac{\partial^2 LL(\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} + \frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal}\right) + \left(-\frac{1}{NT}\frac{\partial^2 LL(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} - \boldsymbol{\Sigma}_{\theta_0}\right)\right)\right\| = o_p(1)$ provided that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = o_p(1)$.[2] It follows that $\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left(\boldsymbol{\Sigma}_{\theta_0}^{-1} + o_p(1)\right)\left(O_p(1) + \sqrt{\frac{N}{T}}\varphi\right)$, and, using Assumption 8, $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O_p\left(\max\left(\sqrt{\frac{1}{NT}}, \frac{1}{T}\right)\right)$. This matches the argument used in the proof of Theorem 3 in Yu et al 2008, and therefore the same reasoning yields

$$\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \sqrt{\frac{N}{T}}\boldsymbol{\Sigma}_{\theta_0}^{-1}\varphi + O_p\left(\max\left(\sqrt{\frac{N}{T^3}}, \sqrt{\frac{1}{T}}\right)\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}_{(k+2+3C)\times 1}, \lim_{T\to\infty}\boldsymbol{\Sigma}_{\theta_0}^{-1}\left(\lim_{T\to\infty}\boldsymbol{\Sigma}_{\theta_0} + \boldsymbol{\Omega}_{\theta_0}\right)\lim_{T\to\infty}\boldsymbol{\Sigma}_{\theta_0}^{-1}\right).$$

4 Conclusions

This paper extends the bias-correction analysis of Yu et al 2008 to a hierarchical spatial dynamic panel model with higher-level-unit-specific spatial dependence in both the dependent variable and the disturbances, together with heterogeneous innovation variances. After concentrating out the fixed effects, the normalized QML score is decomposed into a leading stochastic component and a time-demeaning term that generates the incidental parameters bias.

The expected Hessian, the covariance matrix of the score, and their asymptotic orders are characterized, and the corresponding asymptotic distribution of the QML estimator is stated. The derived score expansion and leading bias term provide the main analytical ingredients required for extending the correction of Yu et al 2008 to the hierarchical setting. Future research will construct a feasible bias-corrected estimator, establish its formal consistency under alternative convergence regimes, and assess its finite-sample performance through MC simulations.

Acknowledgements

The work was supported by the Prague University of Economics and Business project IGS F4/29/2026.

References

Anselin L (1988) Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht

Lee Lf, Yu J (2010) Estimation of spatial autoregressive panel data models with fixed effects. Journal of Econometrics 154(2):165–185. https://doi.org/10.1016/j.jeconom.2009.08.001

Yu J, de Jong R, Lee L (2008) Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and t are large. Journal of Econometrics 146(1):118–134. https://doi.org/10.1016/j.jeconom.2008.08.002

[2] The condition $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = o_p(1)$ requires consistency of $\hat{\boldsymbol{\theta}}$. A formal proof of consistency is outside the scope of this paper and is therefore assumed implicitly.

The Use of a Two-Step Decomposition Approach in Service Vehicle Route Planning

Petr Kozel[1]

Abstract. Route planning for service vehicles is one of the core responsibilities of every dispatcher. Given that individual routes are desirable to be optimally designed based on selected criteria, it is advisable to use quantitative decision-support tools for this purpose. When planning a single route, it is generally assumed that the total demand from customers located at the various nodes of the transport network is less than or equal to the capacity of the service vehicle. However, in many real-world cases, this condition cannot be met, and it is necessary to use one of the appropriate approaches to solve the problem. This paper focuses on the use of a two-step decomposition approach, which falls within the subset of methods known as Cluster First – Route Second. Using this approach, it is possible to decompose the initial set of requirements into subsets within which partial routes are then proposed. The two-step decomposition approach is implemented using linear programming and is tested on both sample and real-world data. The results of these computational experiments are presented in this paper.

Keywords: vehicle routing planning, cluster first-route second, linear programming, waste management, optimization

JEL Classification: C44

1 Introduction

Route planning for service vehicles is one of the core responsibilities of every dispatcher. In practice, this often involves route planning for regularly recurring delivery or collection services, such as the delivery of groceries, lunches, and mail, or the collection of municipal and sorted waste, etc. Given that individual routes are desirable to be optimally designed based on selected criteria, it is advisable to use quantitative decision-support tools for this purpose. Common optimization criteria include, for example, total mileage, total service time, etc. Minimizing these selected criteria can yield not only economic benefits in the form of reduced operating costs but also environmental benefits resulting from the elimination of unnecessary or unproductive trips.

The issue of service vehicle routing falls under the category of routing problems, which can be classified according to various criteria. One of the fundamental criteria is whether the service route serves the edge or the vertex of the transport network. In this text, the focus will be on the latter of these two options. When planning a single route, it is generally assumed that the total demand from customers located at the various vertices of the transport network is less than or equal to the capacity of the service vehicle (Laporte, 1992). However, in many real-world scenarios, this condition cannot be met, and it is necessary to use one of the appropriate approaches, such as the *Cluster First – Route Second* method, the *Route First – Cluster Second* method, the *Saving – Insertion* method, and others (Laporte, 2009). This paper focuses on the use of a two-step decomposition approach, which falls within the subset of methods known as Cluster First – Route Second (Beasley, 1983). Using this approach, it is possible to decompose the initial set of requirements into subsets within which partial routes are then proposed.

The proposed two-step decomposition approach uses a *p-median* method, in which the vertices are treated as customers assigned to individual centers. The number of these centers corresponds to the selected number of subsets into which the initial set of customers and their requirements is decomposed. The approach consists of the following steps:

- Solving the *p-median* problem when relaxing the constraints related to customer requirements and center capacities. In this step, p centers are located, and individual customers are assigned to these centers regardless of the capacities of the centers.
- Solving a reallocation problem in which customers are assigned from overloaded to underloaded centers.

[1] VSB - Technical University of Ostrava/Faculty of Economics, Department of Mathematical Methods in Economics, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, petr.kozel@vsb.cz

2 Formulation of the problem

A transportation network is given, represented by an undirected graph $G = (V, E, d)$. This graph has three attributes: V represents the set of vertices, E represents the set of edges, and d_{ij} represents the weight of an edge ij, where d_{ij} denotes the length of the road in kilometers. Customers j with **unit** requirements b_j are located at the individual vertices of the graph. The goal of the problem is to decompose the set of vertices V into p subsets of predefined capacity a, such that:

$$a = \frac{\sum_{j \in V} b_j}{p}, \tag{1}$$

if the sum of customer requirements is not divisible by the number of centers without a remainder.

2.1 Customer's unit requirement

At this moment, it is necessary to pay attention to the assumption that has been made, namely the customer's unit requirement. This assumption is theoretically very strong; however, it is entirely justified in many real-world applications, such as route planning for municipal waste collection vehicles. In the field of strategic management, where individual subsets of customers to be served during a single service vehicle trip are identified, the analysis typically uses only the total amount of mixed waste generated in the area in question (municipality) over a selected time period (year). Municipalities in the Czech Republic often have a fixed annual waste collection schedule. Based on the total amount of municipal waste generated within the municipality (t/year), the average values corresponding to the individual customers (kg/year) are then determined. By adhering to the specified fixed collection interval and taking into account an appropriate operational buffer, the average unit size of a customer's request (in kilograms) can be determined. In situations where information on the current fill level of the bins is not available (requiring IoT technology) (Faccio et al, 2011) or information on the fill history of the bins is not available (requiring the weighing of individual unloads is used), the customer's average unit requirement is used. From a strategic management perspective, where the average size of a customer's request is determined based on the total annual waste generation for the entire service area of the previous period, this simplification is entirely acceptable. The resulting inaccuracy is offset by the operational reserve built into the capacity of the service vehicle. In real-world situations where the specific request of a customer cannot be justified in this way and this simplification cannot be accepted, the proposed approach cannot be used.

2.2 Problem on the p-median

In the first phase, we will address the classic p-median problem while relaxing the constraints related to customer demands and the capacity of the placement centers. The mathematical model is as follows (Janáček, 2002; Janáček and Kvet, 2016):

$$\text{Min} \sum_{i \in V} \sum_{j \in V} d_{ij} \cdot r_{ij} \tag{2}$$

Subject to:

$$\sum_{i \in V} r_{ij} = 1, \text{ for } j \in V \tag{3}$$

$$r_{ij} \leq y_i, \text{ for } i \in V,\ j \in V \tag{4}$$

$$\sum_{i \in V} y_i = p \tag{5}$$

$$r_{ij} = \{0, 1\}, \text{ for } i \in V,\ j \in V \tag{6}$$

$$y_i = \{0, 1\}, \text{ for } i \in V \tag{7}$$

The mathematical model (2) – (7) includes the bivalent variables r_{ij} and y_i. The variable r_{ij} models whether or not vertex j is assigned to vertex i. The variable y_i models whether or not a center is located at the vertex i. The objective function (2) represents the total distance between the assigned service centers and their assigned customers. The constraints (3) ensure that each customer j is assigned to exactly one vertex i. Condition (4)

ensures that a customer j can be assigned only to a vertex i in which a center is located. The condition (5) ensures that exactly p centers will be located. Conditions (6) and (7) are mandatory conditions and define the variables domains.

2.3 The reallocation problem

After the first phase of the decomposition approach, information on the location of the centers y_i and information on the assignment of customers to individual centers r_{ij} is available. This information will be used in the second phase of the decomposition approach.

Information about the location of the centers will be represented by the set of centers S. This set must be divided into two subsets S^+ and S^-. The first subset consists of centers for which the following holds:

$$\sum_{j \in V} b_j \cdot r_{ij} > \lfloor a \rfloor, \text{ for } i \in V \tag{8}$$

These overloaded centers will be represented by the set S^+. The second subset consists of centers for which the following are valid:

$$\sum_{j \in V} b_j \cdot r_{ij} \leq \lfloor a \rfloor, \text{ for } i \in V \tag{9}$$

These underloaded centers will be represented by the set S^-. It holds that $S = S^+ \cup S^-$. We will now work with a set of customers J consisting of two subsets. The first subset consists of customers who were assigned to the overloaded centers in the first phase. This set is denoted by the symbol J^+. The second subset consists of customers who were assigned to underloaded centers in the first phase. This set is denoted by the symbol J^-. It holds that $V = J^+ \cup J^-$.

Now we need to calculate the cost matrix c_{ij}. As part of the reallocation process, customers who were initially assigned to overloaded centers (customers from the set J^+) will be reassigned to underloaded centers (centers from the set S^-). This situation is illustrated in Figure 1.

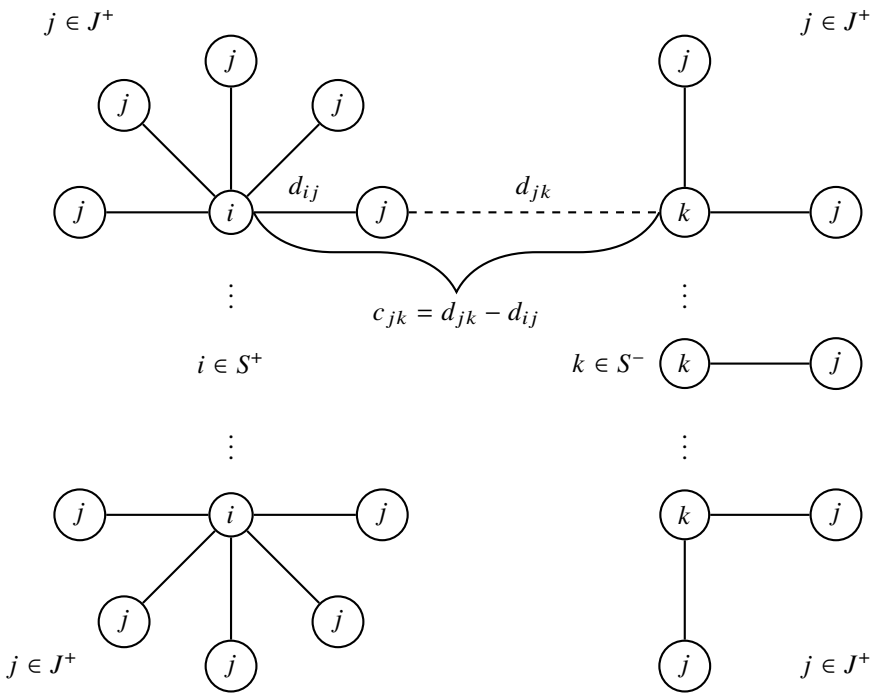


Figure 1 Reallocation process

Vertices labeled with the symbol j represent individual customers from the set J. Vertices labeled with the symbol i represent overloaded centers from the set S^+. Customers j connected by solid lines to vertices i are customers from the set J^+. Vertices labeled with the symbol k represent underloaded centers from the set S^-. Customers j

connected by a solid line to the vertices k are customers from the set J^-. The distance d_{ij} is the distance between the i-th overloaded center and the j-th customer. The distance d_{jk} is then the distance between the j-th customer and the k-th underloaded center. When calculating the costs c_{jk}, the distance d_{ij} will be saved, but the distance d_{jk} will be included. The dashed line represents a possible customer assignment from the i-th overloaded center to the k-th underloaded center. The formula for calculating the cost matrix is as follows:

$$c_{jk} = d_{jk} - d_{ij}, \text{ for } i \in S^+,\ j \in J^+,\ k \in S^-, \text{where } r_{ij} > 0 \tag{10}$$

It is now possible to formulate a mathematical model of the rellocation of customers from overloaded centers to underloaded centers.

$$\text{Min} \sum_{j \in J^+} \sum_{k \in S^-} c_{jk} \cdot z_{jk} \tag{11}$$

Subject to:

$$\sum_{j \in J^+} b_j \cdot z_{jk} \geq \lceil a \rceil - \sum_{j \in J^-} b_j \cdot r_{kj}, \text{ for } k \in S^- \tag{12}$$

$$\sum_{j \in J^+} \sum_{k \in S^-} b_j \cdot z_{jk} \cdot r_{ij} \leq \sum_{j \in J^+} b_j \cdot r_{ij} - \lfloor a \rfloor, \text{ for } i \in S^+ \tag{13}$$

$$\sum_{k \in S^-} z_{jk} \leq 1, \text{ for } j \in J^+ \tag{14}$$

$$z_{jk} \in \{0, 1\}, \text{ for } j \in J^+,\ k \in S^- \tag{15}$$

In the mathematical model (11) – (15), there is a bivalent variable z_{jk} that models the assignment or non-assignment of a customer j from the set J^+ to underloaded centers from the set S^-. Conditions (12) are formulated for each underloaded center k and ensure that the sum of customer requirements assigned to the k-th center is greater than or equal to the difference between the defined capacity of the centers $\lceil a \rceil$ and the sum of requirements already assigned to the center k. Conditions (13) are formulated for each overloaded center i and ensure that the sum of customer requirements assigned to all centers k in the set S^- is less than or equal to the difference between the sum of requirements already assigned to the center i and the defined capacity of the centers $\lfloor a \rfloor$. Conditions (14) ensure that each customer j from the set J^+ can be assigned to at most one underloaded center. Conditions (15) are mandatory conditions and define the domain of the variable z_{jk}.

3 Numerical experiments

Computational experiments using the two-step decomposition approach described above were conducted using both model data and real-world data. For the experiments on model data, we used, among other resources, a library of sample instances for the TSP (Reinelt, 2013). In total, there are 18 problems, which are graded according to their size, or rather the number of vertices. It is assumed that each vertex of the network represents a customer j with a **unit** requirement b_j. For each problem, the number of subsets p into which the initial set of customers must be decomposed was selected, and the capacity $\lfloor a \rfloor$ of that center was also calculated. The parameters of the individual model problems and a summary of the results obtained are presented in Table 1.

As can be seen in Table 1, for all problems where the number of vertices was divisible by the chosen number of centers without remainder, completely uniform clusters were obtained. Otherwise, the clusters consist of the numbers $\lfloor a \rfloor$ and $\lceil a \rceil$. However, in both cases, from a practical perspective, we can consider these clusters to be uniform.

In computational experiments using real-world data, the proposed two-step decomposition approach has been repeatedly applied in practice. Here, we present a selected example from the field of municipal waste collection in the Moravian-Silesian Region. The service area consists of 428 service locations. Each of these service locations is equipped with a specified number of uniform 110-liter bins. However, an analysis of the input data revealed that in 415 cases (97 %) there was only one container. Based on long-term data on the total amount of waste

Problem	Vertices	p	Cluster sizes
1	127	3	43/42/42
2	150	3	50/50/50
3	194	2	97/97
4	299	3	99/100/100
5	318	3	106/106/106
6	400	4	100/100/100/100
7	439	4	110/109/110/110
8	574	4	143/144/143/144

Problem	Vertices	p	Cluster sizes
9	657	3	219/219/219
10	734	3	245/245/244
11	783	4	196/196/196/195
12	813	4	203/204/203/203
13	1002	5	201/200/200/200/201
14	2628	4	657/657/657/657
15	3496	4	874/874/874/874
16	4250	4	850/850/850/850

Table 1 Parameters and summary of results for individual model problems

collected in this area, the waste collection company has determined that the average weight of a bin is 19.8 kg. It also specified the capacity of the collection vehicle, which is 4,500 kg. Based on this initial information, it was determined that a total of approximately 8,500 kg of waste is ready for collection. Given the specified capacity of the service vehicle, it is therefore necessary, as part of the strategic decision, to propose two routes for the service vehicle and to divide the initial set of service locations into two equal subsets.

A two-step decomposition approach was used for this decomposition. In the first phase, the mathematical model (2) – (7) was used to identify two centers and their assigned service locations. However, the number of service locations assigned to individual centers (respectively, clusters) was uneven, as shown in Table 2. Cluster II was overloaded. This imbalance was resolved during the second phase of the two-step decomposition approach using the mathematical model (11) – (15), where 49 service locations from cluster II were added to cluster I.

Cluster	Service locations
I	165
II	263

Table 2 The number of service locations assigned to individual clusters

Given that a two-step decomposition approach was used to solve a practical problem, routes for individual collection vehicles were proposed within the identified clusters. A heuristic approach based on the Lin-Kernighan method was used to design these routes (Lin and Kernighan, 1973). Table 3 summarizes a comparison of the final design and the initial state.

Cluster	Service locations	Load (kg)	Route length (km)
Initial	428	8583.3	63.5
I	214	4336.2	31.9
II	214	4247.1	31.8

Table 3 The comparison of the final design and the initial state

As shown in Table 3, the initial number of service locations was divided into two equal clusters. For each cluster, the total actual load (kg) to be collected is listed, taking into account the actual number of bins. The length of both service vehicle routes is also listed in kilometers. As can be seen in Table 3, in this particular case, the total length of the proposed routes - 63.7 km (31.9+31.8) - corresponds fully to the number of kilometers that a waste collection company providing this service typically has to cover. In this particular case, it was determined that the current waste collection system is effective.

All these above mentioned numerical experiments were solved with the optimization software FICO Xpress 8.3 (64-bit, release 2017) (FICO, 2026). The individual models were run on a PC equipped with the AMD Ryzen 7 2700X Eight-Core Processor 3.7 GHz and 32 GB RAM.

4 Conclusion

This paper focused on the use of a two-step decomposition approach, which falls within the subset of methods known as Cluster First – Route Second. Using this approach, it is possible to decompose the initial set of requirements into subsets within which partial routes for service vehicles can then be designed. The two-step decomposition approach was implemented using linear programming and was tested on both sample data and real-world data from the municipal waste collection sector in the Moravian-Silesian Region. From the perspective of future research, it would be useful to investigate the potential limitations of this approach, particularly when applied to input data containing many times more vertices.

References

Beasley J (1983) Route First – Cluster Second Methods for Vehicle Routing. Omega 11(4):403–408

Faccio M, Persona A, Zanin G (2011) Waste collection multi objective model with real time traceability data. Waste Management 31(12):2391–2405

FICO (2026) Xpress-Mosel: User guide. `https://www.fico.com/fico-xpress-optimization/docs/latest/mosel/UG/dhtml/GUID-5DBA5E00-B9BC-3A85-A9AF-A03B2F564C43.html`

Janáček J (2002) Optimalizace na dopravních sítích. Žilinská univerzita v Žilině, Žilina

Janáček J, Kvet M (2016) Sequential approximate approach to the p-median problem. Computers & Industrial Engineering 94:83–92

Laporte G (1992) The Vehicle Routing Problem: An overview of exact and approximate algorithms. European Journal of Operational Research 59(3):345–358

Laporte G (2009) Fifty Years of Vehicle Routing. INFORMS 43(4):408–416

Lin S, Kernighan BW (1973) An Effective Heuristic Algorithm for the Traveling-Salesman Problem. Operations Research 21(2):498–516

Reinelt G (2013) TSPLIB: Library of sample instances for the TSP (and related problems) from various sources and of various types. `https://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/tsp.html`

Impact of GDP Growth, Inflation, Industrial Employment, and Trade Openness on Foreign Direct Investment in the Visegrad Group Countries

Radmila Krkošková[1], Zuzana Neničková[2], Lucie Waleczek Zotyková[3]

Abstract. This paper analyses the association between selected macroeconomic indicators and foreign direct investment (FDI) in the Visegrad Group countries (the Czech Republic, Hungary, Poland, and Slovakia) during the period 2000-2025. The analysis applies the Panel ARDL model together with country-specific Bounds tests to examine short-run and long-run relationships between FDI, GDP growth, inflation, employment in industry, and trade openness. Panel ARDL methodology was used. The long-run results indicate that GDP growth and employment in industry are positively associated with FDI, while trade openness has a negative coefficient and inflation is not statistically significant. In the short run, inflation is negatively associated with FDI, whereas trade openness has a positive coefficient. The error correction term is negative and statistically significant, suggesting adjustment towards a long-run equilibrium. Country-specific Bounds tests confirm cointegration only for the Czech Republic and Slovakia, highlighting heterogeneity across the V4 countries. The contribution of the study is empirical rather than methodological: it provides updated evidence for the V4 economies, distinguishes between short-run and long-run effects, and explicitly discusses the limitations arising from the small sample, omitted variables, common shocks, and possible endogeneity.

Keywords: ARDL model, foreign direct investment, inflation, trade openness, V4

JEL Classification: F21, C19, C33

1 Introduction

Foreign direct investment (FDI) is an important driver of economic growth, technological progress, and international competitiveness, particularly in transition economies. As a source of capital, innovation, and managerial know-how, FDI contributes to economic development and integration into global value chains. Consequently, identifying the determinants of FDI has become an important topic in economic research.

The Visegrad Group (V4)—the Czech Republic, Hungary, Poland, and Slovakia—provides an appropriate environment for analysing FDI determinants. Although these countries share a similar historical background and European Union membership, they differ in their economic structures, labour markets, and investment attractiveness. Such differences make it possible to examine whether the macroeconomic determinants of FDI are common across the V4 countries.

This study investigates the effects of GDP growth (GDP), inflation (INF), employment in industry (EMI), and trade openness (TRO) on FDI inflows in the Visegrad Group during 2000-2025. The analysis applies the Panel Autoregressive Distributed Lag (Panel ARDL) model, which enables the estimation of both short-run and long-run relationships while accommodating variables integrated of order I(0) and I(1). Country-specific Bounds tests are additionally employed to assess long-run equilibrium relationships.

The contribution of this paper is mainly empirical and should be interpreted in a modest sense. The study distinguishes between short-run dynamics and long-run equilibrium relationships within the Panel ARDL framework and complements pooled estimation with country-specific Bounds tests. This combination allows the paper to show that the estimated long-run relationships are not equally strong across all V4 countries. At the same

[1] School of Business Administration in Karviná, Silesian University in Opava, Department of Informatics and Mathematics, Univerzitní náměstí 1934/3, 733 40 Karviná, Czech Republic, e-mail: krkoskova@opf.slu.cz
[2] School of Business Administration in Karviná, Silesian University in Opava, Department of Informatics and Mathematics, Univerzitní náměstí 1934/3, 733 40 Karviná, Czech Republic, e-mail: nenickova@opf.slu.cz
[3] School of Business Administration in Karviná, Silesian University in Opava, Department of Informatics and Mathematics, Univerzitní náměstí 1934/3, 733 40 Karviná, Czech Republic, e-mail: waleczekzotykova@opf.slu.cz

time, because the number of countries is small and the set of explanatory variables is limited, the results should be read as conditional statistical associations rather than causal estimates.

2 Literature Review and Data

2.1 Literature Review

Foreign direct investment (FDI) has long been recognised as one of the key drivers of economic development, particularly in emerging and transition economies. From a theoretical perspective, one of the most influential explanations of FDI is Dunning's Eclectic (OLI) Paradigm, which states that foreign investment decisions are determined by the interaction of ownership advantages, location advantages, and internalization incentives. Among location-specific advantages, macroeconomic stability, market size, labour market characteristics, and openness to international trade are considered particularly important determinants of FDI.

Market conditions are traditionally represented by GDP level, which captures the absolute size of an economy. Larger economies generally provide higher expected returns, stronger consumer demand, and better investment opportunities. In this paper, however, the empirical variable is GDP growth rather than GDP level. It should therefore be interpreted as an indicator of economic growth and business-cycle dynamics, not as a direct measure of market size. Numerous empirical studies confirm a positive relationship between GDP-related indicators and FDI inflows (Louail, 2019; Rachdi et al., 2016; Mohammad et al., 2022). Economic growth is therefore considered one of the robust macroeconomic correlates of foreign investment.

Macroeconomic stability is another important factor influencing investment decisions. Inflation is commonly used as an indicator of economic uncertainty because persistently high inflation reduces the predictability of future costs and returns. Most empirical studies report a negative relationship between inflation and FDI (Louail, 2019; Dondashe and Phiri, 2018; Mohammad et al., 2022). Nevertheless, several authors argue that moderate inflation may simply reflect expanding economic activity and therefore need not discourage foreign investors (Erdogan and Unver, 2015; Hoang and Bui, 2015). These mixed findings indicate that the effect of inflation depends on the broader macroeconomic environment.

Labour market conditions also play an important role in attracting foreign investors. Countries with a developed industrial sector and a sufficiently skilled labour force are generally more attractive for efficiency-seeking FDI. In this study, employment in industry is used as a proxy for industrial capacity and labour availability. Although employment in industry has received less attention than GDP or inflation, previous studies suggest that industrial development and labour market quality positively affect FDI inflows by increasing production efficiency and reducing investment risk.

Trade openness is another frequently analysed determinant of FDI. Greater openness generally reflects stronger integration into international markets, lower trade barriers, and easier access to foreign customers and suppliers. Most studies therefore report a positive relationship between trade openness and FDI (Rachdi et al., 2016; Dondashe and Phiri, 2018). However, empirical evidence is not unanimous. In some transition economies, high trade openness may substitute for foreign production or reflect structural characteristics that weaken domestic investment incentives, resulting in a negative relationship (Bobenič Hintošová et al., 2018). Consequently, the impact of trade openness appears to depend on country-specific economic conditions.

In addition to the macroeconomic variables examined in this study, previous literature identifies several other factors influencing FDI, including institutional quality, political stability, taxation, exchange rates, labour costs, infrastructure, and the quality of governance (Mohammad et al., 2022) and (Shaari et al., 2023). Since the primary objective of this paper is to analyse the role of selected macroeconomic determinants, these variables are not included in the estimated model.

2.2 Data and Methods

The empirical analysis is based on annual panel data covering the period 2000-2025 for the four Visegrad Group (V4) countries: the Czech Republic, Hungary, Poland, and Slovakia. Data for the Czech Republic, Poland, and Slovakia were obtained from the World Bank database. For Hungary, FDI data adjusted for Special Purpose Entities (SPEs) were collected from the National Bank of Hungary because conventional balance-of-payments statistics are substantially influenced by SPE transactions. This adjustment is economically justified in the Hungarian case, but it may create a comparability limitation because the Hungarian FDI series is not fully identical in construction to the World Bank series used for the other V4 countries.

The selection of explanatory variables is based on economic theory and previous empirical studies discussed in the literature review. Foreign direct investment (FDI), measured as net inflows (% of GDP), is considered the dependent variable. The explanatory variables include GDP growth (GDP), inflation measured by the consumer price index (INF), employment in industry (EMI), and trade openness (TRO). GDP growth captures economic growth and business-cycle conditions rather than absolute market size. Employment in industry is used as a proxy for the industrial structure and labour capacity of the economy. Trade openness captures international economic integration. Their expected effects on FDI are summarised in Table 1. Important FDI determinants such as taxation, wages, exchange rates, institutional quality, infrastructure, and political stability are not included, which creates omitted-variable concerns and limits causal interpretation.

Variable designation	Description and economic interpretation	Expected sign and literature support
Foreign Direct Investment FDI	Foreign direct investment. Net inflows of foreign direct investment in the reporting economy, expressed as a percentage of GDP. Economic interpretation: FDI intensity.	Dependent variable
Inflation INF	Inflation measured by the consumer price index; annual percentage change in the consumer price basket. Economic interpretation: macroeconomic stability and uncertainty.	NEGATIVE Rachdi et al. (2016); Dondashe and Phiri (2018)
Employment in Industry EMI	Employment in industry, expressed as a percentage of total employment. Economic interpretation: industrial structure, industrial capacity, and availability of an industrial labour force.	POSITIVE Bruno et al. (2021); Gorbunova et al. (2012)
GDP growth GDP	Annual percentage growth rate of GDP at market prices based on constant local currency. Economic interpretation: economic growth and business-cycle dynamics, not a direct measure of market size.	POSITIVE Louail (2019); Mohammad et al. (2022)
Trade Openness TRO	Sum of exports and imports of goods and services measured as a share of GDP. Economic interpretation: international economic integration.	POSITIVE Rachdi et al. (2016); Bobenič Hintošová et al. (2018)

Source: The World Bank (2025), National Bank of Hungary (2025)

Table 1 Description of Variables

Since the analyzed variables exhibit a mixed order of integration, the Panel Autoregressive Distributed Lag (Panel ARDL) model, proposed by Pesaran, Shin and Smith (2001), was used. The Panel ARDL approach is suitable when variables are integrated of order I(0) and I(1), provided that none of them is integrated of order I(2). Unlike a static panel regression, the Panel ARDL framework explicitly includes lagged values of the dependent variable and the regressors and therefore separates long-run equilibrium relationships from short-run dynamic adjustments.

The general dynamic Panel ARDL(p, q_1, ..., q_K) specification can be written as

$$FDI_{it} = \alpha_i + \sum_{j=1}^{p} \phi_{ij} FDI_{i,t-j} + \sum_{j=0}^{q_1} \beta_{1ij} GDPG_{i,t-j} + \sum_{j=0}^{q_2} \beta_{2ij} INF_{i,t-j} + \sum_{j=0}^{q_3} \beta_{3ij} EMI_{i,t-j} + \sum_{j=0}^{q_4} \beta_{4ij} TRO_{i,t-j} + \delta_t + \varepsilon_{it}$$

where i denotes the country, t denotes time, X_{it} is the vector of explanatory variables (EMI, GDP, INF, and TRO), α_i captures country-specific intercepts, and ε_{it} is the random error term.

For interpretation purposes, the Panel ARDL model is reparametrized into the Error Correction Model (ECM), which distinguishes between short-run dynamics and the long-run equilibrium relationship:

$$\Delta FDI_{it} = \lambda_i \left(FDI_{i,t-1} - \theta_{1i} GDPG_{i,t-1} - \theta_{2i} INF_{i,t-1} - \theta_{3i} EMI_{i,t-1} - \theta_{4i} TRO_{i,t-1} \right)$$
$$+ \sum_{j=1}^{p-1} \psi_{ij} \Delta FDI_{i,t-j} + \sum_{j=0}^{q_1-1} \gamma_{1ij} \Delta GDPG_{i,t-j} + \sum_{j=0}^{q_2-1} \gamma_{2ij} \Delta INF_{i,t-j} + \sum_{j=0}^{q_3-1} \gamma_{3ij} \Delta EMI_{i,t-j} + \sum_{j=0}^{q_4-1} \gamma_{4ij} \Delta TRO_{i,t-j} + \delta_t + u_{it}$$

where θ_i is the error correction coefficient measuring the speed at which deviations from the long-run equilibrium are eliminated. A negative and statistically significant coefficient confirms adjustment towards a stable long-run relationship. In the PMG specification, the long-run coefficients theta are restricted to be homogeneous across countries, while the short-run coefficients, intercepts, and error variances may differ across countries.

The Panel ARDL model was estimated using the Pooled Mean Group (PMG), Mean Group (MG), and Dynamic Fixed Effects (DFE) estimators proposed by Pesaran et al. (2001). The final specification was selected on the basis of the Hausman specification test, which compares the consistency and efficiency of the PMG and MG estimators. Since the Hausman test did not reject the null hypothesis of long-run parameter homogeneity, the PMG estimator was considered the most appropriate for the subsequent analysis. This estimator assumes homogeneous long-run coefficients while allowing short-run coefficients, intercepts, and error variances to differ across countries. Given the limited sample size (N = 4 countries and T = 26 annual observations), the results of PMG, MG, DFE, and related panel tests should be interpreted cautiously and should not be treated as strong causal evidence.

Before estimating the Panel ARDL model, panel unit root tests were performed to determine the order of integration of all variables. The results indicated that the variables are integrated either of order I(0) or I(1), satisfying the main prerequisite for applying the Panel ARDL methodology. However, because the panel consists of only four countries, the power and finite-sample reliability of panel unit-root tests are limited. The optimal lag structure was selected according to the Akaike Information Criterion.

Finally, the existence of long-run equilibrium relationships was examined using the Bounds testing approach proposed by Pesaran et al. (2001). Unlike traditional residual-based cointegration tests, the Bounds test is consistent with the ARDL framework and allows cointegration analysis when regressors are a mixture of I(0) and I(1) variables. The country-specific Bounds tests were used as a complementary check because the pooled PMG estimator may conceal heterogeneity across individual V4 economies.

3 Data Analysis

3.1 Panel Autoregressive Distributed Lag Model (ARDL)

Before estimating the Panel ARDL model, the order of integration of all variables was examined using four commonly applied panel unit root tests: Levin-Lin-Chu (LLC), Im-Pesaran-Shin (IPS), Fisher-ADF, and Fisher-PP. The results indicate that FDI and GDP growth are stationary in levels (I(0)), whereas INF, EMI, and TRO are integrated of order one (I(1)). Since no variable is integrated of order two, the prerequisite for applying the Panel ARDL methodology is satisfied. Nevertheless, these results should be treated as indicative rather than definitive because the panel contains only four cross-sectional units.

This part of the article presents the results of the Panel ARDL estimates for the V4 countries. The ARDL approach is specified as a dynamic reduced-form equation in which FDI is explained by its own lagged values and by current and lagged values of the selected macroeconomic variables. Given the use of annual data and the short time dimension, one lag was chosen based on the minimum Akaike Information Criterion. Before interpreting the estimation results, the Hausman specification test was performed to compare the PMG and MG estimators. Since the null hypothesis of homogeneous long-run coefficients was not rejected (p-value = 0.431), the PMG estimator was selected for the final analysis. Consequently, only the PMG estimates are discussed below. The model does not include explicit time dummies for common shocks such as EU accession, the global financial crisis, COVID-19, or the energy shock; therefore, part of the estimated association may reflect these unmodelled events.

Variable	Coefficient	Std. Error	t-Statistic	Significance
EMI	0.371	0.068	5.367	0.0000***
GDP	0.226	0.101	2.262	0.0261**
INF	0.092	0.072	1.266	0.2088
TRO	-0.035	0.007	-5.252	0.0000***
C	-5.604	2.295	-2.441	0.0166**

Statistical significance at the 0.01 level (***), at the 0.05 level (**), at the 0.1 level (*).
Source: The World Bank (2025), own calculations.

Table 2 PMG (1,1,1,1,1) Long Run (Pooled) Model

Table 2 presents the estimated long-run coefficients obtained from the PMG estimator. Employment in industry exhibits the strongest positive association with FDI, indicating that economies with a larger industrial employment

share tend to attract higher foreign investment. This finding is consistent with the efficiency-seeking motive of multinational enterprises, which generally prefer locations with an established industrial base and experienced labour force. However, this coefficient should not be interpreted as a pure causal effect because FDI may itself influence industrial employment.

GDP growth has a statistically significant positive coefficient, indicating that stronger economic growth is associated with higher FDI inflows. However, this should not be interpreted as a direct market-size effect, since the model uses GDP growth rather than GDP level. Inflation is not statistically significant in the long run. Trade openness has a negative long-run coefficient, which may reflect a substitution effect between trade and local production. This result may also be influenced by omitted factors such as institutional quality, taxation, exchange rates, and investment incentives.

Variable	Coefficient	Std. Error	t-Statistic	Significance
COINTEQ	-0.906	0.198	-4.878	0.0000***
D(EMI)	-0.519	1.115	-0.466	0.6422
D(GDP)	-1.021	0.989	-1.032	0.3048
D(INF)	-0.974	0.466	-2.086	0.0291**
D(TRO)	0.595	0.294	2.018	0.0311**

Statistical significance at the 0.01 level (***), at the 0.05 level (**), at the 0.1 level (*).
Source: The World Bank (2025), own calculations.

Table 3 PMG Error-Correction Model: Short-Run Coefficients and Adjustment Term

Table 3 reports the error-correction term and the short-run coefficients. Under the PMG estimator, the short-run coefficients are allowed to differ across countries; the reported values therefore summarise the average short-run effects across the V4 economies. The estimated error correction coefficient (-0.906) is negative and highly significant, indicating adjustment towards the long-run equilibrium. Approximately 91% of short-run disequilibrium is corrected within one year, suggesting relatively rapid adjustment. The first differences of employment in industry (-0.519) and GDP growth (-1.021) both show negative coefficients; however, neither coefficient is statistically significant, suggesting no robust short-run effect.

In contrast, inflation has a negative and statistically significant short-run coefficient (-0.974), implying that increases in inflation are associated with lower FDI inflows in the short run. Trade openness, on the other hand, has a positive and statistically significant coefficient (0.595), indicating that higher openness is associated with higher FDI inflows in the short run. These short-run effects should again be interpreted as conditional associations rather than causal impacts.

3.2 Bounds Test for Cointegration

The country-specific Bounds tests reveal considerable heterogeneity across the V4 economies. Cointegration is confirmed only for the Czech Republic and Slovakia, whereas no long-run equilibrium relationship is identified for Hungary and Poland. The results show that the F-statistic equals 4.837 for the Czech Republic and 4.771 for Slovakia, while it is considerably lower for Hungary (2.384) and Poland (1.724). According to the critical values, the upper bound at the 5% significance level is 4.22 (for small samples), meaning that both the Czech Republic and Slovakia exceed this threshold. This confirms the presence of cointegration for these two countries.

In contrast, the F-statistics for Hungary and Poland fall below the lower bound value (3.06 at the 5% level), indicating no evidence of a long-run relationship in these cases. None of the country-specific results fall into the inconclusive region between the lower and upper bounds.

These findings should be considered when interpreting the pooled Panel ARDL estimates. Although the PMG estimator assumes homogeneous long-run coefficients, the Bounds tests suggest differences across countries. Moreover, the V4 economies were affected by common shocks such as EU accession, the global financial crisis, COVID-19, and the energy shock, which are not explicitly controlled for. Therefore, the pooled coefficients may partly reflect these unmodelled effects.

4 Conclusion

This part of the article evaluates the findings and statistical analysis. The results indicate clear differences between short-run and long-run relationships. In the long run, GDP and employment in industry positively affect FDI, whereas inflation is not statistically significant. In contrast, trade openness exhibits a negative long-run effect. In the short run, inflation discourages FDI, while trade openness has a positive impact. The negative and significant error correction term confirms the existence of a stable long-run equilibrium.

The country-specific Bounds tests reveal heterogeneous results across the V4 countries. Cointegration is confirmed only for the Czech Republic and Slovakia, suggesting that long-run relationships differ among individual economies and that investment policies should consider country-specific conditions.

The main contribution of this study is empirical. The paper updates evidence on the selected macroeconomic correlates of FDI in the V4 countries for the period 2000-2025. In addition, the study complements pooled estimation with country-specific Bounds tests, providing evidence of heterogeneity within the Visegrad Group. The contribution should not be understood as methodological novelty, but as updated and carefully qualified regional evidence.

The results should be interpreted with caution due to the small panel of four countries and annual observations, which limits the reliability of unit-root tests, PMG/MG estimation, and Bounds testing. The model may suffer from omitted-variable bias, as taxation, wages, exchange rates, institutional quality, political stability, and investment incentives are not included. Potential endogeneity is also possible, since FDI may influence GDP growth, industrial employment, and trade openness. Comparability may be affected by the use of SPE-adjusted Hungarian FDI data, and major common shocks are not explicitly modelled. Therefore, the coefficients should be interpreted as conditional statistical associations rather than causal effects.

Future research could incorporate additional institutional and structural variables, include explicit controls for common shocks or time effects, apply alternative dynamic panel methods.

Acknowledgements

This research was supported by the Ministry of Education, Youth and Sports Czech Republic within the Institutional Support for Long-term Development of a Research Organization in 2026.

References

Bobenič Hintošová A, Bruothová M, Kubíková Z, Ručinský R (2018) Determinants of foreign direct investment inflows: A case of the Visegrad countries. Journal of International Studies 11(2): 222–235. https://doi.org/10.14254/2071-8330.2018/11-2/15

Bruno R L, Campos N F, Estrin S (2021) The effect on foreign direct investment of membership in the European Union. Journal of Common Market Studies 59(4): 802–821. https://doi.org/10.1111/jcms.13131

Dondashe N, Phiri A (2018) Determinants of FDI in South Africa: Do macroeconomic variables matter? MPRA Paper No 83636. https://mpra.ub.uni-muenchen.de/83636/

Erdogan M, Unver M (2015) Determinants of foreign direct investments: Dynamic panel data evidence. International Journal of Economics and Finance 7(5): 82–95. https://doi.org/10.5539/ijef.v7n5p82

Gorbunova Y, Infante D, Smirnova J (2012) New evidence on FDI determinants: An appraisal over the transition period. Prague Economic Papers 21(2): 129–149. https://doi.org/10.18267/j.pep.415

Hoang H, Bui D (2015) Determinants of foreign direct investment in ASEAN: A panel approach. Management Science Letters 5: 213–222. https://doi.org/10.5267/j.msl.2014.12.015

Louail B (2019) Determinants of foreign direct investment in Arab countries during 1970–2016. International Journal of Advanced and Applied Sciences 6(3): 102–110. https://doi.org/10.21833/ijaas.2019.03.015

Mohammad K, Sapuan N M, Mohammed A F (2022) The impact of political stability and macroeconomic variables on foreign direct investment in Turkey. In Innovation of businesses and digitalization during Covid-19 pandemic Cham: Springer https://doi.org/10.1007/978-3-031-08090-6_31

National Bank of Hungary (2025) Statistical Data Available at https://statisztika.mnb.hu/timeseries

Pedroni P (2004) Panel cointegration: Asymptotic and finite sample properties. Econometric Theory 20(3): 597–625. https://doi.org/10.1017/S0266466604203073

Pesaran M H, Shin Y, Smith R J (2001) Bounds testing approaches to the analysis of level relationships. Journal of Applied Econometrics 16(3): 289-326. https://doi.org/10.1002/jae.616

Rachdi H, Brahim M, Guesmi K (2016) Determinants of foreign direct investment: The case of emerging markets. Journal of Applied Business Research 32(4): 1033–1040. https://doi.org/10.19030/jabr.v32i4.9719

Shaari M S et al (2023) Determinants of foreign direct investment in ASEAN+3 countries: The role of environmental degradation. International Journal of Environmental Research and Public Health 20(3): 1720. https://doi.org/10.3390/ijerph20031720

World Bank (2025) World Bank Data Catalog Available at https://databank.worldbank.org/reports.aspx

Econometric Modelling of Tourism Demand in Collective Accommodation Facilities: Evidence from Czech Tourism Data

Radmila Krkošková[1], Zuzana Neničková[2]

Abstract. Tourism demand in collective accommodation facilities represents an important indicator of economic activity and regional development. The aim of this paper is to analyse the determinants and dynamics of tourism demand in the Czech Republic using an econometric approach. The analysis is based on quarterly data on the number of guests and overnight stays in collective accommodation establishments obtained from the Czech Statistical Office for the period 2010–2024.
To examine both short-term dynamics and long-term relationships, the Autoregressive Distributed Lag (ARDL) model is employed. The model specifies tourism demand as a function of its own lags and selected macroeconomic variables, including gross domestic product, price level, and international tourism flows. The ARDL framework enables the identification of cointegration relationships and the estimation of both long-run equilibrium effects and short-run adjustments through an error correction mechanism.
The results indicate the existence of a stable long-run relationship between tourism demand and macroeconomic factors, while short-term dynamics are significantly influenced by seasonal fluctuations and external shocks. The findings provide valuable insights into the behaviour of tourism demand in the accommodation sector and can support decision-making in tourism policy and regional development.

Keywords: ARDL model, Macroeconomic factors, Time series analysis, Tourism

JEL Classification: L83, C22, C51

1 Introduction

Tourism plays a significant role in economic development, particularly through its impact on regional economies and service sectors. The accommodation industry represents a key component of tourism infrastructure, reflecting both domestic and international travel demand. Understanding the determinants and dynamics of tourism demand is therefore essential for effective policy-making and business decision-making.

The aim of this paper is to analyse tourism demand in collective accommodation facilities in the Czech Republic using econometric modelling. The study focuses on identifying both short-term dynamics and long-term relationships between tourism demand and selected macroeconomic indicators.

2 Literature Review and Variables

2.1 Literature Review

Tourism demand modelling has been widely addressed in the literature using various econometric approaches, including time-series models, panel data analysis, and cointegration techniques. Traditional determinants of tourism demand include income (GDP), prices, exchange rates, and seasonality.

The study (Křečková et al., 2024) focuses on how macroeconomic variables – nominal exchange rate, inflation rate, GDP per capita and employee compensation – affect tourism demand, measured by the number of overnight stays in hotels and similar accommodation facilities. The aim is to provide actionable insights for destination managers and tourism researchers. The research covers the period 2000 - 2022 and the following countries: Czech

[1] School of Business Administration in Karviná, Silesian University in Opava, Department of Informatics and Mathematics, Univerzitní náměstí 1934/3, 733 40 Karviná, Czech Republic, e-mail: krkoskova@opf.slu.cz
[2] School of Business Administration in Karviná, Silesian University in Opava, Department of Informatics and Mathematics, Univerzitní náměstí 1934/3, 733 40 Karviná, Czech Republic, e-mail: nenickova@opf.slu.cz

Republic, Slovakia, Germany, Austria and Poland. The results confirm the importance of macroeconomic conditions for tourism. Higher GDP per capita in countries of origin generally supports tourism demand, while adverse changes in real or nominal exchange rates and rising inflation can reduce the number of overnight stays; sensitivity varies by country of origin of tourists. It offers empirical support for considerations on pricing policy, wage developments and exchange rate policy in the context of the competitiveness of Czech tourism.

According to (Lim, 1999), the article aims to examine a wide range of empirical studies focused on modelling international tourism demand and to synthesise their findings based on key explanatory variables, namely income, transportation costs, and tourism prices. The study evaluates the significance and effect sizes of these variables across published research.

The article (Šálková, 2024) analyses the main determinants of tourism demand and their relevance for researchers, practitioners, and policymakers. It focuses on the impact of macroeconomic variables—such as GDP per capita, inflation, nominal exchange rate, and employee remuneration—on hotel overnight stays. Using a regression model based on the Gauss–Markov framework, the study examines panel data from Czechia, Slovakia, Germany, Austria, and Poland for the period 2000–2022. The results highlight that exchange rates, inflation, and income-related factors significantly influence tourism demand, reflecting changes in purchasing power and economic stability.

The study (Song et al., 2010) compares different measures of tourism demand, such as tourist arrivals and tourist expenditure, in econometric modelling and forecasting. Focusing on Hong Kong tourism from Australia, the UK, and the USA. Results show that arrivals are mainly driven by income and habit persistence, while expenditure depends strongly on relative tourism prices. The study also shows dependence on whether the goal is to predict arrivals or total receipts.

The paper (Jeřábek, 2019) examines the determinants of inbound tourism demand in the South Moravian Region of the Czech Republic, focusing on tourists from Poland, Slovakia, Germany, and Austria. The results show that tourism demand is positively influenced by income, while price and transport cost effects vary across countries. Some tourist groups are sensitive to price and transport changes, while others show no significant response.

Agag et al. (2024) show that tourism demand is very sensitive to crisis situations. During the COVID-19 pandemic, factors like risk perception and uncertainty strongly reduced people's willingness to travel. This supports the use of crisis or structural shock variables in tourism models.

From a macroeconomic point of view, Akarsu (2023) finds strong relationships between tourist arrivals, income, inflation, and exchange rates. This means that tourism demand depends both on economic growth and on price competitiveness. In a similar way, Reivan-Ortiz et al. (2023) show that exchange rate changes and economic uncertainty have a significant effect on international tourist arrivals, which highlights the importance of financial stability and external conditions.

In addition, Ridderstaat J. (2021) finds that household financial wealth is an important factor in tourism demand cycles, supporting the idea that tourism demand is strongly income-driven. Finally, Kumar P. and Ekka P. (2023) confirm that the COVID-19 pandemic had a very strong negative impact on tourism, which shows why it is important to include structural breaks in empirical models.

Most existing studies for Czech tourism either focus on selected source markets, panel settings, or pre-pandemic periods. The present paper contributes by applying a unified quarterly ARDL/ECM framework to aggregate tourism demand in collective accommodation establishments for the Czech Republic over a period that includes the pandemic disruption.

2.2 Variables

The empirical analysis is based on quarterly data covering the period 2010–2024. The data were obtained primarily from the Czech Statistical Office and include indicators related to tourism activity and macroeconomic development in the Czech Republic.

The dependent variable is tourism demand TD_t, measured by the number of overnight stays in collective accommodation establishments. This indicator is widely used in tourism demand research because it reflects the actual utilization of accommodation services and captures both domestic and international tourism activity.

Gross domestic product GDP_t is included as a proxy for overall economic performance and household purchasing power. Economic growth generally increases disposable income and consumers' ability to spend on leisure and travel activities. Therefore, a *positive* relationship between GDP and tourism demand is expected.

The consumer price index CPI_t is used as an indicator of the general price level and inflation. Higher prices reduce real purchasing power and may increase the cost of tourism-related services, including accommodation, transportation, and food. Consequently, a *negative* relationship between inflation and tourism demand is expected.

The variable international tourism INT_t represent the inflow of foreign visitors to the Czech Republic. International tourists constitute an important component of accommodation demand, particularly in major tourist destinations. Therefore, a *positive* relationship between international arrivals and tourism demand is anticipated. To capture the extraordinary effects of the COVID-19 pandemic, a dummy variable is included in the model. The variable takes the value of one during the pandemic disruption period and zero otherwise. The expected effect of this variable is negative because travel restrictions, lockdown measures, and reduced international mobility substantially affected tourism activity.

Tourism demand is characterized by strong seasonal fluctuations. To account for recurring quarterly patterns, seasonal dummy variables are incorporated into the model. Specifically, quarterly indicators for the second, third, and fourth quarters are included, while the first quarter serves as the reference category. The inclusion of seasonal dummies allows the separation of regular seasonal effects from the influence of macroeconomic determinants.

All variables are transformed into natural logarithms, except for dummy variables. The logarithmic specification helps stabilize variance, reduce potential heteroskedasticity, and allows estimated coefficients to be interpreted as elasticities. Consequently, the estimated parameters represent percentage changes in tourism demand associated with percentage changes in explanatory variables.

3 Methodology

To investigate the relationship between tourism demand and selected macroeconomic determinants, the study employs the Autoregressive Distributed Lag (ARDL) modelling framework. The ARDL approach is particularly suitable for analysing economic time series because it allows the simultaneous estimation of short-run dynamics and long-run equilibrium relationships within a single model.

The choice of the ARDL methodology is motivated by several advantages. First, ARDL models can be applied when explanatory variables are integrated of order zero, I(0), or order one, I(1), provided that none of the variables are integrated of order two, I(2). Second, the method performs well with relatively small samples, which is particularly relevant for quarterly macroeconomic data. Third, the framework enables the identification of both long-run equilibrium effects and short-run adjustment processes through the Error Correction Model (ECM) representation.

The general ARDL specification can be expressed as:

$$lnTD_t = \alpha + \sum_{i=1}^{p} \phi_i lnTD_{t-i} + \sum_{j=0}^{q_1} \beta_j lnGDP_{t-j} + \sum_{k=0}^{q_2} \gamma_k lnCPI_{t-k} + \sum_{m=0}^{q_3} \eta_m lnINT_{t-m} + \\ + \theta COVID_t + \sum_{s=2}^{4} \lambda_s Q_s + \varepsilon_t, \quad (1)$$

where TD_t denotes tourism demand, GDP_t represents gross domestic product, CPI_t denotes the consumer price index, INT_t refers to international tourist arrivals, $COVID_t$ is the pandemic dummy variable, Q_s are seasonal dummy variables, and ε_t is the stochastic error term.

After establishing the existence of a long-run relationship, the ARDL model can be reformulated into an Error Correction Model (ECM):

$$\Delta lnTD_t = \alpha + \sum_{i=1}^{p-1} \varrho_i \Delta lnTD_{t-i} + \sum_{j=0}^{q_1-1} \beta_j^* \Delta lnGDP_{t-j} + \sum_{k=0}^{q_2-1} \gamma_k^* \Delta lnCPI_{t-k} + \sum_{m=0}^{q_3-1} \eta_m^* \Delta lnINT_{t-m} + \\ + \psi \Delta COVID_t + \sum_{s=2}^{4} \omega_s \Delta Q_s + \varphi ECT_{t-1} + u_t, \quad (2)$$

where ECT_{t-1} represents the error correction term and φ measures the speed at which short-run deviations from equilibrium are corrected. A negative and statistically significant coefficient of the error correction term indicates convergence towards long-run equilibrium.

Before estimating the model, the stationarity properties of the variables are examined using the Augmented Dickey–Fuller (ADF) unit root test. Subsequently, the ARDL bounds-testing procedure is employed to investigate the existence of cointegration among the variables. Diagnostic tests are then performed to assess the adequacy of the estimated specification.

Although alternative approaches such as Vector Autoregressive (VAR) models, Vector Error Correction Models (VECM), panel cointegration methods, and machine-learning forecasting techniques have been applied in tourism demand research, the ARDL framework was selected because of its flexibility regarding integration orders, its suitability for relatively small samples, and its ability to distinguish between short-run and long-run effects within a unified modelling framework.

4 Data Analysis

Before estimating the ARDL model, the Augmented Dickey–Fuller (ADF) test was applied to examine the stationarity properties of the variables. Testing for unit roots is necessary to avoid spurious regression results and to verify whether the ARDL methodology is appropriate.

Table 1 shows that all variables become stationary after first differencing, meaning that the variables are integrated of order one, I (1). The second column provides information on the model type of testing the unit root (*n* = no trend and level constants /*c* = constant /*c*+*t* = level constant and trend), the third column contains the calculated T-statistics; the following column contains the corresponding level of statistical significance. The last column includes the result of testing: N = non-stationary (H0 not rejected), S = stationary (H0 rejected).

Since none of the variables are integrated of order two, the ARDL bounds testing approach can be applied.

Variable	*n/c/c+t*	*T*-**stat**	**Prob.**	**Result**	**Variable**	*n/c/c+t*	*T*-**stat**	**Prob.**	**Result**
TD	c	−0.578	0.781	N	D(*TD*)	c	−4.951	0.001***	S
GDP	c	−1.465	0.621	N	D(*GDP*)	c	−4.654	0.001***	S
CPI	c+t	−1.566	0.592	N	D(*CPI*)	c+t	−5.815	0.002***	S
INT	n	−0.624	0.803	N	D(*INT*)	n	−4.233	0.037**	S

Statistical significance at the 0.01 level (***), at the 0.05 level (**), at the 0.1 level (*)

Table 1 ADF test

The optimal lag structure was selected using the Akaike Information Criterion (AIC). Based on the information criteria, the ARDL (2,1,1,1) specification was selected as the most appropriate model.

TEST Statistic		**Value**
F - statistic		6.12
Critical Value Bounds	**I(0) Bound**	**I(1) Bound**
5% signifikance level	3.23	4.85
1% significance level	4.29	5.61

Source: CZSO (2025), own calculations.

Table 2 ARDL Bounds Test for Cointegration

To verify the existence of a long-run relationship between tourism demand and explanatory variables, the ARDL bounds test was performed. Since the calculated F-statistic (6.12) exceeds the upper critical bound (4.85), the null hypothesis of no cointegration is rejected. This confirms the existence of a stable long-run relationship between tourism demand and the selected macroeconomic variables.

Table 3 shows long-run coefficients and Table 4 shows short-run relationships.

Variable	Coefficient	Std. Error	t-Statistic	Significance
lnGDP	0.88	0.21	4.19	0.000***
lnCPI	–0.54	0.23	–2.35	0.026**
lnINT	0.76	0.18	4.22	0.000***
COVID	–0.85	0.22	–3.86	0.000***

Statistical significance at the 0.01 level (***), at the 0.05 level (**), at the 0.1 level (*).
Source: CZSO (2025), own calculations.

Table 3 Long-Run Coefficients

The long-run results indicate that GDP has a positive and statistically significant effect on tourism demand. Specifically, a 1% increase in GDP leads to approximately a 0.88% increase in tourism demand, holding other variables constant. Inflation, represented by the CPI, negatively affects tourism demand. Higher price levels reduce consumers' purchasing power and increase tourism-related costs, which may discourage travel activity. International tourism arrivals positively influence tourism demand, confirming the importance of foreign visitors for accommodation services in the Czech Republic. The COVID dummy variable is negative and statistically significant, demonstrating the severe impact of the pandemic on tourism activity.

Variable	Coefficient	Std. Error	t-Statistic	Significance
ΔlnGDP	0.31	0.11	2.81	0.003***
ΔlnCPI	–0.19	0.09	–2.11	0.038**
ΔlnINT	0.42	0.12	3.5	0.000***
ΔCOVID	–0.52	0.22	–2.36	0.034**
ECM(-1)	–0.67	0.13	–5.15	0.000***

Statistical significance at the 0.01 level (***), at the 0.05 level (**), at the 0.1 level (*).
Source: CZSO (2025), own calculations.

Table 4 Error Correction Model Results

The ECM coefficient is negative and statistically significant, which confirms the adjustment process toward long-run equilibrium. The estimated coefficient of –0.67 suggests that approximately 67% of short-run deviations from equilibrium are corrected within one quarter. The short-run results are generally consistent with the long-run findings. Economic growth and international tourism arrivals positively influence tourism demand, while inflation and the COVID-19 pandemic have negative effects.

5 Conclusion

The aim of this paper was to analyze the determinants of tourism demand in collective accommodation establishments in the Czech Republic using an econometric modelling approach. Particular attention was devoted to the relationship between tourism demand and selected macroeconomic variables, including economic activity, price developments, international tourism flows, seasonality, and the impact of the COVID-19 pandemic.

The findings indicate that tourism demand is closely linked to macroeconomic conditions. Economic growth and international tourist arrivals appear to support tourism activity, while rising prices and pandemic-related disruptions are associated with lower demand for accommodation services. These relationships are consistent with theoretical expectations and with the conclusions reported in the tourism demand literature.

The study contributes to the existing literature by applying the ARDL framework to aggregate tourism demand in the Czech Republic over a period that includes both the COVID-19 pandemic and the subsequent recovery phase. While previous studies have frequently focused on specific source markets, regional analyses, or pre-pandemic periods, the present research provides a broader perspective on the dynamics of tourism demand in collective accommodation establishments.

From a methodological perspective, the results support the usefulness of the ARDL approach for analyzing tourism demand under conditions characterized by structural disruptions and changing economic environments. The framework allows the simultaneous examination of long-run relationships and short-run adjustments, which is particularly relevant in tourism markets exposed to external shocks.

The findings may have practical implications for tourism policy and strategic planning. Understanding the role of economic growth, inflation, international mobility, and seasonal patterns may help policymakers and tourism stakeholders design measures that support the resilience and competitiveness of the accommodation sector during periods of economic uncertainty.

Acknowledgements

This research was supported by the Ministry of Education, Youth and Sports Czech Republic within the Institutional Support for Long-term Development of a Research Organization Silesian University in Opava in 2026.

References

Agag G, Abdelmoety Z H, Eid R (2024) Understanding the Factors Affecting Travel Avoidancebehavior During the COVID-19 Pandemic: Findings from a Mixed Method Approach. Journal of Travel Research 63(4):858-882. https://doi.org/10.1177/00472875231182110

Akarsu G (2023) Interrelationships between Tourist Arrivals, Exchange Rate, Inflation, and Economic Growth: Empirical Evidence for Türkiye. World Journal of Applied Economics 9(1):49-76. https://doi.org/10.22440/wjae.9.1.3

CZSO (2025) Czech Statistical Office. Available at https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=metodika-uvod

Jeřábek T (2019) The Effects of Selected Macroeconomic Variables on Tourism Demand for the South Moravian Region of the Czech Republic from Germany, Poland, Austria, and Slovakia, Comparative Economic Research. Central and Eastern Europe 22(3):25-43. https://doi.org/10.2478/cer-2019-0021

Křečková R, Šálková D, Procházková R, Yekimov S (2024) Macroeconomics and Tourism Demand: Evaluating the Role of Economic Indicators in the Czech Republic's Hospitality Industry. In: Proceedings of Economics and Finance Conferences. International Institute of Social and Economic Sciences, University of Hradec Králové, pp 198-214. https://doi.org/0.36689/UHK/HED/2024-01-018

Kumar P, Ekka P (2023) Statistical analysis of the impacts of COVID-19 pandemic on the small and large-scale tourism sectors in developing countries. Environment, Development and Sustainability 26:9625–9659. https://doi.org/10.1007/s10668-023-03112-4

Lim Ch (1999) A Meta-Analytic Review of International Tourism Demand. Journal of Travel Research, 37(3):273–284. https://doi.org/10.1177/004728759903700309

Reivan-Ortiz G G, Cong P T, Wong W K, Ali A, Thu H T T, Akhter S (2023) Role ofgeopolitical risk, currency fluctuation, and economic policy on tourist arrivals: temporal analysis of BRICS economies. Environmental Science and Pollution Research 30(32):78339–78352. https://doi.org/10.1007/s11356-023-27736-1

Ridderstaat J (2021) Households' Net Financial Wealth as a Determinant of Tourism Demand Cycles: Evidence from US Travel to Selected Caribbean Destinations. Journal of Travel Research 60(3):564-582. https://doi.org/10.1177/0047287520925174

Song H Li G, Witt S F, Fei B (2010) Tourism Demand Modelling and Forecasting: How Should Demand Be Measured? Tourism Economics 16(1):63-81. https://doi.org/10.5367/000000010790872213

Šálková D (2024) Impact of Selected Macroeconomics Variables on Tourism Demand in the Czech Republic, Hradec Economic Days . https://doi.org/0.36689/UHK/HED/2024-01-018

Turan T, Karakas M, Ozer H A (2020) How do oil price changes affect the current account balance? Evidence from co-integration and causality tests. Journal of Economics 68(1):55-68

Performance Dynamics of Czech Breweries: An Updated DEA-Based Assessment

Martina Kuncová[1], Roman Fiala[2]

Abstract. The Czech brewing sector has experienced considerable structural and economic shifts in recent years, significantly influenced by the COVID-19 pandemic. This paper investigates the efficiency development of Czech breweries over the period 2018–2023 using Data Envelopment Analysis (DEA). Expanding on earlier studies focused on prior timeframes, the research offers a contemporary perspective on technical efficiency within the industry.
The analysis is based on a dataset combining financial and operational indicators from a sample of Czech breweries. Input-oriented DEA models assuming both constant and variable returns to scale are applied to assess relative efficiency levels and identify performance patterns. The results reveal evolving efficiency trends, highlight persistent inefficiencies, and capture the sector's response to recent external shocks.

Keywords: brewing industry, Czech breweries; efficiency analysis; DEA; technical efficiency; financial performance

JEL Classification: C44, C61, C67

1 Introduction

Maintaining a stable market position is a key challenge for business entities operating in competitive environments. Companies typically pursue long-term objectives aimed at maintaining competitiveness and improving performance. It is important to emphasize that the primary goal of firms is the maximization of market value while maintaining competitiveness and preventing a decline in business efficiency. Due to international economic integration after the Second World War, gradual liberalization, and technological progress in information technologies, companies have been exposed to increasing competition (Knight, 2000). Efficient business performance therefore requires continuous analysis of business activities and possible adjustments of managerial decisions according to market conditions in order to maintain competitiveness and efficiency. Efficiency evaluation is an essential managerial tool for identifying inefficient firms, detecting operational reserves, and supporting managerial decision-making. In competitive markets, higher efficiency contributes to stronger competitiveness and improved financial performance (Jablonský and Dlouhý, 2015). Increasing efficiency therefore contributes not only to stronger competitiveness but also to improved financial results and profitability. This applies to all industries, including the brewing industry, which is the focus of this study.

The brewing industry represents one of the traditional and economically important sectors of the Czech economy. Beer production in the Czech lands has a long historical tradition dating back to the Middle Ages, with the first documented references to brewing activities dating back to the year 993 (Kozák and Kozáková, 2013; Novák Večerníček, 2015). The Czech Republic is among the world leaders in beer consumption per capita and is also one of the major beer exporters within the European Union. Beer exports from the Czech Republic increased from approximately 3.7 million hectolitres in 2007 to 5.3 million hectolitres in 2021, representing an increase of more than 43%. Total production of beer in the Czech Republic has slightly decreased since its high in 2018, but it is still above 20 million hectoliters – see Figure 1 (Czech Beer and Malt Association, 2025). In recent decades, however, the structure of the Czech brewing market has undergone substantial changes. Alongside several dominant industrial breweries, the market has experienced rapid growth in the number of small breweries and microbreweries. According to Tripes (2019), the number of microbreweries in the Czech Republic exceeded 450 in 2019, while the market continued to be dominated by several large industrial breweries. The number of breweries was close to 600 in 2021, then a small decrease was evident (Czech Beer and Malt Association, 2025). While large breweries benefit from economies of scale and strong market positions, smaller breweries often compete through product differentiation, local branding, and flexibility in responding to consumer preferences. Several Czech beer brands are internationally recognized, and the designation "Czech Beer" is protected as a geographical indication within the European Union under Commission Regulation No. 1014/2008.

[1] College of Polytechnics Jihlava, Department of Economic Studies, Tolstého 16, 58601 Jihlava, kuncova@vspj.cz.
[2] College of Polytechnics Jihlava, Department of Economic Studies, Tolstého 16, 58601 Jihlava, roman.fiala@vspj.cz.

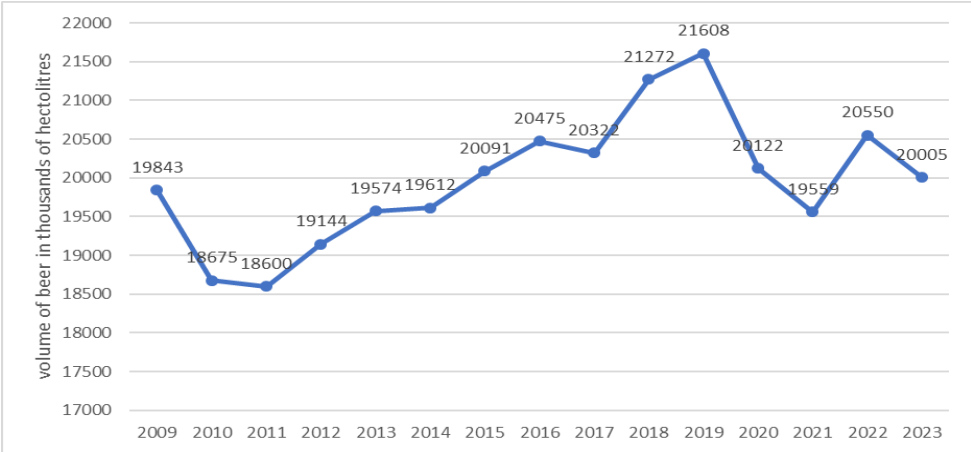


Figure 1 Beer production in the Czech Republic (Czech Beer and Malt Association, 2025)

At the same time, the brewing sector has been exposed to several external shocks in recent years, including the COVID-19 pandemic, increasing energy prices, inflationary pressures, changes in consumer behaviour, and the growing popularity of non-alcoholic beverages. Although the Czech brewing industry demonstrated relatively strong resilience during the pandemic period, the post-pandemic recovery has been accompanied by declining beer consumption and increasing economic uncertainty (Kuncová and Fiala, 2025). These developments created highly dynamic market conditions that may have significantly influenced the efficiency and performance of breweries.

Efficiency analysis allows the identification of inefficient firms, operational reserves, and potential opportunities for improvement. Data Envelopment Analysis (DEA) is one of the most frequently used methods for measuring the relative efficiency of homogeneous production units. DEA models have been widely applied in many industries, including banking, healthcare, tourism, hospitality, and agriculture. In the brewing industry, however, the number of studies remains relatively limited despite the economic and cultural importance of the sector.

Previous studies by the authors analysed Czech breweries before, during and shortly after the COVID-19 pandemic (Fiala and Kuncová, 2021; Kuncová and Fiala, 2023; Kuncová and Fiala, 2025). The present study extends this line of research by providing a unified longitudinal assessment covering the entire 2018–2023 period using DEA, specifically the CCR-I and BCC-I models. Unlike our previous studies, which focused on individual time periods, this paper presents a longitudinal DEA analysis covering the complete period from the pre-pandemic years through the COVID-19 disruption to the subsequent recovery (2018–2023). The contribution therefore lies in analysing efficiency dynamics under substantially different market conditions within a unified methodological framework, allowing changes in relative efficiency to be interpreted over time rather than comparing isolated years.

2 DEA models and the beer industry

The brewing industry is characterized by substantial heterogeneity in production scale, business models, and market orientation. Alongside large industrial breweries operating on international markets, the sector also includes numerous regional breweries and microbreweries that often focus on local production, product differentiation, and specialized customer segments. Measuring efficiency across such heterogeneous production units therefore requires analytical approaches capable of evaluating multiple inputs and outputs simultaneously while accounting for differences in operational scale. In this context, DEA has become one of the most frequently used methods. DEA models were originally introduced by Charnes, Cooper, and Rhodes (1978) for evaluating the relative efficiency of homogeneous decision-making units (DMUs). The method is based on the construction of an efficient frontier representing the best observed combinations of inputs and outputs within the analysed sample. Units located on the frontier are considered efficient, while inefficient units are evaluated according to their distance from the frontier. One of the main advantages of DEA models is their ability to handle multiple inputs and outputs simultaneously without requiring the specification of an explicit production function.

Let us denote $\boldsymbol{Y} = \left(y_{rj}, r = 1, \dots, R, j = 1, \dots n\right)$ a non-negative matrix of outputs and $\boldsymbol{X} = \left(x_{sj}, s = 1, \dots, S, j = 1, \dots n\right)$ a non-negative matrix of inputs. The efficiency score of the unit j_0 is calculated using the model:

$$
\begin{aligned}
\text{Maximise} \quad & U_{j_0} = \frac{\sum_{r=1}^{R} u_r y_{r,j_0}}{\sum_{s=1}^{S} v_s x_{s,j_0}} \\
\text{subject to} \quad & \frac{\sum_{r=1}^{R} u_r y_{r,j}}{\sum_{s=1}^{S} v_s x_{s,j}} \leq 1, \quad j = 1, \dots, n, \\
& u_r \geq \varepsilon, \quad r = 1, \dots, R, \\
& v_s \geq \varepsilon, \quad s = 1, \dots, S,
\end{aligned} \tag{1}
$$

where u_r is a positive weight of the r-th output, v_s is a positive weight of the s-th input, and ε is an infinitesimal constant, U_{j_0} is called the efficiency score for j_0-th unit under evaluation. The resulting efficiency score U_{j_0} equals 1 for efficient units and takes values lower than 1 for inefficient units. This study applies input-oriented DEA models because breweries generally have greater control over their inputs than over market demand and sales volumes. The first applied model is the CCR model proposed by Charnes et al. (1978), which assumes constant returns to scale (CRS). The input-oriented CCR model can be expressed in the following linearized form:

$$
\begin{aligned}
\text{Maximise} \quad & U_{j_0} = \sum_{r=1}^{R} u_r y_{r,j_0} \\
\text{subject to} \quad & \sum_{s=1}^{S} v_s x_{s,j_0} = 1, \\
& \sum_{r=1}^{R} u_r y_{r,j} - \sum_{s=1}^{S} v_s x_{s,j} \leq 0, \quad j = 1, \dots, n, \\
& u_r \geq \varepsilon, \quad r = 1, \dots, R, \\
& v_s \geq \varepsilon, \quad s = 1, \dots, S.
\end{aligned} \tag{2}
$$

Under the CRS assumption, proportional changes in inputs result in proportional changes in outputs. However, in real economic environments this assumption may not always hold, particularly in industries characterized by substantial differences in firm size. Therefore, the analysis also applies the BCC model introduced by Banker, Charnes, and Cooper (1984), which assumes variable returns to scale (VRS). The BCC model allows the separation of pure technical efficiency from scale efficiency and is formulated as follows:

$$
\begin{aligned}
\text{Maximise} \quad & U_{j_0}^{V} = \sum_{r=1}^{R} u_r y_{r,j_0} + \mu \\
\text{subject to} \quad & \sum_{s=1}^{S} v_s x_{s,j_0} = 1, \\
& \sum_{r=1}^{R} u_r y_{r,j} - \sum_{s=1}^{S} v_s x_{s,j} + \mu \leq 0, \quad j = 1, \dots, n, \\
& u_r \geq \varepsilon, \quad r = 1, \dots, R, \\
& v_s \geq \varepsilon, \quad s = 1, \dots, S,
\end{aligned} \tag{3}
$$

Based on a review of existing studies (Charnes et al., 1978; Charnes et al., 1985; Cook and Seiford; 2009), different types of efficiencies could be used depending on the inputs and outputs: for example, technical, production, cost or profit efficiencies. Based on the available data, we have adopted the traditional approach, i.e. for the analysis of technical efficiency of breweries in terms of their ability to convert economic inputs into financial outputs.

Comparing the results of the CCR and BCC models enables the calculation of scale efficiency (SE), which indicates whether firms operate at an optimal scale. Scale efficiency is calculated as the ratio of the CCR efficiency score to the BCC efficiency score. Values close to one indicate operation near the optimal production scale, whereas lower values suggest the presence of scale inefficiencies.

In the brewing industry, DEA models have been applied in studies focused on technical, production, financial, and sustainability efficiency. Bernetti et al. (2020) analysed the efficiency of Italian microbreweries using input-oriented DEA models with labour, debt, and capital indicators as inputs and sales revenues as outputs. Sellers-Rubio (2018) examined advertising efficiency in the Spanish brewing industry using DEA models that incorporated advertising expenditures, labour, and capital variables. Ezan (2015) used a two-stage DEA approach to evaluate the efficiency of Turkish brewing enterprises by combining profitability and sales analyses.

Several studies have also focused specifically on Czech breweries. Kasem et al. (2015) analyzed Czech breweries using financial, environmental, and sustainability indicators within DEA models. Hřebíček et al. (2011)

demonstrated the applicability of DEA methods for evaluating sustainability and performance in Czech brewing companies. Kasem and Trenz (2020) proposed an automated sustainability assessment system for Czech breweries combining DEA analysis with ESG-related indicators. Previous studies by the authors (Fiala and Kuncová, 2021; Kuncová and Fiala, 2023; Kuncová and Fiala 2025) examined the relationship between brewery size and relative efficiency and suggested that although larger breweries tend to achieve slightly higher average efficiency, managerial and operational factors may play a more important role than firm size alone.

3 Data

The analysis focuses on breweries operating in the Czech Republic within the NACE classification 11 – Manufacture of beverages, specifically subgroup 1105 – Manufacture of beer. Financial and economic data were obtained from the Albertina CZ Gold Edition database (for 2023 from the MagnusWeb database). The analyzed period covers the years 2018–2023 and therefore includes three distinct market phases: the pre-pandemic period, the COVID-19 pandemic period, and the post-pandemic adjustment phase.

The total number of active breweries in the Czech Republic increased significantly during the last decade (see Figure 1), mainly due to the expansion of the craft beer segment. However, recent years have also been characterized by declining beer consumption, increasing operating costs, and changing consumer preferences, all of which may have influenced the economic performance and efficiency of breweries. The final dataset contains approximately 100 breweries for each analyzed year (see Table 1), depending on data availability. Only firms with complete information for all selected variables were included in the DEA analysis.

year	2018	2019	2020	2021	2022	2023
Number of companies analyzed	102	98	77	112	106	108

Table 1 Number of breweries analysed

The selection of input and output variables was based on previous DEA studies in the brewing industry (Bernetti et al., 2020; Sellers-Rubio, 2018; Kuncová and Fiala, 2025) and on data availability. The objective was to assess the technical efficiency of breweries in transforming economic and financial resources into financial performance. The selected variables reflect the ability of breweries to utilize labor and financial resources efficiently while generating revenues and profit under changing market conditions. The analyzed period covers substantially different economic conditions, including the pre-pandemic expansion (2019), the COVID-19 pandemic (2020-2021), and the subsequent recovery (2022-2023) characterized by declining beer consumption, inflationary pressures, and rising energy costs. These developments created substantially different economic conditions across the analyzed years and provide an appropriate framework for studying efficiency dynamics in the brewing industry.

For the years 2018–2022, three input variables and two output variables were used:

Inputs:

- personnel costs,
- foreign sources of financing (liabilities),
- equity.

Outputs:

- profit/loss for the accounting period,
- sales of products, goods, and services.

Due to incomplete availability of personnel cost data for 2023, only two inputs (foreign sources and equity) and two outputs were included in the DEA models for that year. As the DEA specification for 2023 differs slightly from previous years, the results for 2023 should primarily be interpreted as indicative of recent developments rather than directly comparable with the previous years. Standard radial DEA models generally assume non-negative input and output variables. Financial variables such as profit/loss may naturally take negative values. Several approaches for handling negative data have been proposed in the DEA literature, including data translation and translation-invariant formulations (Pastor and Ruiz, 2007). In the present study, the DEA models were estimated using Frontier Analyst 4.5.0, a specialized software package for DEA applications, which allows datasets containing negative profit/loss values to be processed directly. Consequently, observations with negative profit/loss values were retained in the analysis and no manual transformation of the original accounting data was performed.

4 Results

The efficiency analysis was performed using input-oriented DEA models assuming both constant returns to scale (CCR-I) and variable returns to scale (BCC-I). The models were estimated separately for each year. between 2018 and 2023 in order to capture changes in brewery efficiency under different market conditions. In addition to technical efficiency scores, scale efficiency (SE) was also calculated as the ratio of CCR and BCC efficiency scores. Table 2 presents the average efficiency scores obtained from the CCR-I and BCC-I models, Table 3 shows the percentage of fully efficient breweries in individual years. Although the DEA frontiers were estimated separately for each year, the analyses were based on comparable samples of Czech breweries and consistent model specifications. The results therefore provide a useful indication of changes in relative efficiency patterns within the industry over time, although they should not be interpreted as direct measures of productivity change.

Year	2018	2019	2020	2021	2022	2023*
Avg. CCR-I score	0.574	0.417	0.516	0.547	0.422	0.324
Avg. BCC-I score	0.698	0.664	0.668	0.632	0.633	0.525
Avg. SE score	0.832	0.621	0.774	0.867	0.674	0.628

Table 2 Results: Average scores (*for 2023, 2 inputs only)

% of efficient breweries / Year	2018	2019	2020	2021	2022	2023*
CCR-I	13%	11%	16%	16%	9%	8%
BCC-I	32%	27%	31%	29%	25%	25%

Table 3 Results: Percentage of efficient breweries (*for 2023, 2 inputs only)

The year 2019 provides a particularly interesting case. Despite being associated with high beer production, strong exports, and the highest number of active breweries in the Czech Republic, the average efficiency scores declined compared to 2018. One possible explanation is the rapid expansion of microbreweries and small breweries during this period. New market entrants may initially operate with lower efficiency due to limited production scale, investment costs, or unstable customer demand. The increasing number of breweries therefore likely contributed to greater heterogeneity and lower average efficiency (mainly SE) within the sector.

In contrast, the pandemic years 2020 and 2021 were associated with a moderate improvement in average efficiency and with a higher proportion of fully efficient breweries. Although the COVID-19 pandemic strongly disrupted hospitality services and on-trade beer consumption, the observed increase in efficiency may reflect several adjustment mechanisms within the sector. First, less efficient breweries may have exited the market or temporarily suspended operations during the pandemic period, resulting in a selection effect that increased the average efficiency of the remaining firms. Second, many breweries implemented cost-reduction measures and adapted their distribution channels toward retail sales, takeaway services, or online distribution. These adjustments may have improved the relationship between inputs and outputs despite declining overall beer consumption.

The post-pandemic year 2022 again shows a deterioration in efficiency indicators. This development corresponds with declining beer consumption, rising energy and production costs, inflationary pressures, and changing consumer preferences, including growing demand for non-alcoholic beverages. Because the DEA specification for 2023 differs from the preceding years, the results for 2023 should primarily be interpreted as a supplementary assessment of the most recent developments in the Czech brewing industry rather than as a fully comparable continuation of the 2018–2022 series. While the results provide useful information about recent developments in the Czech brewing industry, they are not fully comparable with the results obtained for the years 2018–2022.

Figure 2 illustrates the distribution of CCR-I efficiency scores across individual efficiency intervals. The results demonstrate considerable heterogeneity among breweries throughout the analyzed period. In particular, the share of highly inefficient breweries increased substantially in 2019 and also in 2023, while the proportion of breweries achieving efficiency scores above 0.8 decreased compared to earlier years. The analysis also identified several breweries that consistently achieved high efficiency scores throughout the analyzed period. Under the CCR-I specification, only two breweries remained fully efficient during all analyzed years: Gwern, s.r.o. and Mandala CZ, s.r.o., both belonging to the category of small breweries or microbreweries. Under the BCC-I specification, four breweries achieved efficiency in all analyzed years, including Budějovický Budvar, Heineken Czech

Republic, Plzeňský Prazdroj, and Pivovar Nymburk. These results indicate that long-term efficiency is achievable for both small specialized breweries and large industrial producers.

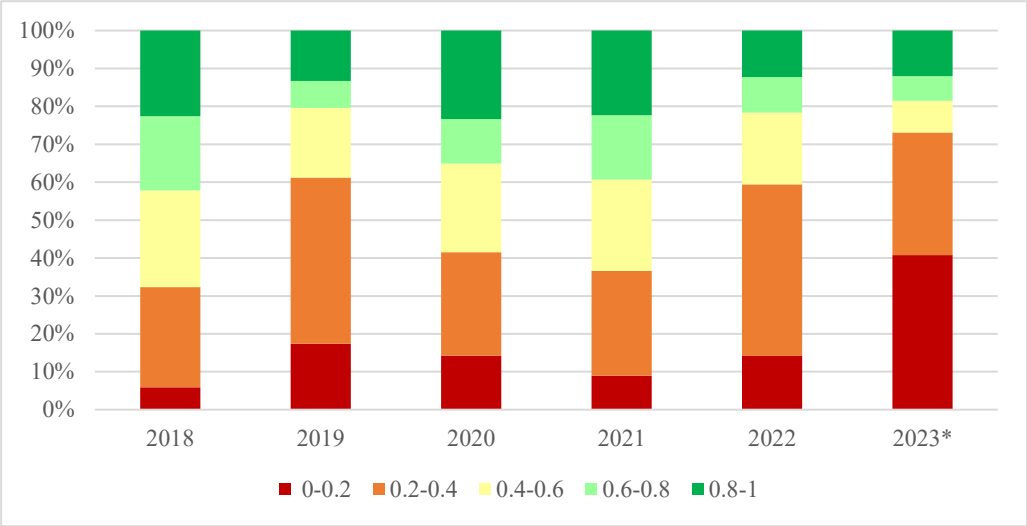


Figure 2 CCR-I scores

5 Conclusions

The main contribution of this paper lies in the evaluation of brewery efficiency across three substantially different economic periods using a DEA framework. The results indicate that the efficiency of Czech breweries changed considerably over time, coinciding with substantial changes in market conditions. The year 2019, despite high beer production and favorable market conditions, showed a decrease in average efficiency. In contrast, the pandemic years 2020 and 2021 were associated with moderate improvements in both average efficiency scores and the proportion of efficient breweries. Although the COVID-19 pandemic significantly disrupted the hospitality sector and on-trade beer sales, the observed efficiency patterns may suggest that many breweries adapted relatively successfully to changing conditions. The temporary increase in efficiency may reflect a combination of market selection effects, cost optimization measures, and adjustments in distribution channels. Less efficient breweries may have suspended operations or exited the market, while the remaining firms adapted their business models to retail and alternative sales channels. The post-pandemic year 2022 again demonstrated a deterioration in efficiency indicators. This development corresponds with declining beer consumption, rising energy and production costs, inflationary pressures, and changing consumer preferences, including increasing demand for non-alcoholic beverages. Several breweries remained fully efficient throughout the analyzed period under both CCR-I and BCC-I specifications, indicating that managerial and operational factors may play a more important role than firm size alone.

The study has several limitations. The analysis was based on available accounting data and therefore included only selected economic indicators. Physical production indicators (e.g. beer output, production capacity or production volume) were not available and therefore could not be included in the DEA models. In addition, personnel cost data were not fully available for 2023. For this reason, the discussion of long-term trends is based primarily on the results for 2018–2022, while the 2023 results should be viewed as supplementary evidence regarding the most recent developments in the sector. Furthermore, the dataset did not include all breweries operating in the Czech Republic, particularly smaller microbreweries with incomplete financial statements.

Future research could extend the analysis by incorporating additional financial indicators, productivity measures, or multi-criteria evaluation methods. A promising direction for further research is also the application of dynamic DEA approaches or Malmquist productivity indices that would allow a more detailed analysis of productivity changes and long-term efficiency development within the Czech brewing industry.

Acknowledgements

The paper was supported by the contribution of long-term institutional support of research activities by the College of Polytechnics Jihlava.

References

Banker R D, Charnes A, Cooper W W (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30(9): 1078-1092. http://dx.doi.org/10.1287/mnsc.30.9.1078

Bernetti I, Alampi Sottini V, Cipollaro M, Menghini S (2020) A survey on the performance of the Italian brewing companies. Italian Review of Agricultural Economics 75: 37–50. https://doi.org/10.13128/rea-11741

Charnes A, Cooper W W, Rhodes E (1978) Measuring the efficiency of decision making units. European Journal of Operational Research 2: 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

Czech Beer and Malt Association (2025) Basic facts. https://ceske-pivo.cz/zakladni-fakta

Ezan A I (2015) Efficiency Measurements in the Turkish Brewing Industry by Using Data Envelopment Analysis. https://kanazawa-u.repo.nii.ac.jp/record/35524/files/Full-H-1321072001-ezan-aykut-ilhan.pdf

Fiala R, Kuncová M (2021) The Czech brewing industry: Is there a relationship between firm efficiency and firm size? In*: 8th SWS International Scientific Conference on Social Sciences - ISCSS 2021, 8: pp* 185–192

Hřebíček J, Soukopová J, Štencl M, Trenz O (2011) Integration of Economic, Environmental, Social and Corporate Governance Performance and Reporting in Enterprises. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis 59: 157-177. https://doi.org/10.11118/actaun201159070157

Jablonský J, Dlouhý M (2015) Modely hodnocení efektivnosti a alokace zdrojů. Professional Publishing

Jiang B, Lio W, Li X (2019) An Uncertain DEA Model for Scale Efficiency Evaluation. IEEE Transactions on Fuzzy Systems 27:1616–1624. https://doi.org/10.1109/TFUZZ.2018.2883546

Kasem E, Trenz O, Hřebíček J, Faldík, O (2015) Key Sustainability Performance Indicator Analysis for Czech Breweries. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis 63: 1937–1944. https://doi.org/10.11118/actaun201563061937

Kasem O, Trenz O (2020) Automated Sustainability Assessment System for Small and Medium Enterprises Reporting. Sustainability 12: 1–23. https://doi.org/10.3390/su12145687

Karagiannis G, Sarris A (2005) Measuring and explaining scale efficiency with the parametric approach: the case of Greek tobacco growers. Agricultural Economics 33:441–451. https://doi.org/10.1111/j.1574-0864.2005.00084.x

Knight G. (2000) Entrepreneurship and Marketing Strategy: The SME Under Globalization. Journal of International Marketing 8(2): 12-32. https://doi.org/10.1509/jimk.8.2.12.19620

Kozák V, Kozáková V (2013) Změny v českém pivovarství na přelomu tisíciletí [Changes in the Czech Brewing Industry at the Turn of the Millennium]. VeRBuM, Prague

Kuncová M, Fiala R (2023) Analysis of the efficiency of Czech breweries in the years 2016-2020. In: 15th International Conference on Strategic Management and its Support by Information Systems 2023, Conference Proceedings. VSB Technical University of Ostrava, Ostrava, pp 209–218

Kuncová M, Fiala R (2025) Efficiency Dynamics in the Czech Brewing Industry: A Post-Pandemic Comparison of Small, Medium, and Large Breweries. In: Conference Proceedings. The 43rd International Conference on Mathematical Methods in Economics (MME 2025), The Czech Society for Operations Research, pp 198–205

Novák Večerníček J (2015) Dějiny piva: od zrození až po konec středověku [History of Beer: From Its Origins to the End of the Middle Ages], 2nd edn. CPress, Brno

Pastor J T, Ruiz J L (2007) Variables with Negative Values in DEA. In: Cook W D, Zhu J (eds) Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis. Springer, New York, pp 63–84. https://doi.org/10.1007/978-0-387-71607-7_4

Sellers-Rubio R (2018) Advertising efficiency in the Spanish beer industry: spending too much? International Journal of Wine Business Research 30(4), 410–427. https://doi.org/10.1108/IJWBR-11-2017-0066

Tripes S (2019) Leaders of Czech Craft Breweries and Their Ability to Attract the Customer. In: Mesquita A, Silva P (eds) Proceedings of the 15th European Conference on Management, Leadership and Governance. Academic Conferences and Publishing International Limited, Reading, pp 381–390

Metrics in Scrum: Quantitative Model for Evaluating Metrics Capability in Scrum Teams

Josef Kunhart [1], Jan Bartoška [2]

Abstract. Scrum is the leading agile method for software development and a suitable alternative to traditional project management. However, Scrum teams often struggle with metrics selection and utilization to evaluate and improve their efficiency. A gap exists between the metrics recommended in the literature and those that teams actually utilize in practice. This pilot study addresses this gap by proposing a quantitative metrics capability model that uses 21 key Scrum efficiency metrics identified in prior research. The model evaluates metrics capability across three categories: Sprint, Team, and Tech, and defines four levels of metrics maturity: low, medium, advanced, and high. We conducted a pilot study with 15 Scrum teams to refine our framework, using an online application called Metrics in Scrum. Empirical metric weights derived from team responses diverged substantially from literature-based weights. These findings reveal that teams prioritize different metrics than the literature suggests, and our refined model now reflects real-world practice more accurately. The tool is available online and enables Scrum teams to assess their current metrics maturity and identify areas for improvement.

Keywords: empirical validation, metrics capability model, metrics utilization, Scrum, Scrum metrics, team efficiency

JEL Classification: C93

1 Introduction

Scrum teams rely on metrics to assess their efficiency and maturity, yet a gap exists between metrics that literature mentions and those teams actually prioritize. Existing approaches to metrics measurement in agile projects often lack a systematic framework for evaluating how efficiently teams utilize these metrics. Many teams collect metrics, but few systematically evaluate how efficiently they are being used. We developed a quantitative model and software application for evaluating Scrum metrics capability and refined it with 15 Scrum teams. Our research addresses three research questions:

- RQ1: How do empirical metric weights differ from those described in literature?
- RQ2: How can empirical data refine our model for evaluating Scrum metrics capability?
- RQ3: What practical insights emerge from empirical findings for Scrum teams?

The research contributes a quantitative metrics capability model validated with 15 Scrum teams that reveals significant gaps between literature-based and empirically prioritized metrics.

2 Related work

Agile project and software development management is an innovative approach to project management that differs from traditional project management procedures (Shore, 2021). Scrum is currently widely used not only for software application development (Hobbs and Petit, 2017), but has also found applications in various fields beyond traditional project management, for example in research (Hron and Obwegeser, 2022). According to Kadenic et al. (2023), the most widely used agile methods are currently Scrum and hybrid methods based on Scrum, andsuccess in Scrum depends on the maturity of the team and the team composition. Strode et al. (2022) state that effective teamwork is crucial in agile software development. Research by Almeida and Carneiro (2023) suggests that metrics are key elements that provide valuable information about the efficiency and effectiveness of agile software development processes. Efficiency metrics translate qualitative observations into quantifiable evidence, enabling organizations to objectively evaluate the impact of agile adoption (Natarajan and Pichai, 2024). Verwijs and Russo (2023) found that continuous improvement is only possible when teams have the data to inspect their own work processes through collected metrics. Project management maturity models (PMMM) are

[1] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, kunhart@pef.czu.cz
[2] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, bartoska@pef.czu.cz

structured frameworks designed to evaluate, guide, and enhance organizational capability by defining maturity levels and assessing organizational practices against predefined benchmarks or standards (Al-Marri et al., 2025). Key examples include the Capability Maturity Model Integration (CMMI) (CMMI, 2026) and the Agile Maturity Model (AMM) (Chetankumar and Ramachandran, 2009). PMMM typically consist of several maturity levels, which represent different stages of organizational development and capability. As presented in table 1, both CMMI and AMM define five distinct maturity levels.

Level	CMMI	AMM
Level 1	Initial	Initial
Level 2	Managed	Just started
Level 3	Defined	Defined
Level 4	Quantitatively managed	Measured
Level 5	Optimizing	Optimal

Table 1 Project management maturity models (CMMI (2026); Chetankumar and Ramachandran (2009))

3 Data and methods

We developed a metrics capability model based on 21 key Scrum efficiency metrics and implemented it in an online application to collect empirical data from Scrum teams. We based our research on the list of metrics that affect Scrum team efficiency (table 2). The list derives from our (unpublished) prior research, in which we identified 61 metrics based on academic and professional literature. We narrowed the list to the 21 key metrics based on their weights, clustering, and a practical survey form length. According to Revilla and Höhne (2020), a reasonable survey scope has a completion time of 10–15 minutes. The table presents both original weights corresponding to the full 61-metric list and normalized weights adjusted to the 21 selected metrics.

Metric	Domain	Group	Original weight	Normalized weight
Velocity	Team	Performance	0.0971	0.1550
Cycle time	Team	Performance	0.0706	0.1127
Work capacity	Team	Performance	0.0530	0.0846
Focus factor	Team	Motivation	0.0353	0.0563
Targeted value increase	Team	Performance	0.0353	0.0563
Estimation accuracy	Sprint	Review/Demo	0.0259	0.0414
Sprint Burn-down	Artifacts	Sprint Backlog	0.0259	0.0414
Defect count	Tech	Quality	0.0259	0.0414
Defect count per User Story	Tech	Quality	0.0259	0.0414
Test automation percentage	Tech	Testing	0.0207	0.0331
Defect removal percentage	Tech	Quality	0.0207	0.0331
Defects severity	Tech	Quality	0.0207	0.0331
Source lines of code	Tech	Quality	0.0207	0.0331
Team size	Sprint	Planning	0.0207	0.0331
Number of accepted User Stories	Sprint	Review/Demo	0.0207	0.0331
Number of rejected User Stories	Sprint	Review/Demo	0.0207	0.0331
Team satisfaction score	Team	Motivation	0.0177	0.0283
Team member turnover	Team	Performance	0.0177	0.0283
Interruption time	Team	Performance	0.0177	0.0283
Team efficiency	Team	Performance	0.0177	0.0283
Number of User Stories in the Sprint	Artifacts	Sprint Backlog	0.0155	0.0246

Table 2 Metrics that affect Scrum team efficiency

To construct our metrics capability model, we defined three key parameters: Metric Maturity Level (MML), Metric Maturity Score (MMS), and Metrics Capability Level (MCL). MML indicates the extent to which a Scrum team uses a selected metric. We defined MML as an ordered set of tuples (formula 1).

$$mml \in \{(1, \text{not_used}), (2, \text{collected}), (3, \text{evaluated}), (4, \text{improved})\} \tag{1}$$

The set includes four maturity levels derived from project management maturity models: not used, collected, evaluated, and improved. We adapted the MML levels from CMMI and AMM maturity models and applied them to Scrum metrics. Consequently, the levels focus on metric capability and utilization rather than organizational management or processes. In the survey, MML ratings correspond to a four-point Likert scale with values of 1–4. MMS provides a percentual numerical value of a team's maturity in a selected area (formula 2).

$$mms \in [0, 100] \quad \text{(percentage)} \tag{2}$$

MMS can be calculated for all metrics or for a specific category (subset) of metrics. In both cases, MMS is calculated as the ratio of absolute and maximum scores (formula 3). In the formula, m indicates the number of metrics, i is the metric index, mml_i is the maturity level of the given metric, and w_i is the metric weight. We transform MML values to 0–3 by subtracting one, ensuring the minimum possible MMS is zero.

$$mms = \frac{\sum_{i=0}^{m}(mml_i - 1)w_i}{\sum_{i=0}^{m}(\max_j mml_{j,i} - 1)w_i} \tag{3}$$

We defined MCL as an ordered set of string values (formula 4), which correspond to low, medium, advanced, and high MCL.

$$mcl \in \{\text{low}, \text{medium}, \text{advanced}, \text{high}\} \text{ with low} < \text{medium} < \text{advanced} < \text{high} \tag{4}$$

MCL can be calculated for a specific category (subset) of metrics (mcl_c) or for all metrics (mcl_{total}). For specific categories, MCL maps the MMS value to one of four levels of MCL (formula 5).

$$mcl_c = \begin{cases} \text{high} & \text{if } 75 \leq mms_c \leq 100 \\ \text{advanced} & \text{if } 50 \leq mms_c < 75 \\ \text{medium} & \text{if } 25 \leq mms_c < 50 \\ \text{low} & \text{if } 0 \leq mms_c < 25 \end{cases} \tag{5}$$

For all metrics, we check both the MMS criteria and the minimum baseline across all categories. High total MCL requires 75–100% MMS and minimum category MCL of at least advanced. Advanced total MCL requires 50–<75% MMS and minimum category MCL of at least medium. Medium total MCL requires 25–<50% MMS.

$$mcl_{total} = \begin{cases} \text{high} & \text{if } 75 \leq mms_c \leq 100 \wedge \min_c(mcl_c) \geq \text{advanced} \\ \text{advanced} & \text{if } 50 \leq mms_c < 75 \wedge \min_c(mcl_c) \geq \text{medium} \\ \text{medium} & \text{if } 25 \leq mms_c < 50 \\ \text{low} & \text{otherwise} \end{cases} \tag{6}$$

We used Metrics in Scrum, a software application developed by one of the authors, to collect data (Kunhart, 2026). The application implements the proposed model with front-end and back-end components. The back-end component includes the database, administration interface, and application programming interface. The front-end includes the form page, results page, and research design page. It is fully responsive and localized to English and Czech. The survey form prioritizes simplicity and user-friendliness to maximize response rates and protect privacy. It includes an e-mail field, form fields for metrics listed in table 2, and related instructions. The form collects no personal data and anonymizes e-mail addresses using a one-way hash function, which serves as a unique identifier for each response. Upon submission, the application saves data to the database and redirects respondents to the results page displaying their total MML, total MMS, and MMS for individual categories. We disseminated the survey through local project management associations, professional social media platforms, agile-focused events, and professional networking contacts.

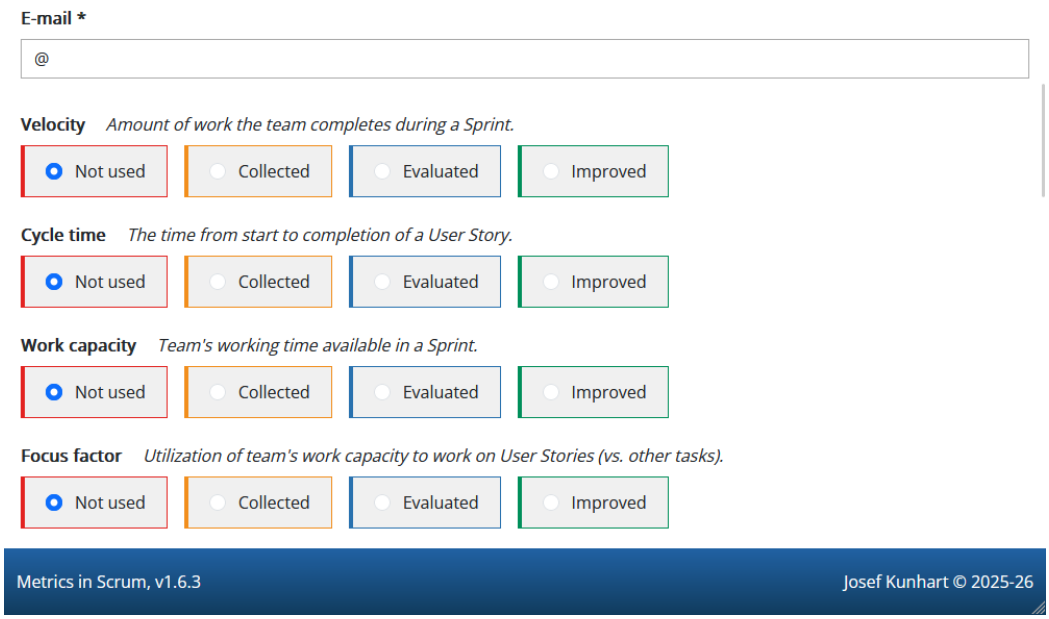


Figure 1 Metrics in Scrum

We designed the model to implement the proposed framework and provide a practical tool for Scrum teams. The model calculates MMS and MCL based on the MML of individual metrics and their respective weights, which respondents provided via survey form. We derived the initial weights of the metrics from our prior research that synthesized academic and professional literature. We categorized the metrics into three categories: Team, Sprint, and Tech, which correspond to the domains and groups in table 2. To maintain a balanced distribution of metrics across categories, we consolidated the original Sprint and Artifact domains into a single Sprint category. We designed the model to be adaptable, allowing for metric weight adjustments based on empirical data we collected from Scrum teams. We subsequently recalibrated the model weights accordingly to better reflect the practical importance of metrics in real-world practice.

4 Results

In the pilot study, we collected data from 15 Scrum teams, each consisting of approximately 5–9 members (Sutherland, 2014). Our data included MML ratings (1–4) for the 21 key metrics listed in table 2. We calculated empirical metric weights (w'_i) based on the collected data using formula 7.

$$w'_i = \frac{\sum_{j=0}^{r}(mml_{j,i} - 1)}{\sum_{i=0}^{m}\sum_{j=0}^{r}(mml_{j,i} - 1)} \tag{7}$$

In the formula, m indicates the number of metrics, r indicates the number of records, i is the metric index, j is the record index, $mml_{j,i}$ is the maturity level of metric i for record j. We subtracted one from the MML values to transform them to a 0–3 scale, to keep it consistent with the MMS calculation. Table 3 presents the metrics and metric categories along with normalized initial metric weights, the normalized empirical metric weights, and the percentage difference between them. Initial metric weights correspond to the normalized metric weights from prior research, while empirical metric weights are calculated based on the collected data. We answered the first research question by comparing the initial and empirical metric weights.

Metric	Category	Initial weight	Empirical weight	Difference (%)
Sprint Burn-down	Sprint	0.0414	0.0841	+103.14%
Velocity	Team	0.1550	0.0810	-47.74%
Work capacity	Team	0.0846	0.0717	-15.25%
Cycle time	Team	0.1127	0.0685	-39.22%
Team size	Sprint	0.0331	0.0623	+88.22%
Number of User Stories in the Sprint	Sprint	0.0246	0.0592	+140.65%
Estimation accuracy	Sprint	0.0414	0.0561	+35.51%
Defects severity	Tech	0.0331	0.0561	+69.49%
Number of accepted User Stories	Sprint	0.0331	0.0561	+69.49%
Team satisfaction score	Team	0.0283	0.0561	+98.23%
Defect count	Tech	0.0414	0.0498	+20.29%
Defect removal percentage	Tech	0.0331	0.0467	+41.08%
Test automation percentage	Tech	0.0331	0.0436	+31.72%
Number of rejected User Stories	Sprint	0.0331	0.0405	+22.35%
Defect count per User Story	Tech	0.0414	0.0343	-17.15%
Team efficiency	Team	0.0283	0.0343	+21.20%
Targeted value increase	Team	0.0563	0.0312	-44.58%
Focus factor	Team	0.0563	0.0312	-44.58%
Team member turnover	Team	0.0283	0.0187	-33.93%
Interruption time	Team	0.0283	0.0156	-44.88%
Source lines of code	Tech	0.0331	0.0031	-90.63%

Table 3 Initial and empirical metric weights

We calculated descriptive statistics to compare weight distributions (table 4). The identical means of 0.0476 reflect the mathematical constraint that normalized weights sum to 1.0 across 21 metrics. Despite this, empirical weights showed substantially lower variability (standard deviation: 0.0237 vs. 0.0324) and a more concentrated distribution with a lower maximum value.

Statistic	Initial weight	Empirical weight
Mean	0.0476	0.0476
Median	0.0331	0.0498
Standard deviation	0.0324	0.0237
Variance	0.0011	0.0006
Minimum	0.0246	0.0031
Maximum	0.1550	0.0841

Table 4 Statistical comparison of initial and empirical metric weights

Based on the empirical weights, we refined the model by updating the metric weights in the MMS calculation formula (formula 8).

$$mms' = \frac{\sum_{i=0}^{m}(mml_i - 1)w'_i}{\sum_{i=0}^{m}(\max_j mml_{j,i} - 1)w'_i} \tag{8}$$

The formula is an updated version of the original formula for calculating MMS (formula 3), where we replaced the initial metric weights with the empirical metric weights. Hence, m indicates the number of metrics, i is the metric index, mml_i is the maturity level of the given metric, and w'_i is the empirical metric weight. We answered the second research question by refining the model and updating the metric weights accordingly.

5 Discussion

We analyzed results from 15 Scrum teams, evaluating 21 agile metrics using a four point maturity level (MML) scale. We calculated empirical metric weights from the collected data and compared them against initial weights

derived from prior research. Notably, empirical weights diverged significantly from the academic and practical literature findings, directly addressing RQ1. The largest increases appeared in Number of User Stories in the Sprint (↑ 140.65%), Sprint Burn down (↑ 103.14%), and Team satisfaction score (↑ 98.23%), while the steepest declines affected Source lines of code (↓ 90.63%), Velocity (↓ 47.74%), and Interruption time (↓ 44.88%). This reordering reflects a qualitatively different conception of what matters in practice rather than merely a reweighting of the same priorities. Literature based weights prioritized Velocity (0.1550) as the dominant metric, but empirical data reduced it to 0.0810, while Number of User Stories surged from 0.0246 to 0.0592.

Empirical weighting compressed the distribution substantially. Variance decreased by 45% (from 0.0011 to 0.0006), the maximum weight fell from 0.1550 to 0.0841, and the minimum dropped from 0.0246 to 0.0031. This compression suggests that practicing teams treat metrics more holistically than literature suggests. However, this compression also masks a critical finding for RQ2. While the highest empirically weighted metrics are Sprint Burn-down (0.0841), Velocity (0.0810), Work capacity (0.0717), Cycle time (0.0685), and Team size (0.0623), the metrics with the largest magnitude of change differ from this ranking. Number of User Stories (↑ 140.65%) and Team Size (↑ 88.22%) show the most dramatic increases yet rank lower in final weights, whereas Velocity (↓ 47.74%) appears highly weighted despite losing nearly half its priority relative to literature. This suggests metric importance is context dependent. Teams may prioritize Sprint level tracking for immediate decision-making while reducing emphasis on Velocity as a strategic planning tool, a refinement that our refined MMS formula now captures.

Our model suggests a simpler, more practical approach to Scrum metrics. Rather than trying to implement all possible metrics, teams can concentrate on metrics that empirically matter in practice. The Metrics in Scrum application brings this model to life. A team completes a short survey and immediately receives their MMS and MML across three categories (Team, Sprint, Tech), along with a total MML. This addresses a real gap between what academic and professional literature recommends and how teams actually operate. The significant decline in Velocity's weight is particularly noteworthy, as it indicates that teams may be recognizing limitations of this traditionally dominant metric such as variance across Sprints or misalignment with delivery outcomes. Similarly, in our sample, the near elimination of Source lines of code reflects away from code volume toward quality and delivery outcomes. We answered RQ3 by deriving these practical insights from the empirical findings. Notably, a low category MML may signal a problematic area worth investigating, giving teams a diagnostic baseline grounded in empirical data.

This pilot study has several important limitations. The sample of 15 Scrum teams, while adequate for model refinement, may be too small to generalize our findings. We cannot yet determine whether these weight patterns would hold with other Scrum teams or whether they are specific to our sample. Our survey design emphasizes simplicity and brevity to maximize response rates, because response rates are notoriously low for these types of surveys. To improve participation, participants provided only email addresses, which we directly anonymized. We saved only the hashed e-mail address, insertion date, and language in which the respondent displayed the form. This approach protects respondent privacy but limits our ability to analyze how metrics may differ based on team characteristics or demographic factors. We may address this limitation by conducting case studies in specific agile teams. Finally, the MML ratings are self-reported rather than based on direct observation or objective measurement of actual metric utilization. Teams may perceive their metric maturity differently than external observers would assess it. Despite these limitations, the pilot study provides a valuable empirical foundation for refining the model and demonstrates its feasibility for larger scale validation.

6 Conclusion

This pilot study demonstrates that empirical data from practicing Scrum teams significantly reshapes how we understand metric priorities in agile environments. Our refined model contradicts several assumptions from the literature, particularly regarding Velocity's dominance and the irrelevance of code metrics. By grounding our study in how teams actually use metrics, we have created a more practical tool for assessing teams' metrics maturity. The Metrics in Scrum application enables teams to assess their current state and identify improvement areas aligned with how they actually work. In follow-up research, we will pursue three complementary directions. First, we will continue collecting data through the application to build a more comprehensive sample and refine the model iteratively. Second, we will conduct a detailed statistical analysis of individual metrics and examine how metric categories influence overall metric utilization patterns. Third, we will implement additional features, including self-adaptation based on empirical data and user feedback, as well as enhanced result-sharing capabilities.

References

Al-Marri R, Abdalla G, Mahdi E (2025) Project management maturity in project-based organizations: frameworks, drivers, and the role of sustainability. Future Business Journal 11: https://doi.org/10.1186/s43093-025-00670-z

Almeida F, Carneiro P (2023) Perceived Importance of Metrics for Agile Scrum Environments. Information 14:327. https://doi.org/10.3390/info14060327

Chetankumar P, Ramachandran M (2009) Agile Maturity Model (AMM): A software process improvement framework for agile software development practices. International Journal of Software Engineering 2(1):32–43.

CMMI Institute (2026) Capability Maturity Model Integration. Available from: https://cmmiinstitute.com/

Hobbs B, Petit Y (2017) Agile Methods on Large Projects in Large Organizations. Project Management Journal 48:3–19. https://doi.org/10.1177/875697281704800301

Hron M, Obwegeser N (2022) Why and how is Scrum being adapted in practice: A systematic review. Journal of Systems and Software 183:111110. https://doi.org/10.1016/j.jss.2021.111110

Kadenic MD, Koumaditis K, Junker-Jensen L (2023) Mastering Scrum with a focus on team maturity and key components of Scrum. Information and Software Technology 153:107079. https://doi.org/10.1016/j.infsof.2022.107079

Kunhart J (2026) Metrics in Scrum: Survey. Available from: https://metrics.kunhart.eu/

Natarajan T, Pichai S (2024) Behaviour-driven development and metrics framework for enhanced agile practices in Scrum teams. Information and Software Technology 170:107435. https://doi.org/10.1016/j.infsof.2024.107435

Revilla M, Höhne JK (2020) How long do respondents think online surveys should be? New evidence from two online panels in Germany. International Journal of Market Research 62:538–545. https://doi.org/10.1177/1470785320943049

Shore J (2021) The art of agile development, Second edition. O'Reilly, Beijing

Strode D, Dingsøyr T, Lindsjorn Y (2022) A teamwork effectiveness model for agile software development. Empirical Software Engineering 27: https://doi.org/10.1007/s10664-021-10115-0

Sutherland J (2014) Scrum, First edition. Currency, New York

Verwijs C, Russo D (2023) A Theory of Scrum Team Effectiveness. ACM Transactions on Software Engineering and Methodology 32:1–51. https://doi.org/10.1145/3571849

Ambulance Fleet Minimization without Loss of System Quality

Marek Kvet[1], Jaroslav Janáček[2]

Abstract. Most problems dealing with the design of emergency medical care systems are formulated in such a way that a limited number of ambulances is given and the goal is to deploy the ambulances at a set of possible locations such that the average response time of the system is minimal. In this paper, the reverse approach is studied. A desired limit on the average response time is set, and a minimal number of ambulances is sought to ensure that the average response time is below this limit. The presented computational study considers only those values of the time limit that lie between the current average response time and the time that can be achieved by optimally deploying the current number of ambulances.

Keywords: Prehospital urgent healthcare management, strategic decision, optimization, radial approach, generalized model

JEL Classification: C44, C61

1 Introduction

One of the key application domains where many sophisticated operations research techniques can frequently be applied is the creation of various service systems. We can now address a wide range of difficult problems that were previously unsolvable because of tremendous advancements in a variety of technical domains, from sophisticated software technologies to hardware. One such example is the design of service systems as a challenging combinatorial issue (Brotcorne et al., 2003; Chanta et al., 2014; Doerner et al., 2005; Ingolfsson et al., 2008; Karatas and Yakıcıa E, 2019). Let's take a closer look at service systems.

A service system often consists of a few components (people, buildings, equipment, and/or software programs) with a purpose, behavior (which may be characterized as a business process), and a structure. The idea that a service system can be thought of as a work system that generates a particular type of specified services is a simpler but more constrained description (Jánošíková, 2007; Jánošíková et al., 2021; Reuter-Oppermann et al., 2017). In order to provide the served population with related services like medical care, education, consultation, etc., a number of procedures are carried out by human participants and/or machines using the resources that are available.

Generally speaking, there are two major categories of service systems. *Public service systems* operate under different presumptions than *private service systems*, which are typically designed to maximize profits for their managers, founders, shareholders, and operators regardless of the number of serviced clients or equal access to given services. Regardless of financial gain or loss, the goal of public service systems is to ensure that all local citizens receive services. The law grants them existence. Emergency medical services (EMS), state administration, and several other systems are examples of such systems. From a scientific perspective, the public service system design problem is a member of the family of location problems that numerous researchers have examined and successfully resolved (Czimmermann and Kohání, 2018; Matiaško and Kvet, 2017; Revelle et al., 2008). In this research article, we focus specifically on emergency service systems as the main application domain.

Several facts must be taken into account while looking for an appropriate solution for these kinds of large location problems: The problems' combinatorial nature suggests the use of mathematical programming tools, which carry certain particularities. The scarcity of resources - human, environmental, and technical - is another significant factor. Additionally, the service must be concentrated in multiple service centers rather than being offered elsewhere. Naturally, a team may go from a service center to the emergency location, or the served consumers may travel to these centers. Therefore, finding the best service center network topology to maximize a given criterion is typically the problem of a public service system design. The weighted p-median problem formulation is one of the most widely used modeling approaches based on the previously given preliminaries.

[1] University of Žilina, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovakia, marek.kvet@fri.uniza.sk
[2] University of Žilina, Faculty of Management Science and Informatics, Univerzitná 8215/1, 010 26 Žilina, Slovakia, jaroslav.janacek@fri.uniza.sk

The solvability of the problem depends on the specific form of the objective function and the modeling method itself. The generally known location-allocation form can be replaced by so-called *radial approach*, which makes it possible to solve much larger problem instances (Avella et al., 2007; García et al., 2011; Janáček, 2008). If the used optimization criterion takes a min-sum form, then the problem is much easier to be solved than the case, the objective function of which takes a min-max form accompanied with some link-up constraints (Janáček and Kvet, 2016; Kvet, 2015).

The requirement to abstract too many aspects of a real system is one potential drawback of typical mathematical models used in the design of public service systems. The so-called *generalized disutility*, which has been developed to accommodate stochastic demands occurring in the planned system within the decision-making process, can partially solve this problem, as explained in (Jankovič, 2016; Kvet, 2014). With the advent of the concept of general disutility, it is necessary to consider the possibility of placing multiple ambulances at one location, which makes the model more complex. Another potential extension is to limit the number of vehicles at each location in the network by a lower and upper bound.

As mentioned in previous paragraphs, we concentrate here only on the main application domain of research, which consists in urgent pre-hospital healthcare management.

The Emergency Medical Service (EMS) was established to provide emergency pre-hospital care in severe situations where people's lives or health are in danger. Resuscitation and intensive care provide the greatest level of individualized patient care, and it is an essential and indispensable component of urgent care (Brotcorne et al., 2003; Ingolfsson et al., 2008).

The need for equal access to the service is typically taken into consideration when designing public service systems using standard quality criteria. They might also include specifications for the average time that each consumer can access the service. Time accessibility can be defined as either the maximum time required to reach the user in the worst situation or the average time required to reach the client. Time accessibility can also be defined as the need that every prospective client be contacted from the closest service center within a specific time frame (for instance, 15 minutes). For simplicity, in this article we will limit ourselves to only one criterion - the average travel time of the nearest ambulance crew from its station to the patient.

The main topic we are addressing takes into account the economic aspects of optimizing the rescue system. We will investigate the smallest number of ambulances with which the desired average response time can be achieved.

This paper is organized into four main sections. The theoretical section, which describes the optimization problem's mathematical context, comes after the introduction. The numerical experiment findings for the benchmarks under test are presented in the third section. The final section offers pertinent findings and a summary of the objectives that were accomplished.

2 Mathematical Background

2.1 Concept of generalized disutility

Almost all mathematical models that are appropriate for optimizing public service systems have one drawback: they are unable to adequately capture the complexity of the systems they represent. Because the weighted p-median model presupposes that supplying the demand occurred in some place always from the center assigned to it by solving the problem, deterministic models do not adequately grasp the actual method of handling the demands for service. Put another way, it is anticipated that the closest service center will always provide coverage for each user.

However, operational administration of actual service systems, particularly in the medical field, does not adhere to this premise of static assignment. It must be acknowledged that service requests are sporadic and typically unpredictable. The aforementioned suggests that the closest support center could not be accessible at the time of the request since another system user is being satisfied. In this situation, the request is sent to a different, nearby, and free service center. The demand fulfillment process typically involves r closest centers, where r is an integer. The probability that the k-th closest EMS station with the required equipment is the closest one is represented by q_k in the criterion. This idea greatly increases the model's realism. Figure 1 makes it easy to understand the primary idea.

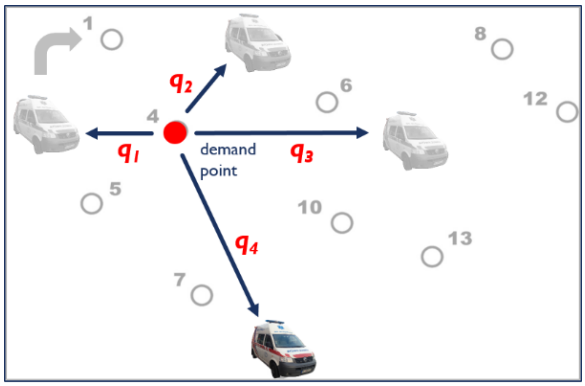


Figure 1 The idea of generalized disutility, in which the distances from clients to service providers are multiplied by weight coefficients q_k. The smallest distance is weighted by the largest coefficient.

2.2 Inputs of the model

The following inputs and variables are required to be introduced to fulfill the requirement of providing the readers with the mathematical background of the associated optimization problem. To maintain the highest level of clarity, we have summarized the symbols used in the following Table 1 and Table 2 respectively.

Symbol	Range	Meaning
I		finite set of places, at which an ambulance can be located
J		finite set of places of the serviced region, where demands for service are generated
r		maximal considered number of ambulances increasing time distance from the demand location (how many nearest ambulances can cover current request), in the example depicted in Figure 1, the value of r would be 4
K		range $\{1, .., r\}$
u_i	$i \in I$	upper bound of the number of ambulances, which can be placed at location $i \in I$
f_i	$i \in I$	lower bound of the number of ambulances, which can be placed at location $i \in I$
w_j	$j \in J$	frequency of demands for service generated at place $j \in J$
q_k	$k \in K$	probability that the k-th nearest ambulance is the first available for $k \in K$
V		finite set of radii used in the associated radial model of the location problem
e		time difference between a pair of neighboring radii
T^d		demanded average response time of the system
a_{ijv}	$i \in I$, $j \in J$, $v \in V$	gets its value from the set $\{0, 1\}$ and the value 1 means that possible ambulance location $i \in I$.is inside the radius $v \in V$ from the place $j \in J$.

Table 1 Used symbols and coefficients

Variable	Range	Values	Meaning
y_i	$i \in I$	N_0	number of ambulances, which should be located at the place $i \in I$
x_{jvk}	$j \in J$, $v \in V$, $k \in K$	$\{0, 1\}$	takes the value of 1 if the k-th closest ambulance is located outside the radius $v \in V$ from the demand location $j \in J$

Table 2 Summary of used decision variables

Having introduced all necessary input sets, symbols, coefficients and decision variables, we can continue by formulating the mathematical model.

2.3 Mathematical model

The mathematical model follows from the common weighted p-median problem formulation. It has been extended by the concept of generalized disutility with the possibility of locating more ambulances in the same location. The complete model takes the following form:

$$Minimize \quad \sum_{i \in I} y_i \tag{1}$$

$$Subject\ to \quad \sum_{k \in K} x_{jvk} + \sum_{i \in I} a_{ijv} y_i \geq r \quad for\ j \in J,\ v \in V \tag{2}$$

$$\frac{e \sum_{j \in J} w_j \sum_{k \in K} q_k \sum_{v \in V} x_{jvk}}{\sum_{j \in J} w_j} \leq T^d \tag{3}$$

$$y_i \geq f_i \quad for\ i \in I \tag{4}$$

$$y_i \leq u_i \quad for\ i \in I \tag{5}$$

$$y_i \in N_0 \quad for\ i \in I \tag{6}$$

$$x_{jvk} \in \{0, 1\} \quad for\ j \in J,\ v \in V,\ k \in K \tag{7}$$

The objective function (1) expresses the total number of deployed ambulances. The series of constraints (2) ensures the relation among the decisions on locating ambulances and the auxiliary decision variables x_{jvk}.The left hand side of inequality (3) expresses the average response time of the system and the constraint assures that the average response time is less than or equal to the demanded value T^d. The mathematical expression describing the average time to reach the request location from the closest available facility is based on the radial formulation, which is described in detail in (Janáček, 2008; Kvet, 2015). The series of constraints (4) ensures that the minimal number of ambulances will be located at the locations, where it is necessary. The series of constraints (5) limits the number of located ambulances at location i. The constraints (6) and (7) are obligatory constraints determining the feasible values of decision variables.

Model (1)-(7) is based on the radial approach, which has many advantages over the location-allocation approach. Its essence lies in replacing distances with their approximation, and instead of directly assigning customers to located service centers, in this case the zone (radius) in which the center closest to the customer is located is identified. The issue of the radial approach and its improvements is addressed in many publications, such as (Avella et al., 2007; García et al., 2011; Janáček, 2008).

3 Computational Experiments

3.1 Benchmarks and optimization environment

Road network data was used in this study's numerical experiments. In Slovakia's self-governing region of Žilina (ZA) and the Czech Republic's Moravsko-sliezský kraj (MSK), the Emergency Medical Service system's customized radial model, created to satisfy intervention needs, was utilized to improve system performance.

The associated mathematical model was solved by the optimization software FICO Xpress 8.3 (64-bit, release 2017) on a common computer equipped with the Intel® Core™ i7 8565U 1.80 GHz processor and 128 GB RAM.

3.2 Analysis of achieved results

Before running the model (1)-(7), it was necessary to analyze the interval, in which the demanded average response time T^d may move. The highest possible value was taken from the current stations' deployment, and it is further denoted by T^v. No optimization should worsen the measurable indicators of the current state. On the contrary, it is done to improve them. The lowest possible average response time T^{opt} achievable with current number p of located EMS stations was computed exactly by solving the following model (8)-(14). Special description and explanation of particular constraints is not necessary, most of them are identical to the previous model (1)-(7).

$$Minimize \quad \frac{e\sum_{j\in J} w_j \sum_{k\in K} q_k \sum_{v\in V} x_{jvk}}{\sum_{j\in J} w_j} \tag{8}$$

$$Subject\ to \quad \sum_{k\in K} x_{jvk} + \sum_{i\in I} a_{ijv} y_i \geq r \quad for\ j \in J,\ v \in V \tag{9}$$

$$\sum_{i\in I} y_i = p \tag{10}$$

$$y_i \geq f_i \quad for\ i \in I \tag{11}$$

$$y_i \leq u_i \quad for\ i \in I \tag{12}$$

$$y_i \in N_0 \quad for\ i \in I \tag{13}$$

$$x_{jvk} \in \{0, 1\} \quad for\ j \in J,\ v \in V,\ k \in K \tag{14}$$

The original number p of ambulances at disposal was 64 for the region MSK and 37 for the region ZA. The cardinalities of the candidate set I were 300 for the benchmark MSK and 300 for ZA. Mentioned set I consists of all cities and villages of the associated region. For the purpose of computational study, the set J was identical to the set I for all problem instances. The values of current and optimal average response times in minutes for MSK are $T^c = 7.076$ and $T^{opt} = 6.430$. The values of current and optimal average response times for ZA are $T^c = 6.819$ and $T^{opt} = 6.00$. The computational time consumed to obtain the optimal solution T^{opt} was 57.4 seconds for MSK and 30.4 seconds for ZA.

If we know the values of T^c and T^{opt}, the demanded average response time T^d may get its value from the interval $[T^{opt}, T^c]$ based on the following expression: $T^d = \alpha T^c + (1-\alpha)T^{opt}$ for $\alpha = 0.0, 0.1, \ldots, 1.0$.

The results for the benchmarks ZA and MSK are plotted in Table 3. Symbol *NoTot* denotes the number of ambulances necessary for reaching the demanded average response time, *CT* stands for computational time in seconds and T^* is the resulting average response time in minutes.

α	**Region MSK**			**Region ZA**		
	NoTot	*CT*	T^*	*NoTot*	*CT*	T^*
0.0	64	2041.4	6.430	37	1846.6	6.000
0.1	62	2274.1	6.494	36	2221.7	6.078
0.2	61	1913.2	6.528	35	2211.8	6.160
0.3	59	2143.1	6.603	34	2054.5	6.245
0.4	57	1648.3	6.685	34	1695.4	6.245
0.5	56	1365.0	6.731	33	1295.5	6.331
0.6	55	1504.0	6.781	32	2789.9	6.418
0.7	54	1199.4	6.833	31	2365.0	6.554
0.8	53	1760.5	6.893	30	1629.0	6.600
0.9	51	2089.2	7.010	29	1946.1	6.694
1.0	50	1815.5	7.074	28	2521.4	6.800

Table 3 Results of performed numerical experiments

4 Conclusions

This paper addressed the problem of a manager of an emergency medical care system being forced to reduce the current ambulance fleet. Reduction generally means a deterioration in system efficiency, where system efficiency is measured by the average response time of the system to the users' requests. The average response time can range from a value corresponding to the current deployment of ambulances to the optimal value achieved by unrestricted

system reorganization. Several problems have been solved with the aim of achieving the prescribed average response time with a minimum number of ambulances.

The experimental results show that the current average response time can be achieved with a significantly smaller ambulance fleet. Moreover, an average response time equal to the average of the optimal and current response times can be achieved with a fleet the size of which is approximately 12 percent smaller than the current fleet size.

Future research in the field of public service system optimization can be aimed at verification of the obtained results by simulation of a real system. The goal is to develop a suitable optimization method, the results of which could be not only confirmed by simulations experiments, but the quality of which could enable their practical application in real world conditions.

Acknowledgements

This paper was supported by the following Slovak national research grants: VEGA 1/0220/25 “Regional equality in access to public rescue services provided from service centers on the transport network”, VEGA 1/0696/26 “Predictions of delays and vehicle occupancies to inform the planning of public transport”, and VEGA 1/0516/26 “Optimal design of charging infrastructure for electric vehicles and its efficient operation in urban mobility conditions”.

References

Avella P, Sassano A, Vasil'ev I (2007) Computational study of large scale p-median problems. Mathematical Programming 109: 89-114. https://doi.org/10.1007/s10107-005-0700-6

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. European Journal of Operations research 147: 451-463. https://doi.org/10.1016/S0377-2217(02)00364-8

Chanta, S, Mayorga, M E, McLay, L A (2014). Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. Annals of Operations Research 221: 133-159

Czimmermann P, Koháni M (2018) Characteristics of changes of transportation performance for pairs of critical edges. Communications 20: 84-87

Doerner K F, Gutjahr W J, Hartl R F, Karall M, Reimann M (2005) Heuristic Solution of an Extended Double-Coverage Ambulance Location Problem for Austria. Central European Journal of Operations Research 13: 325-340

García S, Labbé M, Marín A (2011) Solving large p-median problems with a radius formulation. INFORMS Journal on Computing 23: 546-556. https://doi.org/10.1287/ijoc.1100.0418

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. Health care management science 11: 262-274. https://doi.org/10.1007/s10729-007-9048-1

Janáček J (2008) Approximate Covering Models of Location Problems. In: Lecture Notes in Management Science: Proceedings of the 1st International Conference ICAOR, Yerevan, Armenia, pp 53-61

Janáček J, Kvet M (2016) Sequential approximate approach to the p-median problem. Computers & industrial engineering 94: 83-92. https://doi.org/10.1016/j.cie.2016.02.004

Jankovič P (2016) Calculating Reduction Coefficients for Optimization of Emergency Service System Using Microscopic Simulation Model. In: CINTI 2016. Conference Proceedings, pp 163-167. https://doi.org/10.1109/CINTI.2016.7846397

Jánošíková Ľ (2007) Emergency Medical Service Planning. Communications 9: 64-68

Jánošíková Ľ, Jankovič P, Kvet M, Zajacová F (2021) Coverage versus response time objectives in ambulance location. International Journal of Health Geographics 20: 1-16. https://doi.org/10.1186/s12942-021-00285-x

Karatas M, Yakıcıa E (2019) An analysis of p-median location problem: Effects of backup service level and demand assignment policy. European Journal of Operational Research 272: 207-218

Kvet M (2014) Computational Study of Radial Approach to Public Service System Design with Generalized Utility. In: Digital Technologies 2014. Conference Proceedings, Žilina, Slovakia, pp 198-208

Kvet M (2015) Advanced Radial Approach to Resource Location Problems. In: Studies in Computational Intelligence: Developments and Advances in Intelligent Systems and Applications, pp 29-48. https://doi.org/10.1007/978-3-319-58965-7_3

Matiaško K, Kvet M (2017) Medical data management. Informatics 2017. Conference Proceedings, pp 253-257

Reuter-Oppermann M, van den Berg P L, Vile J L (2017) Logistics for Emergency Medical Service systems, Health Systems 6: 187-208. https://doi.org/10.1057/s41306-017-0023-x

Revelle C S, Eiselt H A, Dakin M S (2008) A bibliography for some fundamental problem categories in discrete location science, European Journal of Operational Research 184: 817–848. https://doi.org/10.1016/j.ejor.2006.12.044

A Multilevel Analysis of Post-Harvest Food Loss in Sub-Saharan Africa: Commodity, Food Supply Chain Stage, and Regional Interactions

Elvis Kwakye[1]

Abstract. Post-harvest food loss (PHL) is widely recognized as a critical constraint on food security in sub-Saharan Africa (SSA), where an estimated 20% of agricultural output is lost through supply-chain inefficiencies. This study analyses how losses emerge through the multilevel interaction of commodities, food supply chain stages, and regions using a multilevel linear mixed-effects model (LMM) applied to 2,461 observations from 36 SSA countries covering five cereal commodities across seventeen food supply chain stages (FAO Data Lab, 2019–2024). The food supply chain stage is the strongest predictor of PHL: processing losses exceed 15% in Western and Central Africa, storage losses are highest in Eastern Africa (4.26%), and farm-level losses remain above 4% across all regions. The commodity–stage interaction shows that all five cereals exhibit their highest predicted losses at the Farm–Farm stage, with secondary peaks for maize and rice at husking and harvesting, and for millet and sorghum at threshing and winnowing. These findings contradict the assumption that a uniform post-harvest intervention is sufficient and demonstrate that effective policy must be stage-specific and region-specific, with direct implications for SDG 2 (Zero Hunger) and SDG 12.3 (reducing food loss and waste by 2030).

Keywords: post-harvest food loss, sub-Saharan Africa, mixed-effects model, cereal commodities, SDG 2, SDG 12.3

JEL Classification: Q18, C23

1 Introduction

Post-harvest food loss (PHL) has significant consequences in sub-Saharan Africa (SSA), where agricultural production contributes 20–50% of GDP and is the primary livelihood source for a substantial share of the population (Giller, 2020). SSA loses an estimated 20% of its agricultural output through supply-chain inefficiencies (Lufu et al., 2025), against a backdrop of growing nutritional deficits, fragile market systems, and increasing climate-related shocks (Gatto and Chepeliev, 2024). Globally, food loss and waste accounts for an estimated 24% of food energy content and more than $1 trillion in annual economic losses (Economou et al., 2024). Reducing PHL is therefore not only a food security imperative but also a pathway to economic resilience and environmental sustainability.

Despite growing research, the PHL literature in SSA remains fragmented, dominated by single-commodity or single-country investigations that overlook broader regional patterns. As a result, policymakers lack the comprehensive evidence needed to design well-targeted interventions. Affognon et al. (2015) and Mawoneke et al. (2025) explicitly call for multi-commodity, multi-stage empirical studies to address this gap.

This study responds to that call. Using harmonized FAO Data Lab estimates (2019–2024) from 36 SSA countries, it applies a multilevel LMM that simultaneously estimates fixed effects for commodity, stage, region, and their interactions, and random effects for country-level heterogeneity. Three key contributions: (i) comprehensive cross-commodity, cross-stage, cross-region PHL analysis; (ii) multilevel modelling capturing interaction effects and country-level variation; (iii) identification of stage-specific and region-specific vulnerabilities to inform targeted interventions.

[1] VSB–Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic. elvis.kwakye.st@vsb.cz

2 Literature Review and Conceptual Framework

Food systems theory views food loss as the product of interconnected social, technical, and environmental processes (Garnett, 2013). Losses are not evenly distributed: perishable commodities experience 30–60% losses concentrated at farm and early-distribution phases (Abdullah and Wang, 2024), while cereals suffer high losses at drying and storage due to inadequate facilities, high moisture content, and pest infestation (Dumitru et al., 2020). Regional disparities reflect differences in climate, infrastructure, and market access: West Africa's humidity renders traditional drying inefficient, while East Africa's reliance on sun-drying affects grain quality and market value. Existing commodity-level studies illuminate specific loss points without capturing broader patterns: research on maize identifies drying and storage as critical loss points (Erenstein et al., 2022); groundnut studies highlight aflatoxin contamination during storage and transport (Daba et al., 2023); and reviews of fruit and vegetable chains note the absence of systematic stage-level quantification (Owino et al., 2025). This fragmentation leaves policymakers without the comparative evidence needed for well-targeted interventions (Parfitt et al., 2010).

A critical gap in the existing literature is the absence of studies that simultaneously examine multiple commodities, multiple food supply chain stages, and multiple regions within a single analytical framework. Single-commodity or single-country studies cannot reveal whether the stage at which losses peak varies systematically by crop type, or whether regional context moderates the stage–loss relationship. The present study directly addresses this by estimating Commodity×Stage and Stage×Region interaction effects within a multilevel model. Country-level structural factors (governance, infrastructure, institutional capacity) are modelled as random effects, following supply chain resilience theory (Ponomarov and Holcomb, 2009), which argues that lasting loss reductions depend on the ability of food supply networks to withstand disruptions at multiple points simultaneously.

3 Methodology

3.1 Data Source and Variables

Secondary data from the FAO Data Lab (2019–2024) comprise 2,461 observations from 36 SSA countries. The FAO Data Lab integrates national field surveys, modelled estimates, and expert assessments. Because measurement methods and definitions of 'food loss' are not uniform across countries, the dataset introduces potential measurement error; country-level random intercepts partially absorb these systematic baseline differences. Where a country lacked a direct field survey, the FAO Data Lab substitutes a modelled estimate rather than a blank cell; this study uses these records as provided, without further imputation, and the final 2,461-row dataset contains no missing values on any model variable. Five staple cereals are covered: maize, rice, millet, sorghum, and wheat. Countries are clustered into four regions following FAO/UN classifications: Central, Eastern, Southern, and Western Africa. Year is included as a continuous fixed effect (2019–2024) to partially capture temporal variation including the COVID-19 period, during which supply-chain disruptions may have elevated losses (Vos et al., 2025).

3.2 Food Supply Chain Stage Classification

The FAO Data Lab records PHL at seventeen activity–stage categories, used directly as the stage factor levels in the model: Distribution, Distribution (reference category, the omitted baseline level in the fixed-effects table); Drying, Farm; Drying, Harvesting, Harvest; Farm, Farm; Husking, Farm; Shelling, Farm; Shelling, Threshing, Farm; Threshing, Farm; Winnowing, Farm; Storage, Farm; Storage, Storage; Processing, Processing; Packaging, Packing; Marketing, Market; Transportation, Farm; and Transportation, Transport (16 estimated coefficients, Table 2). In Table 4, these seventeen categories are aggregated into eight summary labels using simple averages: 'Farm' combines Drying-Farm, Farm-Farm, Husking-Farm, Shelling-Farm, Shelling/Threshing-Farm, Threshing-Farm, and Winnowing-Farm (7 categories); 'Harvest' combines Harvesting-Harvest and Drying/Harvesting-Harvest (2); 'Storage' combines Storage-Farm and Storage-Storage (2); 'Transport' combines Transportation-Farm and Transportation-Transport (2); 'Processing', 'Packing', 'Market', and 'Distribution' each correspond to a single underlying category.

3.3 Model Specification and Diagnostics

Three properties of the data make the multilevel LMM statistically necessary over OLS or fixed-effects regression: (i) non-independence of observations nested within countries, producing downward-biased standard errors if ignored (Bolker, 2015); (ii) unbalanced hierarchical structure favoring random intercepts over country

dummies (Kanyama Busanga and Njuho, 2024); (iii) the need to estimate Commodity×Stage and Stage×Region interactions. The model is:

$$Loss_{ijkc} = \beta_0 + \beta_{com,i} + \beta_{stage,j} + \beta_{reg,k} + \beta_{year} \cdot Year + \beta_{com\times stage,ij} + \beta_{stage\times reg,jk} + u_{0c} + \varepsilon_{ijkc} \quad (1)$$

where $u_{0c} \sim N(0, \sigma^2_u)$ is the country-level random intercept and $\varepsilon \sim N(0, \sigma^2)$ is the residual. The model was estimated in R using lme4 (Bates et al., 2015) via REML for unbiased variance component estimation; Type II Wald χ^2 tests assess fixed-effect significance (Fox and Weisberg, 2019). Three nested models were compared under ML using AIC, BIC, and likelihood-ratio tests before re-estimation under REML.

Residual diagnostics revealed mild heteroskedasticity (Breusch–Pagan: BP = 89.07, $p < .001$) and positive autocorrelation (Durbin–Watson: DW = 0.937, $p < .001$). These features do not invalidate coefficient estimates; the LMM estimator remains consistent, but reported standard errors may be slightly understated. A robustness check re-estimating Model 3 with heteroskedasticity-consistent (HC3) standard errors confirmed unchanged substantive conclusions. The Breusch–Pagan test is applied here as a diagnostic indicator rather than a formal test of LMM assumptions, and the large sample (n = 2,461) ensures asymptotic reliability via the central limit theorem (Bates et al., 2015).

4 Results

4.1 Model Selection

Adding Commodity×Stage (Model 2) produced a significant improvement over the main-effects model ($\chi^2 = 381.62$, df = 27, $p < .001$); adding Stage×Region (Model 3) improved fit further ($\chi^2 = 87.29$, df = 26, $p < .001$). Model 3 achieved the lowest AIC (10,204) and BIC (10,686) and was selected as the final model (Table 1). The final REML model (2,461 observations; 36 country groups) yielded country-level intercept variance $\sigma^2_u = 0.074$ (SD = 0.27) and residual variance $\sigma^2 = 3.543$ (SD = 1.88). The intraclass correlation coefficient (ICC = 2.0%) confirms that 98% of PHL variance is within-country. Marginal $R^2 = 0.436$; conditional $R^2 = 0.447$.

Model	Params	AIC	BIC	logLik	χ^2	df	p
Model 1 (main effects)	30	10,567	10,741	−5,253.5	–	–	–
Model 2 (+Com.×Stage)	57	10,239	10,571	−5,062.7	381.62	27	<.001
Model 3 (+Stage×Region)	83	10,204	10,686	−5,019.1	87.29	26	<.001

Note: χ^2 and p-values are for the likelihood-ratio test vs. the preceding model.

Table 1 Model comparison using maximum likelihood

4.2 Significance of Predictors and Fixed-Effect Estimates

Type II Wald χ^2 tests (Table 2) confirmed the food supply chain stage as the dominant predictor ($\chi^2(16) = 1{,}097.04$, $p < .001$), substantially larger than any other effect. Commodity×Stage ranked second ($\chi^2(27) = 452.71$), followed by commodity ($\chi^2(4) = 134.81$) and Stage×Region ($\chi^2(26) = 94.08$; all $p < .001$). Region reached significance ($p = .021$); year did not ($p = .095$), indicating no systematic PHL trend over 2019–2024. The processing stage showed the largest fixed effect above the Distribution baseline (+16.67 pp; 95% CI: [14.92, 18.42]), followed by Farm–Farm (+15.55 pp; [13.89, 17.21]) and Husking–Farm (+9.74 pp; [8.23, 11.25]). Regional main effects were not individually significant, confirming that regional variation operates through the Stage×Region interaction.

Effect	χ^2	df	*p*-value
Commodity	134.81	4	< .001
Stage	1,097.04	16	< .001
Region	9.74	3	.021
Year	2.78	1	.095
Commodity × Stage	452.71	27	< .001
Stage × Region	94.08	26	< .001

Note: Each χ^2 is a Type II Wald statistic testing whether that fixed effect explains significant variance in PHL, conditional on all other effects; df = number of parameters for that effect. Unlike the likelihood-ratio χ^2 in Table 1 (which compares nested models), these Wald tests assess individual effects within the final REML-estimated Model 3.

Table 2 Type II Wald χ^2 tests for fixed effects

4.3 Commodity × Food Supply Chain Stage Interaction

Despite similar overall mean losses across the five cereals (4.38–5.13%), stage-specific predictions reveal starkly different vulnerability profiles (Table 3). Rice recorded the highest maximum predicted loss (17.10%; 95% CI: [15.3, 18.9]), driven by milling inefficiencies and grain breakage at Processing–Mill (Tong et al., 2019). Maize peaked at 16.40% ([14.6, 18.2]) at Husking–Farm. Millet and sorghum exhibited their highest losses during Winnowing and Threshing, consistent with their dependence on manual processing methods. Wheat's maximum (15.90%; [14.1, 17.7]) occurs at Processing–Mill. These patterns indicate that commodity-specific, stage-targeted technology interventions are needed rather than generic programmes.

Commodity	Mean (%)	SD	Min (%)	Max (%)	*n* Stages	Peak stage (95% CI)
Maize	5.03	3.85	0.90	16.40	17	Farm–Farm [14.1, 18.8]
Millet	4.38	3.92	0.58	16.10	17	Farm–Farm [12.8, 17.5]
Rice	5.13	4.24	0.72	17.10	17	Farm–Farm [13.3, 18.0]
Sorghum	4.48	3.87	0.38	15.90	17	Farm–Farm [13.7, 18.4]
Wheat	4.49	3.74	0.33	15.90	17	Farm–Farm [13.7, 18.5]

Note: Predictions from Model 3 (REML). 95% CIs from fixed-effect variance–covariance matrix. Stage abbreviations follow Section 3.2.

Table 3 Predicted PHL (%) by commodity across all food supply chain stages

4.4 Food Supply Chain Stage × Region Interaction and Country-Level Random Effects

PHL does not follow a uniform pattern across SSA (Table 4). Western and Central Africa record the highest processing losses (17.98% and 17.35% respectively), reflecting systemic underinvestment in processing infrastructure and technology that cannot be attributed to commodity mix or country governance, both of which are controlled. Eastern Africa exhibits the highest storage loss (4.26%; 95% CI: [3.5, 5.0]), nearly double of Southern Africa's (2.61%; [1.9, 3.3]) pointing to inadequate hermetic storage adoption, persistent pest pressure, and humidity (Paterson and Lima, 2011; Baributsa and Njoroge, 2020). Farm-level losses are strikingly consistent across all four regions (5.12–5.89%), indicating structural pan-SSA on-farm handling challenges. Transport, distribution, and market stages perform comparatively well across all regions, confirming that upstream stages are the primary targets for intervention.

Region	Distrib.	Farm	Harvest	Market	Packing	Processing	Storage	Transport
Central	0.68 [0.3, 1.1]	5.55 [4.8, 6.3]	6.16 [5.4, 7.0]	2.42 [1.8, 3.1]	6.30 [5.5, 7.1]	17.35 [15.4, 19.3]	2.66 [2.0, 3.4]	1.04 [0.6, 1.5]
Eastern	1.06 [0.6, 1.5]	5.79 [5.0, 6.6]	4.90 [4.1, 5.7]	2.80 [2.1, 3.5]	6.68 [5.9, 7.5]	4.78 [3.9, 5.7]	4.26 [3.5, 5.0]	1.25 [0.8, 1.7]
Southern	0.09 [0.0, 0.3]	5.12 [4.3, 5.9]	5.44 [4.6, 6.3]	1.83 [1.2, 2.5]	5.70 [4.9, 6.5]	0.73 [0.2, 1.3]	2.61 [1.9, 3.3]	1.40 [0.9, 1.9]
Western	1.31 [0.8, 1.8]	5.89 [5.1, 6.7]	4.84 [4.0, 5.7]	3.06 [2.3, 3.8]	6.93 [6.1, 7.7]	17.98 [16.1, 19.9]	3.56 [2.8, 4.3]	2.46 [1.8, 3.1]

Note: Distrib. = distribution; Transp. = transport. Food supply chain stage groupings follow the eight-category aggregation defined in Section 3.2. CIs from fixed-effect variance–covariance matrix.

Table 4 Predicted PHL (%) by region and food supply chain stage (approx. 95% CIs in brackets)

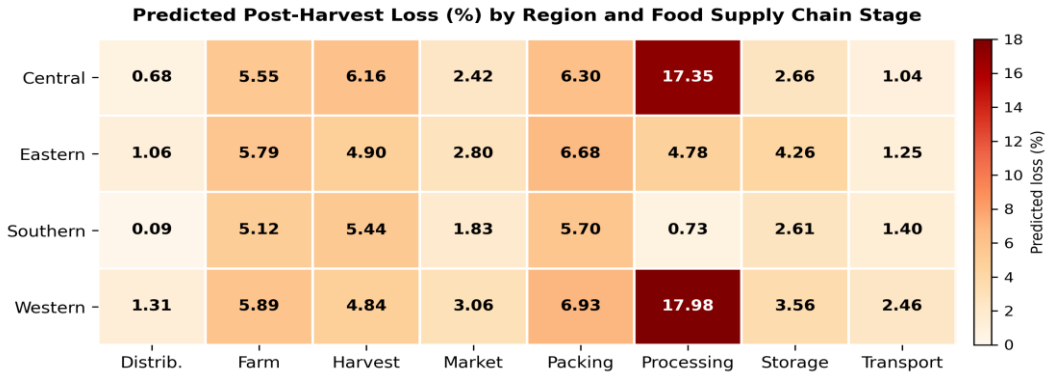


Figure 1 Predicted post-harvest loss (%) by region and food supply chain stage. Darker cells indicate higher predicted loss. Processing losses in Western (17.98%) and Central Africa (17.35%) stand out as the dominant hotspot; Eastern Africa's elevated storage loss (4.26%) is visible as a secondary, regionally distinct pattern. Values correspond to Table 4.

Country-level intercept variance was small ($\sigma^2_u = 0.074$; ICC = 2.0%), confirming that stage- and region-level effects dominate PHL variation. Nigeria ($\hat{w}_{0c} = 0.650$; 95% CI: [0.316, 0.984]) and Ethiopia (0.567; [0.300, 0.834]) showed significant positive deviations; the remaining 34 countries' confidence intervals crossed zero.

5 Discussion

The food supply chain stage is the variable most strongly associated with PHL in SSA, accounting for more variance than commodity type, region, or country-level governance structures—advancing the case for multi-stage, multi-commodity analyses (Affognon et al., 2015; Mawoneke et al., 2025). The multilevel approach reveals three substantive patterns that conventional OLS or fixed-effects regressions would obscure.

First, processing losses in Western (17.98%) and Central Africa (17.35%) far exceed those in Eastern (4.78%) and Southern Africa (0.73%). This contrast reflects longstanding systemic underinvestment in processing infrastructure (Hodges et al., 2011; Dumitru et al., 2020)—not commodity mix or governance, both of which are controlled. Unlike prior studies treating processing loss as a background factor, this analysis quantifies the regional asymmetry while controlling for country-level heterogeneity.

Second, Eastern Africa's storage loss (4.26%) is nearly double Southern Africa's (2.61%), consistent with documented inadequate hermetic storage adoption and persistent mycotoxin contamination (Paterson and Lima, 2011; Baributsa and Njoroge, 2020); this gap has not been previously quantified at this level of methodological rigor.

Third, farm-level losses are uniformly high across all four regions (5.12–5.89%), indicating structural pan-SSA challenges amenable to broad technology solutions such as solar drying (Erenstein et al., 2022; Mohammed et al., 2020).

The small ICC (2%) implies that stage-specific interventions can deliver measurable PHL reductions within existing governance structures, without requiring broad institutional reform—challenging narratives that treat weak institutions as the primary barrier. These findings are associative, not causal; unobserved confounders may remain despite controlling for commodity, region, and country effects.

6 Conclusion

Three policy-relevant findings emerge. First, Western and Central Africa require urgent investment in processing infrastructure (losses 17–18%): processing losses at this scale reflect systemic underinvestment in milling technology, and equipment upgrading combined with technical training are the most practical entry points (Hodges et al., 2011). Second, Eastern Africa's priority is improving storage conditions (4.26%): hermetic storage remains the evidence-based solution despite cost and access barriers, and its adoption gap relative to Southern Africa (2.61%) is quantified here for the first time within a controlled multilevel framework (Benimana et al., 2023; Baributsa and Njoroge, 2020). Third, farm-level losses constitute a pan-SSA structural challenge (5.12–5.89% across all four regions): their uniformity confirms that no single country or sub-region bears disproportionate responsibility, and solar drying offers a low-cost, widely deployable option (Mohammed et al., 2020).

These findings carry a direct governance implication: the intraclass correlation of 2% shows that country-level institutional variation explains only a small fraction of PHL variance. Stage-specific interventions can therefore deliver measurable reductions within existing governance structures, without waiting for broad institutional reform. Regional cooperation among countries that share similar loss profiles—processing-intensive West and Central Africa; storage-vulnerable Eastern Africa—directly advances SDG 2 (Zero Hunger) and SDG 12.3 (halving food loss and waste by 2030), and is a more actionable pathway than pan-SSA uniform programmes (Stathers et al., 2020; Mawoneke et al., 2025).

Limitations include: non-standardized FAO measurement methods across 36 countries; restriction to five cereals; absence of economic valuation of losses; COVID-19 period overlap; and omission of gender disaggregation (women bear a disproportionate share of labor at farm, drying, and processing stages—precisely where this study identifies the strongest loss associations). Future research should extend to non-cereal commodities, incorporate gender-disaggregated data, and pursue quasi-experimental designs for causal inference.

Acknowledgements

This article has been produced by the SGS project (SP2026/071) of the Faculty of Economics, VSB-Technical University of Ostrava. The author thanks Prof. Jana Hančlová for guidance and supervision throughout this research, and the FAO Data Lab for making the post-harvest loss estimates publicly available.

References

Abdullah W Z (2024) Rural roadways and regional sustenance: transportation infrastructure and food insecurity in Pakistan. Environment, Development and Sustainability 143. https://doi.org/10.1007/s10668-024-05838-1

Affognon H, Mutungi C, Sanginga P, Borgemeister C (2015) Unpacking postharvest losses in sub-Saharan Africa: a meta-analysis. World Development 66:49–68. https://doi.org/10.1016/j.worlddev.2014.08.002

Baributsa D, Njoroge A (2020) Hermetic technologies for grain storage among smallholder farmers in Eastern Kenya. Journal of Stored Products Research 87:101618. https://doi.org/10.1016/j.jspr.2020.101618

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67:1–48. https://doi.org/10.18637/jss.v067.i01

Benimana G U, Ritho C, Irungu P (2023) Impact of adopting maize hermetic storage technologies in Rwanda. Heliyon 9:e14592. https://doi.org/10.1016/j.heliyon.2023.e14592

Bolker BM (2015) Linear and generalized linear mixed models. In: Ecological Statistics: Contemporary Theory and Application. Oxford University Press, pp 309–333

Daba H G, Delele M A, Fanta S W, Satheesh N (2023) The extent of groundnut post-harvest loss in Africa. Journal of Agriculture and Food Research 14:100826. https://doi.org/10.1016/j.jafr.2023.100826

Dumitru O M, Iorga S, Vladut N V, Bracaescu C (2020) Food losses in primary cereal production. INMATEH-Agricultural Engineering 62. https://doi.org/10.35633/inmateh-62-14

Economou F, Voukkali I, Papamichael I et al. (2024) Turning food loss and food waste into watts. Energies 17:3191. https://doi.org/10.3390/en17133191

Erenstein O, Jaleta M, Sonder K et al. (2022) Global maize production, consumption and trade. Food Security 14:1295–1319. https://doi.org/10.1007/s12571-022-01288-7

Fox J, Weisberg S (2019) An R Companion to Applied Regression, 3rd ed. SAGE Publications, Thousand Oaks. https://www.john-fox.ca/Companion/

Garnett T (2013) Food sustainability: problems, perspectives and solutions. Proceedings of the Nutrition Society 72:29–39. https://doi.org/10.1017/S0029665112002947

Gatto A, Chepeliev M (2024) Global food loss and waste estimates show increasing pressures. Nature Food 5:136–147. https://doi.org/10.1038/s43016-023-00915-6

Giller K E (2020) The food security conundrum of sub-Saharan Africa. Global Food Security 26:100431. https://doi.org/10.1016/j.gfs.2020.100431

Hodges R J, Buzby J C, Bennett B (2011) Postharvest losses and waste in developed and less developed countries. Journal of Agricultural Science 149:37–45. https://doi.org/10.1017/S0021859610000936

Kanyama Busanga J, Njuho P (2024) Linear mixed model effects in meta-analysis of agricultural data. International Journal of Agricultural & Statistical Sciences 20:511–520. https://doi.org/10.59467/IJASS.2024.20.511

Lufu R, Ambaw A, Opara U L (2025) Evaporative cooling systems for perishables in sub-Saharan Africa. Journal of Food Process Engineering 48:e70127. https://doi.org/10.1111/jfpe.70127

Mawoneke K G, Rusike T, Makumbirofa H M, Mukwidigwi J (2025) From farm to fork: strategies for sustainable reduction of post-harvest losses in SSA. Cogent Food & Agriculture 11:2588851. https://doi.org/10.1080/23311932.2025.2588851

Mohammed S, Fatumah N, Shadia N (2020) Drying performance and economic analysis of solar dryers for drying fruits in East Africa. Journal of Stored Products Research 88:101634. https://doi.org/10.1016/j.jspr.2020.101634

Owino W, Kahenya P, Wafula E et al. (2025) Food loss and waste reduction in fruit and vegetable value chains in Eastern Africa. Foods 14:3938. https://doi.org/10.3390/foods14223938

Parfitt J, Barthel M, Macnaughton S (2010) Food waste within food supply chains. Philosophical Transactions of the Royal Society B 365:3065–3081. https://doi.org/10.1098/rstb.2010.0126

Paterson R R M, Lima N (2011) Further mycotoxin effects from climate change. Food Research International 44:2555–2566. https://doi.org/10.1016/j.foodres.2011.05.038

Ponomarov S Y, Holcomb M C (2009) Understanding supply chain resilience. International Journal of Logistics Management 20:124–143. https://doi.org/10.1108/09574090910954873

Stathers T, Holcroft D, Kitinoja L et al. (2020) A scoping review of interventions for crop postharvest loss reduction in SSA and South Asia. Nature Sustainability 3:821–835. https://doi.org/10.1038/s41893-020-00622-1

Tong C, Gao H, Luo S et al. (2019) Impact of postharvest operations on rice grain quality. Comprehensive Reviews in Food Science and Food Safety 18:626–640. https://doi.org/10.1111/1541-4337.12439

Vos R, Glauber J, Hebebrand C, Rice B (2025) Global shocks to fertilizer markets. Food Policy 133:102790. https://doi.org/10.1016/j.foodpol.2024.102790

Comparative Analysis of Mathematical Programming and Metaheuristic Approaches for the Vehicle Routing Problem with Multiple Soft Time Windows

Farouk Olatunji Ligali[1], Dušan Hrabec[2], Lars Magnus Hvattum[3]

Abstract. This paper presents a controlled experimental comparison of a Mixed-Integer Linear Program (MILP) solved by Gurobi, a restricted route-pool Lagrangian procedure (LR), a Genetic Algorithm (GA), and Simulated Annealing (SA) applied to the Vehicle Routing Problem with Multiple Soft Time Windows. The benchmark is the published C101-MTW multiple-time-window extension of the Solomon C101 instance, restricted to nested customer subsets of size $N \in \{10, 25, 50\}$. All stochastic methods are evaluated over ten seeded runs per instance size under identical cold-start wall-clock budgets. The MILP incumbent provides the best objective value at every size, although it does not certify global optimality and its LP-based lower bounds remain weak. Among the heuristics, GA attains the lowest mean relative deviation from the MILP incumbent at $N = 10$ (0.10%) and $N = 25$ (2.56%), with seed-paired advantages over SA that are large and statistically significant; at $N = 50$ the deviations of GA (5.02%) and SA (4.93%) are statistically indistinguishable. LR exhausts its budget during route-pool generation at $N \geq 25$ and returns its constructive fallback solution. The study offers a reproducible protocol and size-dependent guidance for method selection under moderate computational budgets, with conclusions restricted to the studied instance family.

Keywords: Vehicle Routing Problem, Soft Time Windows, Branch-and-Bound, Lagrangian Relaxation, Genetic Algorithm, Simulated Annealing, Solomon Benchmark

JEL Classification: C44

1 Introduction

In contemporary e-commerce and urban logistics, last-mile delivery accounts for a disproportionate share of total supply chain cost and complexity (Savelsbergh and Van Woensel, 2016). Carriers must coordinate fleets of capacitated vehicles to serve hundreds of customers daily, each with individual demand and preferred service windows, and failing to arrive within those windows incurs both operational penalties and reputational costs. The theoretical framework for this class of problems is the Vehicle Routing Problem with Time Windows (VRPTW) (Solomon, 1987). In practice, rigid hard time windows are often impractical: traffic variability, stochastic service durations, and customers who accept alternative delivery slots motivate the soft time window variant, in which violations are penalized rather than forbidden. Extending this further, many real-world delivery systems allow customers to specify multiple preferred windows per day - a morning slot or an afternoon slot - giving rise to the Vehicle Routing Problem with Multiple Soft Time Windows (VRP-MSTW). Despite its practical relevance, the VRP-MSTW has received less attention than its single-window counterpart, and systematic comparisons of exact versus heuristic methods on standard benchmark instances are scarce.

This paper addresses that gap by benchmarking four representative optimization methods - spanning exact mathematical programming, classical dual decomposition, and two modern metaheuristics - on the C101-MTW instance from a published multiple-time-window extension of the Solomon datasets. Throughout the paper, the MILP reference denotes the best feasible incumbent found by the solver within the wall-clock budget, not a proven global optimum, and all heuristic quality measures are relative deviations from this incumbent. The main contribution is not a new problem class or algorithm but a standardized, cross-method benchmarking study for the VRP-MSTW under symmetric wall-clock budgets, which, to the best of our knowledge, distinguishes this work from existing VRPTW and VRP-STW benchmark studies. The specific contributions are:

[1] Tomas Bata University in Zlín, Faculty of Applied Informatics, Nad Stráněmi 4511 760 05 Zlín, Czech Republic, f_ligali@utb.cz
[2] Tomas Bata University in Zlín, Faculty of Applied Informatics, Nad Stráněmi 4511 760 05 Zlín, Czech Republic, hrabec@utb.cz
[3] Molde University College, Faculty of Logistics, Britvegen 2, NO-6410 Molde, Norway, lars.m.hvattum@himolde.no

1. A consistent mathematical formulation of the VRP-MSTW with a unified cost function enabling fair cross-method comparison.
2. A controlled experimental design in which all heuristic methods share the same time budget as the Mixed-Integer Linear Program (MILP) solver at each instance scale.
3. A statistical evaluation over ten seeded replications per stochastic method, reporting mean relative deviation from the MILP incumbent, dispersion measures, and seed-paired comparisons between heuristics.
4. A scalability analysis across $N \in \{10, 25, 50\}$ that links each method's degradation pattern to its recorded search depth, indicating that, on this instance family, the heuristic ranking is size-dependent rather than universal.

2 Literature Review

The theoretical foundation of this study is the VRPTW, formally established by Solomon, 1987. Over the past three decades, the VRPTW has become one of the most intensively studied paradigms in combinatorial optimization (Braekers et al, 2016). However, traditional formulations impose hard time windows, treating any temporal violation as strictly infeasible. As urban logistics and e-commerce have evolved (Savelsbergh and Van Woensel, 2016), rigid constraints have become increasingly inadequate for modelling real-world stochasticity and customer preferences. This limitation motivated the development of the Vehicle Routing Problem with Soft Time Windows (VRP-STW), wherein delivery outside the preferred window is permitted but incurs a penalized cost (Taş et al, 2014). Effective methods for the soft-window variant include multistage very large-scale neighborhood search (Mouthuy et al, 2015). A further, practically important extension is the Vehicle Routing Problem with Multiple Time Windows (VRP-MTW). In modern attended home delivery, customers frequently specify non-contiguous availability periods (e.g., a morning slot and an evening slot). Kourank Beheshti et al, 2015 show, in a distribution case study, that allowing multiple prioritized time windows increases routing flexibility and operational efficiency, albeit at the cost of expanding the search space. For the hard-window VRP-MTW, hybrid variable neighborhood tabu search (Belhaiza et al, 2014), multi-start evolutionary heuristics (Belhaiza et al, 2019), and constant-time neighborhood evaluations (Hoogeboom et al, 2020) define the recent state of the art. The combination of these two features - the VRP-MSTW - represents a practically relevant yet computationally challenging variant. While the literature contains formulations for both STW and MTW independently, their intersection remains relatively sparse, often addressed only as sub-problems within broader rich routing frameworks.

Methodologically, the VRP literature is broadly divided into exact methods and approximate heuristics. Exact approaches, traditionally reliant on Branch-and-Cut or Branch-and-Price algorithms, excel at providing tight lower bounds and certified optima for small to medium instances; generic branch-cut-and-price frameworks now solve many VRPTW variants to optimality (Pessoa et al, 2020). However, they struggle with the combinatorial explosion inherent in multiple soft windows. To circumvent this, classical dual-decomposition techniques, such as Lagrangian Relaxation (Fisher, 1981), have historically been employed to isolate complicating constraints and yield tractable sub-problems. Conversely, modern routing implementations largely favor metaheuristics. Population-based methods, such as Genetic Algorithms (Goldberg, 1989) and their state-of-the-art hybrid descendants (Vidal, 2022), are valued for their global search capabilities and spatial recombination, while trajectory-based methods, such as Simulated Annealing (Kirkpatrick et al, 1983) and Adaptive Large Neighborhood Search (ALNS), tend to excel at rapid, localized exploitation. Despite the proliferation of sophisticated algorithms tailored to specific routing variants, standardized, cross-methodological benchmarking for the VRP-MSTW remains scarce. Studies introducing novel metaheuristics frequently compare their performance against disparate algorithms run on different hardware, with varying time budgets and termination criteria - practices at odds with established guidance on fair experimental comparison in optimization research (Kendall et al, 2016), and ones that make the trade-offs between exact, dual-decomposition, and heuristic techniques difficult to evaluate objectively. This paper addresses this gap by subjecting four representative optimization paradigms to a strictly controlled, time-symmetric evaluation on a standardized VRP-MSTW benchmark.

3 Problem Formulation

The VRP-MSTW models a distribution network where a central depot serves N urban customers. A fleet of identical, capacitated vehicles departs from the depot, delivers goods, and returns. Each customer specifies one or more preferred delivery windows; arriving outside all of them incurs a penalty proportional to the deviation. Vehicle capacity must not be exceeded. The operator minimizes the total cost of vehicle deployment, travel, and time-window violations. The problem is defined on a directed graph $G = (\mathcal{V}, \mathcal{A})$. The node set $\mathcal{V} = \{0\} \cup \mathcal{C}$ contains the depot (node 0) and the customer set $\mathcal{C} = \{1, \ldots, N\}$. The arc set $\mathcal{A} = \{(i, j) : i \neq j,\ i, j \in \mathcal{V}\}$

represents all feasible direct travel links.

Parameters

- $\mathcal{K}$: Set of homogeneous vehicles, indexed by k
- d_{ij}, t_{ij}: Travel distance and travel time from node i to node j. In the implementation, these are computed from Euclidean coordinates. The model generalizes to any metric satisfying the triangle inequality.
- q_i, s_i: Demand and service time at node $i \in C$
- Q: Maximum vehicle capacity
- $[a_{i,m}, b_{i,m}]$: m-th soft time window for customer i; $m \in \mathcal{W}_i$ (window set for customer i)
- α_i, β_i: Earliness and lateness penalty coefficients
- κ: Fixed vehicle activation cost
- U: Service-time upper bound, $U = B + Ns + (N+1)d_{\max}$, where $B = \max_{i,m} b_{i,m}$ and $d_{\max} = \max_{i,j} d_{ij}$
- M: Big-M constant for time propagation, $M = U + s + d_{\max}$

Decision Variables

$$
\begin{aligned}
x_{ijk} &\in \{0,1\}, \quad (i,j) \in \mathcal{A},\ k \in \mathcal{K} && \text{1 if vehicle } k \text{ traverses arc } (i,j) \\
y_k &\in \{0,1\}, \quad k \in \mathcal{K} && \text{1 if vehicle } k \text{ is used} \\
z_{i,m} &\in \{0,1\}, \quad i \in C,\ m \in \mathcal{W}_i && \text{1 if window } m \text{ is active for customer } i \\
T_i, E_i, L_i &\geq 0, \quad i \in C && \text{service-start time, earliness, lateness slacks}
\end{aligned}
$$

Objective Function

$$
\min\ Z = \kappa \sum_{k \in \mathcal{K}} y_k + \sum_{k \in \mathcal{K}} \sum_{(i,j) \in \mathcal{A}} d_{ij}\, x_{ijk} + \sum_{i \in C} \big(\alpha_i E_i + \beta_i L_i\big) \tag{1}
$$

Constraints

$$
\begin{aligned}
&\sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{V} \setminus \{j\}} x_{ijk} = 1 && \forall\, j \in C && \text{(C1)} \\
&\sum_{i \in \mathcal{V} \setminus \{p\}} x_{ipk} = \sum_{j \in \mathcal{V} \setminus \{p\}} x_{pjk} && \forall\, p \in C,\ k \in \mathcal{K} && \text{(C2)} \\
&\sum_{i \in C} q_i \sum_{j \in \mathcal{V} \setminus \{i\}} x_{ijk} \leq Q && \forall\, k \in \mathcal{K} && \text{(C3)} \\
&\sum_{m \in \mathcal{W}_i} z_{i,m} = 1 && \forall\, i \in C && \text{(C4)} \\
&E_i \geq \sum_{m \in \mathcal{W}_i} a_{i,m} z_{i,m} - T_i && \forall\, i \in C && \text{(C5)} \\
&L_i \geq T_i - \sum_{m \in \mathcal{W}_i} b_{i,m} z_{i,m} && \forall\, i \in C && \text{(C6)} \\
&\sum_{j \in C} x_{0jk} = y_k && \forall\, k \in \mathcal{K} && \text{(C7a)} \\
&\sum_{i \in C} x_{i0k} = y_k && \forall\, k \in \mathcal{K} && \text{(C7b)} \\
&T_j \geq d_{0j} - M(1 - x_{0jk}) && \forall\, j \in C,\ k \in \mathcal{K} && \text{(C8)} \\
&T_j \geq T_i + s + d_{ij} - M(1 - x_{ijk}) && \forall\, i, j \in C,\ i \neq j,\ k \in \mathcal{K} && \text{(C9)} \\
&x_{ijk}, y_k, z_{i,m} \in \{0,1\}, \quad T_i,\ E_i,\ L_i \geq 0 && && \text{(C10)}
\end{aligned}
$$

Constraint Explanation: (C1) visits each customer exactly once; (C2) conserves vehicle-specific flow; (C3) enforces capacity. (C4)–(C6) handle multiple soft windows: (C4) selects exactly one active window per customer, and (C5)–(C6) measure earliness and lateness against the selected window, so the solver chooses the penalty-minimizing window without explicit min/max operators. (C7a)–(C7b) link each vehicle's single depot departure and return to its activation variable, and (C8)–(C9) propagate service-start times T_i along routes; because no optimal schedule requires service beyond $U = B + Ns + (N+1)d_{\max}$, the constant $M = U + s + d_{\max}$ fully deactivates every unused constraint in (C9). Waiting is free of direct cost, yet service may deliberately begin before the selected window when the earliness incurred reduces downstream lateness; for each fixed route, service times

and window choices are optimized under this convention by a common evaluator shared by LR, GA, and SA, while the MILP enforces identical timing relationships directly. (C10) states the variable domains.

3.1 Optimization Methods

To evaluate solution strategies for the VRP-MSTW under a common protocol, this study benchmarks four distinct approaches. The reference exact method is a MILP solved via Gurobi's branch-and-bound engine (Gurobi Optimization, LLC, 2023). Because the VRP-MSTW is NP-hard, Gurobi is used to establish a high-quality reference benchmark solution (Z^{MILP}) under a strict time budget rather than to certify global optima. The decomposition baseline is a restricted route-pool Lagrangian procedure (Fisher, 1981). Singleton routes and a nearest-neighbor split provide feasible primal incumbents; additional routes are sampled from seeded permutations. Customer-cover equalities are dualized and updated by a subgradient method. Because the pool is not exhaustive, the resulting dual value bounds only the restricted pool master and is not reported as a lower bound for the original VRP-MSTW. To represent modern metaheuristics, we implemented a standard Genetic Algorithm (GA) (Goldberg, 1989) using order crossover (OX) and tournament selection, alongside a Simulated Annealing (SA) algorithm (Kirkpatrick et al, 1983) that explores the local neighborhood via Relocate, Swap, and 2-opt moves governed by a geometric cooling schedule. These textbook variants are retained deliberately, rather than state-of-the-art hybrids such as hybrid genetic search (Vidal, 2022), so that the comparison reflects paradigm behavior under equal budgets rather than differences in tuning effort.

4 Experimental Setup

4.1 Benchmark Instance: C101-MTW

The experiments use C101-MTW from the Multiple Time Window VRP dataset of Gülmez et al, 2025, an extension derived from the clustered Solomon C101 instance (Solomon, 1987). It retains the coordinates and demands of 100 customers on a $[0, 100]^2$ grid and provides three pairwise non-overlapping 60-unit windows per customer, with start times on a 30-unit grid from 480 to 900. The published instance is used without alteration. Scalability is examined through nested subsets containing the first $N \in \{10, 25, 50\}$ customer identifiers.

4.2 Parameter Values

All computational experiments use a uniform service time $s = 10$, vehicle capacity $Q = 200$, penalty coefficients $\alpha_i = 0.5$ (earliness) and $\beta_i = 2.0$ (lateness) per time unit, vehicle activation cost $\kappa = 1000$, and Euclidean distances on the $[0, 100]^2$ grid. The asymmetric penalties reflect a cost structure common in last-mile delivery, where lateness typically carries a higher weight than early arrival.

4.3 Common Wall-Clock Protocol

For each customer subset, every method was subject to the same wall-clock upper bound: 60, 150, and 450 seconds for $N \in \{10, 25, 50\}$, respectively. The measured runtime includes initialization, preprocessing, route generation or model construction, and optimization. Each algorithm–seed combination was executed in an independent process on the same machine using one computational thread, without warm starts or shared caches. For the MILP, model construction was included in the allotted time, and only the remaining time was passed to Gurobi.

All experiments were executed sequentially on one machine (Apple M4 Pro, 24 GB RAM, macOS 26.5) using Python 3.9 and Gurobi 12.0.3 with default settings apart from the seed, single-thread, and time-limit parameters. The GA uses a population of 24 with order crossover, tournament selection, mutation rate 0.25, and elitism of 4; SA uses initial temperature 120 with geometric cooling at rate 0.997 over Relocate, Swap, and 2-opt moves; LR samples up to 120 seeded route permutations and performs up to 120 subgradient iterations. All parameter values were fixed before the campaign and never tuned to its outcomes. The MILP was executed once for each subset, whereas the stochastic LR, GA, and SA implementations were evaluated over ten runs using the common seed set $\{101, \ldots, 110\}$ at every subset size. GA and SA ran until the wall-clock limit; LR stopped at its convergence, iteration, or wall-clock limit. For every run, we recorded the objective value, feasibility status, measured runtime, termination condition, seed, and algorithm settings. The best feasible Gurobi incumbent obtained within the time limit, Z_N^{MILP}, was used as an empirical reference and not as a proven optimum.

4.4 Optimality Status and Evaluation Metrics

The exact solver could not certify global optimality within the allocated limits, and its LP relaxation provided an exceptionally weak lower bound (MIP gaps exceeding 96%); all reported MILP results are therefore the best

feasible solutions found before termination. The primary quality metric is the relative deviation from the MILP incumbent, $\Delta_{\text{rel}} = (Z_{\text{alg}} - Z^{\text{MILP}})/Z^{\text{MILP}} \times 100\%$; it is not an optimality gap, since the incumbent is not proven optimal, and positive values indicate a higher (inferior) routing cost.

5 Computational Results and Discussion

5.1 Performance and Scalability

The primary objective is to evaluate algorithmic performance under strict, symmetric time budgets. Table 1 reports the objective values, runtimes, and relative deviations (Δ_{rel}) from the MILP incumbent across all instance scales.

	$N = 10$			$N = 25$			$N = 50$		
Algorithm	Z	t (s)	Δ_{rel} (%)	Z	t (s)	Δ_{rel} (%)	Z	t (s)	Δ_{rel} (%)
MILP (Gurobi)	1058.0	60.0	0.000	3194.0	150.0	0.000	5402.5	450.0	0.000
Genetic Algorithm	**1059.0**	60.1	**0.095**	**3275.7**	150.6	**2.556**	5673.5	454.4	5.016
Simulated Annealing	1064.3	60.1	0.591	3297.3	150.6	3.234	**5668.6**	450.9	**4.925**
Lagrangian Relax.	1068.1	60.4	0.955	3299.5	151.4	3.303	5684.0	457.5	5.211

Table 1 Computational results on the C101-MTW subsets. Heuristic entries are means over ten seeded runs; the MILP row is a single run; best mean heuristic per scale in **bold**

All heuristics remain within one percent of the MILP incumbent at the smallest scale ($N = 10$), and their mean relative deviations grow with instance size (Tables 1 and 2). A Friedman test across the three heuristics, using the ten shared seeds as blocks, indicates significant ranking differences at $N = 10$ ($\chi^2 = 18.20$, $p < 0.001$) and $N = 25$ ($\chi^2 = 15.75$, $p < 0.001$), but not at $N = 50$ ($\chi^2 = 4.73$, $p = 0.094$). Pairwise comparisons refine this picture: the Genetic Algorithm (GA) attains the lowest mean deviation at $N = 10$ and $N = 25$, with a seed-paired advantage over Simulated Annealing (SA) of 5.3 objective units at $N = 10$ (95% CI $[3.7, 6.8]$, Wilcoxon signed-rank $p = 0.002$) and 21.7 units at $N = 25$ (CI $[7.7, 35.6]$, $p = 0.010$). At $N = 50$, however, the paired GA–SA difference is small relative to its variability (4.9 units, CI $[-21.1, 30.9]$, $p = 0.61$), so the two metaheuristics appear statistically indistinguishable at that scale, and many runs of both methods terminate at the same objective value as the Lagrangian fallback solution. With ten paired runs per size, these comparisons should be read through their effect sizes and confidence intervals rather than as universal rankings.

5.2 Algorithm Failure Modes and Consistency

To evaluate performance consistency, Table 2 reports per-size descriptive statistics of the ten seeded objective values for each stochastic method. The three instance sizes are reported separately and are not treated as independent statistical replications, since the subsets are nested within a single instance. GA shows the largest seed-to-seed variability at $N = 25$, whereas at $N = 50$ the two metaheuristics exhibit similar dispersion; both have coefficients of variation below 0.6% at every size. The Lagrangian procedure has zero variance at $N \geq 25$ because every run exhausted its budget during route-pool generation and returned the same deterministic nearest-neighbor fallback solution.

	$N = 10$			$N = 25$			$N = 50$		
Algorithm	Δ_{rel}	**SD**	**CV**	Δ_{rel}	**SD**	**CV**	Δ_{rel}	**SD**	**CV**
Genetic Algorithm	**0.095**	0.82	0.08	**2.556**	17.03	0.52	5.016	23.22	0.41
Simulated Annealing	0.591	1.87	0.18	3.234	6.45	0.20	**4.925**	23.61	0.42
Lagrangian Relax.	0.955	1.20	0.11	3.303	0.00	0.00	5.211	0.00	0.00

Table 2 Per-size statistics over ten seeded runs: mean Δ_{rel} (%), standard deviation of Z, and coefficient of variation (%)

The recorded search-depth diagnostics offer a plausible explanation for these patterns. At $N = 10$, the GA completes between 80 and 2,555 generations across seeds and SA completes roughly 400 cooling iterations, so both methods search far beyond their constructive initial solutions. As the per-candidate evaluation cost grows with N, the same budgets admit only 6–10 GA generations at $N = 25$ and exactly 3 at $N = 50$, while SA's completed iterations fall to roughly 110–140. With so little effective search at the largest size, both metaheuristics remain

close to their constructive initialization, which explains why many $N = 50$ runs of GA and SA coincide exactly with the objective of the Lagrangian nearest-neighbor fallback. These observations suggest that, under equal cold-start budgets, the binding factor at larger sizes is the effective search depth a method achieves per second of wall-clock time, rather than its asymptotic design.

The restricted route-pool procedure maintains a feasible primal incumbent using singleton routes and a nearest-neighbor split. Consequently, termination during route-pool generation, restricted-master optimization, or dual updating does not imply that the instance is infeasible. At $N = 10$ the procedure completes its full 120-sequence pool and performs 11–98 subgradient iterations, whereas at $N = 25$ and $N = 50$ it completes only 73–79 and 41–42 pool sequences, respectively, and never reaches the dual phase. We report the best feasible incumbent available at termination together with the termination phase. Restricted-master dual values are reported separately and are not interpreted as lower bounds for the original VRP-MSTW because the route pool is incomplete. Finally, the MILP incumbents carry essentially no time-window penalty (zero at $N \in \{10, 25\}$ and 0.5 objective units at $N = 50$), yet Gurobi's reported MIP gaps (97.0%, 97.6%, and 97.2% for $N = 10$, 25, and 50) underscore that the LP relaxation provides an exceptionally weak lower bound. Consequently, all heuristic deviations in this study are measured against an unproven reference solution that nonetheless dominated all heuristic results, and the solver's MIP gap is reported separately from those deviations.

6 Conclusion and Future Work

This study presents a controlled comparative analysis of four optimization methods applied to the Vehicle Routing Problem with Multiple Soft Time Windows (VRP-MSTW). By evaluating exact mathematical programming, dual decomposition, and population- and trajectory-based metaheuristics under a shared, cold-start wall-clock protocol with ten seeded replications per stochastic method, this research provides size-dependent evidence on solver behavior for nested subsets of the C101-MTW instance. Because the evidence comes from a single clustered instance family, the conclusions are restricted to these implementations and subsets and are not claimed to generalize to random (R) or mixed (RC) Solomon-type instances.

The computational results do not support a single universal ranking of the heuristics. The Mixed-Integer Linear Program solver produced the best feasible solution at every tested scale ($N \in \{10, 25, 50\}$), but could not certify global optimality within the allotted budgets, and its LP-based lower bounds were weak. Among the heuristics, the Genetic Algorithm attained the lowest mean deviation from the MILP incumbent at $N = 10$ and $N = 25$, with statistically significant seed-paired advantages over Simulated Annealing, whereas at $N = 50$ the two metaheuristics were statistically indistinguishable. The restricted route-pool Lagrangian procedure exhausted its budget during route-pool generation at $N \geq 25$ and returned its deterministic constructive fallback, so it is interpreted through its feasible incumbent and termination phase rather than as a complete exact decomposition method. The recorded search-depth diagnostics suggest that, under equal cold-start budgets, the practically decisive property at larger sizes is how much effective search a method completes per second, a consideration that may be as important for method selection as nominal algorithmic sophistication.

Building upon these findings, several promising avenues for future research emerge. A primary objective is scaling the evaluation to the full 100-customer Solomon instances to further stress-test heuristic scalability. To make this dual-decomposition approach competitive at larger scales, future work should transition from the static route pool to a full branch-and-price framework with exact column generation. Additionally, there is significant potential in hybridizing the search algorithms—such as embedding population-based diversification directly into the Simulated Annealing framework. Finally, transitioning from deterministic models to those incorporating stochastic travel times would be an important step toward evaluating the operational robustness of these routing strategies in true real-world logistics environments.

Acknowledgements

This work was supported by the Czech Science Foundation (GAČR) project No. 25-16870S. The research was also supported by the Internal Grant Agency of Tomas Bata University in Zlín, under project number IGA/CebiaTech/2026/004.

References

Belhaiza S, Hansen P, Laporte G (2014) A hybrid variable neighborhood tabu search heuristic for the vehicle routing problem with multiple time windows. Computers & Operations Research 52:269–281. `https://doi.`

org/10.1016/j.cor.2013.08.010

Belhaiza S, M'Hallah R, Ben Brahim G, et al (2019) Three multi-start data-driven evolutionary heuristics for the vehicle routing problem with multiple time windows. Journal of Heuristics 25(3):485–515. https://doi.org/10.1007/s10732-019-09412-1

Braekers K, Ramaekers K, Van Nieuwenhuyse I (2016) The vehicle routing problem: State of the art classification and review. Computers & Industrial Engineering 99:300–313. https://doi.org/10.1016/j.cie.2015.12.007

Fisher ML (1981) The lagrangian relaxation method for solving integer programming problems. Management Science 27(1):1–18. https://doi.org/10.1287/mnsc.27.1.1

Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA

Gurobi Optimization, LLC (2023) Gurobi Optimizer Reference Manual. https://www.gurobi.com

Gülmez B, Emmerich M, Fan Y (2025) Multiple time window vrp dataset. https://doi.org/10.5281/zenodo.17515543

Hoogeboom M, Dullaert W, Lai D, et al (2020) Efficient neighborhood evaluations for the vehicle routing problem with multiple time windows. Transportation Science 54(2):400–416. https://doi.org/10.1287/trsc.2019.0912

Kendall G, Bai R, Błażewicz J, et al (2016) Good laboratory practice for optimization research. Journal of the Operational Research Society 67(4):676–689. https://doi.org/10.1057/jors.2015.77

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680. https://doi.org/10.1126/science.220.4598.671

Kourank Beheshti A, Hejazi SR, Alinaghian M (2015) The vehicle routing problem with multiple prioritized time windows: A case study. Computers & Industrial Engineering 90:402–413. https://doi.org/10.1016/j.cie.2015.10.005

Mouthuy S, Massen F, Deville Y, et al (2015) A multistage very large-scale neighborhood search for the vehicle routing problem with soft time windows. Transportation Science 49(2):223–238. https://doi.org/10.1287/trsc.2014.0558

Pessoa A, Sadykov R, Uchoa E, et al (2020) A generic exact solver for vehicle routing and related problems. Mathematical Programming 183(1–2):483–523. https://doi.org/10.1007/s10107-020-01523-z

Savelsbergh M, Van Woensel T (2016) 50th anniversary invited article—city logistics: Challenges and opportunities. Transportation Science 50(2):579–590. https://doi.org/10.1287/trsc.2016.0675

Solomon MM (1987) Algorithms for the vehicle routing and scheduling problems with time window constraints. Operations Research 35(2):254–265. https://doi.org/10.1287/opre.35.2.254

Taş D, Gendreau M, Dellaert N, et al (2014) Vehicle routing with soft time windows and stochastic travel times: A column generation and branch-and-price solution approach. European Journal of Operational Research 236(3):789–799. https://doi.org/10.1016/j.ejor.2013.05.024

Vidal T (2022) Hybrid genetic search for the CVRP: Open-source implementation and SWAP* neighborhood. Computers & Operations Research 140:105643. https://doi.org/10.1016/j.cor.2021.105643

Firm level evidence on investor sentiment and cross-sectional stock market characteristics

Bahate Maidiya[1]

Abstract. This study examines the relationship between investor sentiment and stock characteristics, focusing on returns, volatility, and trading volume across firm size and industry dimensions. Using weekly data from 2000 to 2025 and firm-level OLS regressions, we analyze Russell 1000 and 2000 constituents classified by GICS sectors. The results reveal significant heterogeneity: sentiment effects vary systematically across sectors and between big and small size firms. The findings indicate that sentiment influences are structurally distributed rather than uniform, highlighting the importance of considering cross-sectional market characteristics.

Keywords: investor sentiment, behavioral finance, Russell 1000, Russell 2000

JEL Classification: G12, G14, G41

1 Introduction

Traditional finance treats markets as disciplined information processors: prices, under the efficient market hypothesis, absorb publicly available information almost instantaneously. Real markets, however, behave less neatly. Waves of optimism and anxiety repeatedly leave fingerprints on asset prices, a pattern documented throughout the behavioral finance literature. Investor sentiment, typically defined as expectations detached from observable fundamentals, has become a widely used explanation for anomalies in stock pricing and cross-sectional return patterns (Baker et al., 2007; Azeredo et al., 2024).

Evidence linking sentiment to market outcomes is extensive, though rarely uniform. Brown and Cliff (2005) show that periods of elevated optimism often precede weaker subsequent returns, consistent with temporary overpricing that later retreats toward fundamentals. Yang et al. (2017) reach a different conclusion, finding that stronger sentiment coincides with higher short-term returns, especially among smaller and more volatile firms where speculative trading faces fewer constraints. Sentiment also appears in measures beyond returns. Using Chinese A-share data, Jiang et al. (2021) report that higher sentiment accompanies greater return volatility, while Pan et al. (2022) identify a close connection between sentiment fluctuations and trading activity during periods of market stress. The effect is not evenly distributed: small-capitalization stocks respond more sharply to sentiment shifts (McGurk et al., 2020), and sector-level reactions vary according to differences in investor composition and information transparency (Chen et al., 2013).

What emerges from this literature is not a coherent map but a collection of isolated coordinates. Most studies examine a single outcome variable or one structural dimension. Comparatively little research evaluates returns, volatility, and trading volume simultaneously within a unified cross-sectional framework spanning both size and industry. The broader distribution of sentiment effects therefore remains only partially understood. This paper addresses that omission. Drawing on Russell 1000 and Russell 2000 constituents grouped into 11 GICS sectors, we examine how sentiment shapes the three outcome variables across the size spectrum and industry classifications. Sentiment does not move through the market uniformly: its effects vary across stock characteristics and follow identifiable cross-sectional patterns tied to firm size and sector affiliation.

The remainder is organized as follows: Section 2 describes the data and methodology; Section 3 presents the main findings; Section 4 concludes.

2 Data and methodology

Investor sentiment is measured using the AAII Sentiment Survey, a weekly poll conducted since 1987 in which individual investors classify their six-month market outlook as bullish, neutral, or bearish. The sentiment indicator is constructed as the bullish-minus-bearish percentage spread. Over the sample period from January 2000 to

[1] VSB – Technical University of Ostrava, Department of Finance, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic, bahate.maidiya@vsb.cz.

December 2024 (approximately 1,300 weekly observations), this spread averaged 0.065 with a standard deviation of 0.184, ranging from −0.54 to 0.63.

The stock universe draws on Russell 1000 and Russell 2000 constituents, representing large and small-capitalization firms respectively. All stocks are classified into 11 GICS sectors. Price and volume data are sourced from Yahoo Finance using daily adjusted closing prices. Stocks with fewer than 30 weekly observations are excluded, yielding a final sample of 2,119 firms: 851 from the Russell 1000 and 1,268 from the Russell 2000. Figure 1 shows that the sectoral distribution across both groups is largely similar across the two size groups.

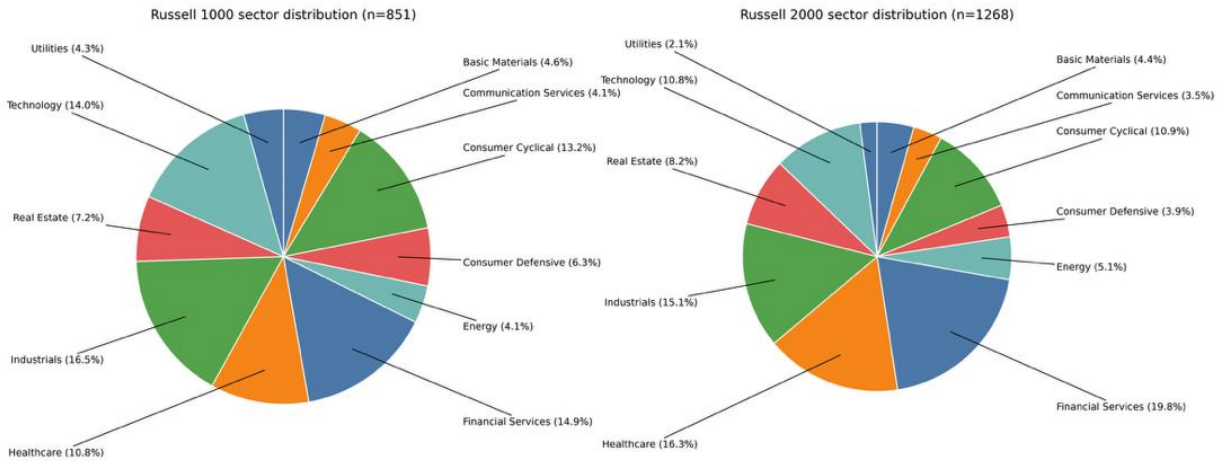


Figure 1 Sector distribution of sample firms

Since the AAII survey is released each Thursday, all data are converted to weekly frequency with each week ending on Thursday. Three market characteristic variables are constructed. Weekly returns are defined as the percentage change in adjusted closing prices between consecutive Thursdays. Weekly volatility is estimated via a GARCH (1,1) model fitted to each stock's daily return series:

$$R_{i,t} = \mu_i + \sigma_{i,t} \cdot \epsilon_{i,t} \tag{1}$$

$$\sigma_{i,t}^2 = \omega_i + a_i \cdot \sigma_{i,t-1}^2 + b_i \cdot \epsilon_{i,t-1}^2 \tag{2}$$

where $\omega_i > 0$, $a_i \geqslant 0$, $b_i \geqslant 0$ and $a_i + b_i < 1$ to ensure a positive and stationary conditional variance. The resulting daily conditional variances are summed within each trading week to obtain the weekly volatility measure. Weekly dollar volume is constructed as the sum of daily closing price multiplied by traded shares within each trading week.

For each stock i, the following model is estimated:

$$E(y_{i,k,t}) = \alpha_{i,k} + \beta_{i,k} \cdot Sentiment_t \tag{3}$$

where $y_{i,k,t}$ denotes outcome variable k for stock i in week t, with $k \in \{\text{return}, \text{volatility}, \text{volume}\}$. $Sentiment_t$ is the AAII spread from the prior week. Parameters are estimated by OLS with Newey-West robust standard errors using a lag of four weeks to account for serial correlation in weekly return series. The resulting stock level coefficients $\alpha_{i,k}$ and $\beta_{i,k}$ are then compared across size groups and GICS sectors using box plots and dumbbell plots to characterize the cross-sectional distribution of sentiment effects.

3 Result

Across the three market outcomes, investor sentiment displays different explanatory power. For stock returns, Table 1 shows that significant slope coefficients remain below 10% in every sector and size group. Intercepts are more frequently significant, particularly among big stocks where eight of eleven sectors exceed 60%. This pattern suggests that average return levels differ meaningfully across firms and sectors, but week ahead AAII sentiment contains little systematic signal for individual return variation. Short-horizon returns are likely dominated by firm-specific information shocks such as earnings announcements and other idiosyncratic events, while the survey based AAII spread cannot separate sentiment signals from market noise (Cai et al., 2024).

The volatility and volume columns of Table 1 tell a different story. Intercept significance reaches 100% universally across both variables, all sectors, and both size groups. For volatility slopes, big stocks average 79.834% significance and small stocks 73.830%, with seven big stock sectors exceeding 80%, led by Financial Services, Basic Materials, Energy, Real Estate, and Utilities. Healthcare and Technology record the lowest proportions in both size groups, yet close to half or more of their stocks still show significant slopes. For volume slopes, big stocks proportions remain comparably high at an average of 79.310%, while small stocks proportions fall to 54.881%, producing a wider size gap than observed in the volatility results.

Sector	Size Group	Return		Volatility		Volume	
		Intercept	Slope	Intercept	Slope	Intercept	Slope
Basic Materials	BIG	58.974%	0.000%	100%	94.872%	100%	84.615%
	SMALL	46.429%	3.571%	100%	82.143%	100%	58.929%
Communication Services	BIG	40.000%	0.000%	100%	65.714%	100%	57.143%
	SMALL	9.091%	4.545%	100%	72.727%	100%	40.909%
Consumer Cyclical	BIG	66.071%	2.679%	100%	85.714%	100%	76.786%
	SMALL	25.362%	2.899%	100%	84.783%	100%	60.870%
Consumer Defensive	BIG	62.963%	0.000%	100%	68.519%	100%	83.333%
	SMALL	32.653%	4.082%	100%	73.469%	100%	59.184%
Energy	BIG	54.286%	2.857%	100%	94.286%	100%	80.000%
	SMALL	18.462%	0.000%	100%	81.538%	100%	43.077%
Financial Services	BIG	66.929%	1.575%	100%	96.063%	100%	80.315%
	SMALL	40.637%	4.781%	100%	76.892%	100%	64.940%
Healthcare	BIG	70.652%	5.435%	100%	51.087%	100%	84.783%
	SMALL	18.841%	6.280%	100%	44.928%	100%	33.333%
Industrials	BIG	77.857%	0.000%	100%	87.143%	100%	78.571%
	SMALL	36.649%	5.236%	100%	68.063%	100%	53.927%
Real Estate	BIG	60.656%	0.000%	100%	91.803%	100%	93.443%
	SMALL	19.231%	2.885%	100%	87.500%	100%	62.500%
Technology	BIG	66.387%	2.521%	100%	53.782%	100%	58.824%
	SMALL	28.467%	0.730%	100%	55.474%	100%	45.255%
Utilities	BIG	78.378%	5.405%	100%	89.189%	100%	94.595%
	SMALL	73.077%	7.692%	100%	84.615%	100%	80.769%

Table 1 Significant proportion of intercept and slope for three stock characteristics distribution

Figure 2 presents the full distribution of volatility regression coefficients across sectors and size groups. Intercept medians for big stocks cluster tightly around 0.02 across most sectors, with Energy and Technology at the higher end and Utilities at the lower end. Small stocks show markedly higher intercept levels in every sector, with Healthcare (median 0.101) and Energy (0.079) recording the largest absolute values. Panel B confirms this gap is universal: small stock intercepts exceed big stock intercepts across all eleven sectors, with the widest separations in Healthcare (difference 0.080) and Energy (0.053), reflecting the structurally higher baseline volatility of smaller firms shaped by thinner trading and greater information uncertainty.

Slope medians are negative across all sectors and both size groups, indicating that higher sentiment systematically precedes lower return volatility in the following week. Small stocks consistently register more negative slopes than big stocks, with Energy (small median -0.041, big -0.015) and Real Estate (small -0.035, big -0.015) showing the largest size gaps, a pattern visible across all eleven sectors in Panel B. This asymmetry aligns with the behavioral finance prediction that smaller firms, facing greater information frictions and weaker arbitrage conditions, are more sensitive to sentiment-driven shifts in market expectations (Baker et al., 2007). Energy and Real Estate stand out as the most sentiment responsive sectors, consistent with prior evidence that these industries are particularly exposed to macroeconomic uncertainty and risk transmission channels (Jawadi et al., 2024; Mensi et al., 2023).

Figure 3 presents the coefficient distribution for trading volume. The intercept picture reverses: big stocks dominate in every sector, with Technology (median 5.51×10^8) and Energy (5.28×10^8) at the top and Real Estate (1.88×10^8) and Industrials (2.55×10^8) at the lower end. Small stock intercepts are compressed near zero across the board, reflecting the orders-of-magnitude difference in dollar volume between the two size groups. Panel B confirms the big-over-small intercept gap holds universally, with Technology, Energy, and Communication Services showing the widest separations.

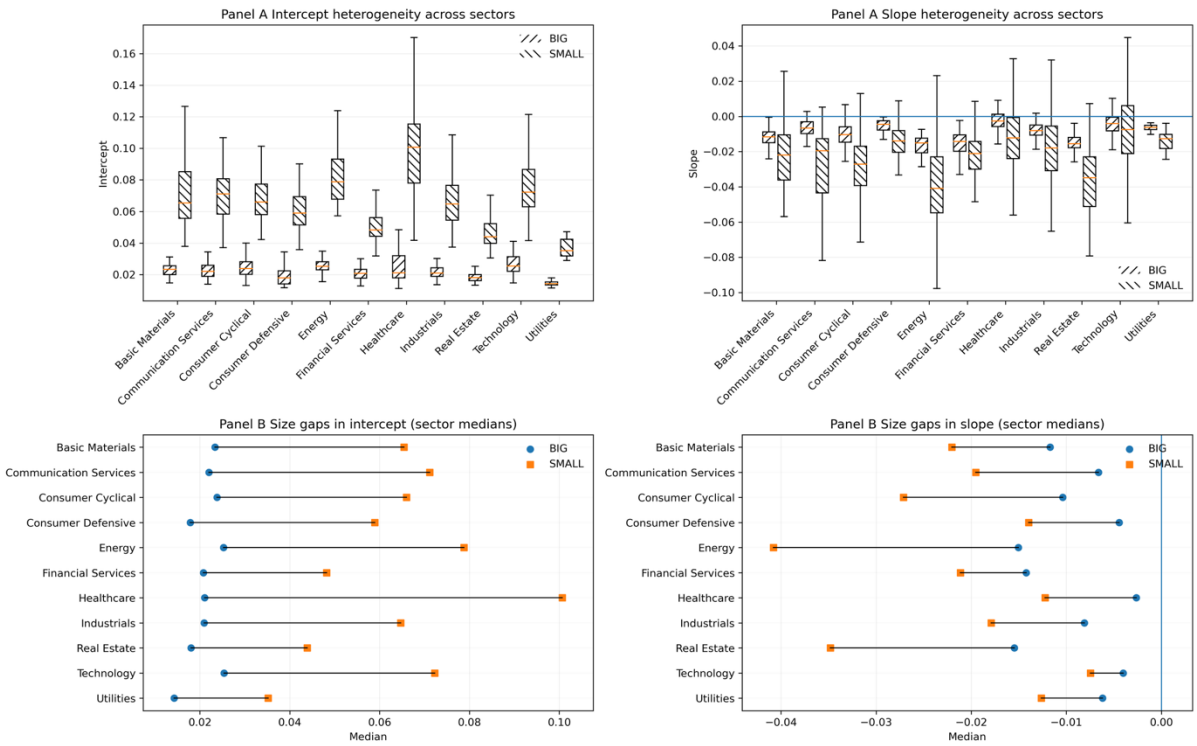


Figure 2 Investor sentiment effects to volatility across sectors and firm size

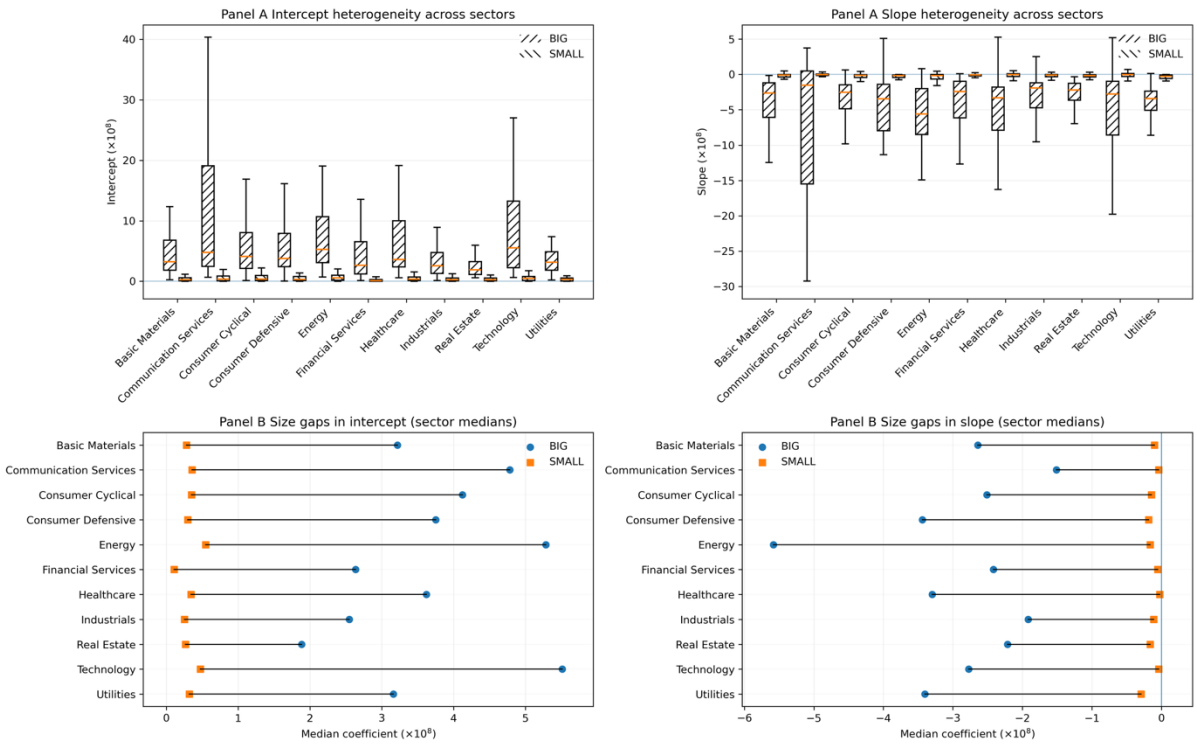


Figure 3 Investor sentiment effects to trading volume across sectors and firm size

Slope medians are also negative across all sectors and both size groups, but the size pattern runs opposite to that of volatility: big stocks exhibit more negative slopes in every sector. Energy records the starkest contrast, with big stock medians near -5.58×10^8 against small stock medians near -0.16×10^8. Because the AAII spread captures aggregate market sentiment, its signal transmits most forcefully through big stock index constituents with high price synchronicity (Kim et al., 2021). Energy again shows the strongest size asymmetry, reinforcing a pattern that runs consistently across both figures: sentiment effects in this sector are not only stronger in absolute terms but also more structurally differentiated by firm size.

The results reveal a clear hierarchy. Sentiment explains little of next week returns but is closely related to both volatility and trading volume. These effects are not uniform. Intercepts show that small stocks carry structurally higher baseline volatility while big stocks command far larger baseline dollar volume, differences that reflect the underlying information environment and market depth of each size group. At the slope level, volatility sensitivity is stronger among small stocks while volume sensitivity is stronger among big stocks, a size asymmetry that holds across all eleven sectors. Sector patterns are also visible, particularly in Energy sector. Sentiment operates less as a universal pricing force than as a cross-sectional mechanism whose influence depends on the market characteristic being examined, the firm size group, and the sector in which the firm operates.

4 Conclusion

This paper asks a question: does investor sentiment affect all stocks equally? Examining 2,119 U.S. firms across 11 GICS sectors, the answer is no. Sentiment carries negligible predictive power for individual stock returns, yet it is systematically and negatively associated with both volatility and trading volume, relationships that hold across sectors while varying in strength by firm size. Small stocks absorb sentiment shocks more acutely in volatility, while big stocks show stronger volume contraction, a size asymmetry that cuts across the market cross-section. By examining three outcome variables simultaneously within a unified framework, this paper offers more structured empirical evidence on how sentiment distributes its effects across the stock market.

Two limitations deserve mention. The analysis rests on a single survey-based sentiment measure, and the linear specification may not capture state dependent dynamics that emerge during market stress or regime shifts. Whether the structural patterns documented here hold across international markets, alternative sentiment proxies, or nonlinear model specifications remains an open question worth pursuing.

Acknowledgements

The research was financially supported by the SGS project no. SP2026/003 of VSB – Technical University of Ostrava.

References

Azeredo da Silveira R, Sung Y, Woodford M (2024) Optimally imprecise memory and biased forecasts. American Economic Review 114:3075-3118. https://doi.org/10.1257/aer.20201806

Baker M, Wurgler J (2007) Investor sentiment in the stock market. Journal of Economic Perspectives 21:129-151. https://doi.org/10.1257/jep.21.2.129

Brown G W, Cliff M T (2005) Investor sentiment and asset valuation. The Journal of Business 78:405-440. https://doi.org/10.1086/427633

Cai Y, Tang Z, Chen Y (2024) Can real-time investor sentiment help predict the high-frequency stock returns? Evidence from a mixed-frequency-rolling decomposition forecasting method. The North American Journal of Economics and Finance 72:102147. https://doi.org/10.1016/j.najef.2024.102147

Chen M P, Chen P F, Lee C C (2013) Asymmetric effects of investor sentiment on industry stock returns: panel data evidence. Emerging Markets Review 14:35-54. https://doi.org/10.1016/j.ememar.2012.11.001

Jawadi F, Bourghelle D, Rozin P, Cheffou A I, Uddin G S (2024) Sentiment and energy price volatility: a nonlinear high frequency analysis. Energy Economics 133:107465. https://doi.org/10.1016/j.eneco.2024.107465

Jiang S, Jin X (2021) Effects of investor sentiment on stock return volatility: a spatio-temporal dynamic panel model. Economic Modelling 97:298-306. https://doi.org/10.1016/j.econmod.2020.04.002

Kim K, Ryu D (2021) Term structure of sentiment effect on investor trading behavior. Finance Research Letters 43:102005. https://doi.org/10.1016/j.frl.2021.102005

Mensi W, Gubareva M, Teplova T, Kang S H (2023) Spillover and connectedness among G7 real estate investment trusts: the effects of investor sentiment and global factors. The North American Journal of Economics and Finance 66:101919. https://doi.org/10.1016/j.najef.2023.101919

McGurk Z, Nowak A, Hall J C (2020) Stock returns and investor sentiment: textual analysis and social media. Journal of Economics and Finance 44:458-485. https://doi.org/10.1007/s12197-019-09494-4

Pan Y, Hou L, Pan X (2022) Interplay between stock trading volume, policy, and investor sentiment: a multifractal approach. Physica A: Statistical Mechanics and its Applications 603:127706. https://doi.org/10.1016/j.physa.2022.127706

Yang H, Ryu D, Ryu D (2017) Investor sentiment, asset returns and firm characteristics: evidence from the Korean stock market. Investment Analysts Journal 46:132-147. https://doi.org/10.1080/10293523.2016.1277850

Counting passengers in public transport buses using door optical sensors

Tomáš Majer[1], Dalibor Gonda[2], Ivana Ondrušová[3]

Abstract. In many Slovak cities, public bus transport was only boarded through the front doors and tickets were purchased from the driver. Based on data on tickets sold, it was possible to calculate the load of trips with great accuracy, and there were a number of models for calculating the Origin-Destination (OD) matrix. Nowadays, more and more passengers are using prepaid time tickets. Therefore, measuring transport demand can no longer rely only on the sale of single tickets and other methods need to be prepared. One such method is to place optical sensors in the vehicle doors that will count the number of passengers. However, some errors can occur, which results in some subsequent problems. For example, it may happen that the vehicle has a negative number of passengers or that not all passengers get off at the final s top. In our paper, we will present a method for best correcting passenger counting errors to ultimately refine the estimate of the number of passengers in the OD matrix.

Keywords: public transport, Origin-Destination matrix, optical sensors, mathematical programing

JEL Classification: C61, R40

1 Introduction

The increase in the intensity of individual car transport and the associated problems such as traffic jams and lack of parking spaces clearly call for improving the quality and efficiency of public transport. In order for public transport to be able to satisfy the demands of passengers, these demands need to be quantified as best and most accurately as possible. The demand for transport services is most often expressed in the form of an OD matrix (origin – destination), the individual elements of which (T_{ij}) express the number of passengers from zone or stop i to zone or stop j.

The best, but also the most expensive, way to obtain such a matrix is considered to be a complete transport survey. In addition to the high price and the large number of workers (enumerators) deployed at all stops, the disadvantage is that the survey can only measure demand where there is at least some public transport service available.

Another option for measuring demand for public transport services is a survey. This allows for capturing emerging demand between areas where no public transport service has been operating so far. The problem with surveys is that only a small part of the travelling public participates in them and it is also very difficult to process the survey forms. Alternatively, these two options can be combined in a suitable way.

The OD matrix can also be calculated using a partial traffic survey (when only the total number of passengers boarding and alighting at individual stops is measured without information about the direction to or from where they are traveling), historical demand data, and some other parameters of the transport system, such as the distance between stops.

Several methods for estimating the elements of the OD matrix such as growth and gravity models can be found in the paper (Majer and Palúch, 2015).

Since the OD matrix focuses more on the characteristics and activities of transport between the origin–destination area, and due to the great progress and development of software technologies, the task of data collection has become extensive, the OD matrix has many applications in transport planning (Abdulsalam and Asmael, 2025).

In the recent past, it was possible to estimate the number of boarding passengers based on the number of tickets

[1] University of Žilina, Faculty of management science and informatics, Department of mathematical methods and operational research, Univerzitná 8215/1, 010 26 Žilina, tomas.majer@fri.uniza.sk
[2] University of Žilina, Faculty of management science and informatics, Department of mathematical methods and operational research, Univerzitná 8215/1, 010 26 Žilina, dalibor.gonda@fri.uniza.sk
[3] University of Žilina, Faculty of management science and informatics, Department of mathematical methods and operational research, Univerzitná 8215/1, 010 26 Žilina, ondrusova24@stud.uniza.sk

marked in the vehicle, as each passenger had this obligation under the transport regulations. Since most of these tickets were recorded electronically on a smart transport card, many procedures were developed to fairly accurately predict the place of exit from the vehicle of such a passenger based on the serial number of the card.

As integrated transport systems are currently being built and prepaid time tickets are being used extensively without marking individual journeys, it is necessary to look for other options for counting passengers. One option is to have optical sensors installed in every door of all vehicles that will record the passage of each person in both directions in and out of the vehicle.

Automatic optical sensors can count one passenger more than once, or if there are many passengers boarding at one stop, some passengers may be overlooked by the sensor. This leads to an inequality between the number of boarding and alighting passangers for one trip. Such discrepancies in the data are not suitable if we want to use the data for other purposes, namely for using methods for estimating OD matrix.

To eliminate this problem, the principle of the maximum likelihood estimation method (MLE) can be used.

Thanks to our cooperation with the Žilina City Transport Company (DPMŽ for *"Dopravný podnik mesta Žilina"*), homepage (DPMŽ, 2026)), we obtained data exported from their information system. This includes timetable data in the standardized GTFS (General Transit Feed Specification) format, (MobilityData, 2026)), as well as the detailed number of boardings and exits on individual trips and stops counted by optical sensors at every bus door.

This datasets are used to estimate OD marix of passengers using public transport in our city Žilina.

2 Maximum likelihood estimation (MLE) method

It is an important tool used in statistics, where the principle of this method is to find an estimate of the parameter θ (theta), or a vector of parameters, that maximizes the probability that the observed values come from the assumed probability distribution. In other words, we are trying to find a value of θ for which the probability that the observed values come from the assumed distribution is maximum. The estimate therefore tried to fit the observed data as closely as possible, which is logical when we assume that the data represent the only source of information about our unknown parameter (Wikipedia, 2026).

Because the more precise characteristics of the sensors are not known, each record can be erroneous with equal probability, the gaps between errors will have an exponential probability distribution, and the frequency of errors over a given period (given by the number of records) will have a Poisson probability distribution.

However, again, the entire success of the maximum likelihood estimation method depends on the correct specification of the probability behavior, i.e. the choice of a specific probability distribution (Wikipedia, 2026).

3 Application in practice

As already explained, the numbers of passenger boardings and alightings at individual stops are Poisson random variables.

$$N_i \sim \text{Poisson}(\lambda_i), \quad V_i \sim \text{Poisson}(\mu_i) \tag{1}$$

where $i = 1, \ldots, n$ is index of the stop within one trip, N_i is actual number of arrivals at stop i (Poisson random variable), V_i is actual number of exits at stop i (Poisson random variable), $\lambda_i \geq 0$ is intensity of arrivals at stop i, $\mu_i \geq 0$ is intensity of exits at stop i.

Due to errors in estimates and noise in the data, negative values are also discovered, which can be eliminated within the parameter estimation using the non-negative MLE estimation given by the relations:

$$\hat{\lambda}_i = \max(0, \tilde{N}_i), \quad \hat{\mu}_i = \max(0, \tilde{V}_i) \tag{2}$$

where $\hat{\lambda}_i$ is expected intensity of arrivals at stop i, $\hat{\mu}_i$ is expected intensity of exits at stop i, $\tilde{N}_i$ is observed number of arrivals at stop i, $\tilde{V}_i$ is observed number of exits at stop i.

The consistency of the individual trip is ensured by the global constraint of equality of total arrivals and exits, expressed by the condition:

$$\sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \mu_i \tag{3}$$

The parameter estimation itself is formulated as a problem of maximizing the log-likelihood function:

$$\ell(\lambda, \mu) = \sum_{i=1}^{n} \left(\hat{\lambda}_i \log \lambda_i - \lambda_i\right) + \sum_{i=1}^{n} \left(\hat{\mu}_i \log \mu_i - \mu_i\right) \tag{4}$$

whereby the terms independent of the optimizes parameters were omitted during the maximization. The problem is solved using the Lagrange multiplier α, by introducing the Lagrange function:

$$\mathcal{L}(\lambda, \mu, \alpha) = \ell(\lambda, \mu) + \alpha \left(\sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \mu_i \right) \tag{5}$$

The solution can be implemented by discrete iterative correction, where the global difference between the sums is gradually redistributed by units between individual stops. The procedure corresponds to the discrete implementation of the MLE solution, assuming that the stops have the same weights. The result leads to a consistent and statistically justified equalization of passengers boardings and alighting as we can see on Table 1 and Table 2.

Line: 1 Variant: B Trip: 07:32

Stop #	Stop name	Number of paper tickets	Number of smart cards	Tickets total	Passengers boarding	Passengers alighting	Passengers in vehicle
1	Matice slovenskej		5	5	15		15
2	Obchodná		2	2	4		19
3	Sv. Cyrila a Metoda	1	5	6	7		26
4	Košická, TESCO hypermarket				2	2	26
5	Predmestská				1	4	23
6	Veľká okružná, AUPARK				1	7	17
7	Polícia					7	10
8	Hálkova					6	4
9	Závodská					3	1
10	Kvačalova, DPMŽ					2	**-1**
Total:		1	12	13	**30**	**31**	

Table 1 Inequality between boarding and alighting passengers

Line: 1 Variant: B Trip: 07:32

Stop #	Stop name	Passengers boarding	Passengers alighting	Passengers in vehicle	Citizens in zone	Quality of live	Job opportunities and culture	Negatives	City zone
1	Matice slovenskej	16	0	16	17998	9	6	3	Vlčince
2	Obchodná	4	0	20	17998	6	8	2	Vlčince
3	Sv. Cyrila a Metoda	7	0	27	17998	8	7	3	Vlčince
4	Košická, TESCO hypermarket	2	2	27	2286	0	8	4	Žilina
5	Predmestská	1	4	24	8898	5	6	5	centrum
6	Veľká okružná, AUPARK	1	7	18	8898	4	9	2	centrum
7	Polícia	0	7	11	8898	2	7	4	centrum
8	Hálkova	0	6	5	8898	2	6	4	centrum
9	Závodská	0	3	2	2286	2	6	5	Žilina
10	Kvačalova, DPMŽ	0	2	**0**	2286	0	5	5	Žilina
Total: Boarding=31 Alighting=31									

Table 2 Equality after MLE correction

In the case of different weights of stops, the log-likelihood function would be weighted according to the importance of individual stops or according to the reliability of individual observations. The resulting MLE solution leads to

a redistribution proportional to the weights, with stops that have a higher weight also having a greater impact on the correction of the global constraint.

Data fixed by MLE correction will be used in gravity model for the estimation of OD matrix. As well as data collected and calculated in order to estimate travel impedance. Both are really important when it comes to estimating traffic flow. We have experimented with several opportunities for the calculation of gravity model.

4 Calculation of gravity models

For the estimation of the OD matrix, which is an output of the gravity model, the number of boarding and alightning passengers is not enough. Traffic flow is also defined by many other factors. We have decided to collect data about the number of inhabitants in each neighborhood as well as the quality of living, work and entertainment, which also involved public service offer. Traffic zones were defined according to the districts of the city of Žilina. Quality information was collected through map data and each category was scored separately. Finally, a single number was calculated that represents the overall quality of life around a given stop or transport zone. Final dataset can be seen on Table 3.

Stop name	Stop name DPMŽ	Passengers boarding	Passengers alighting	Quality of live	Citizens in zone	Zone name	Bus lines
Rázusova	Rázusova	28	27	4	8898	centrum	27,31
Bytčica	Bytčica	230	206	-5	2286	Žilina	20,22
Pietna, Krematórium	Pietna, Krematórium	7	6	-4	2286	Žilina	31
Centrálna	Centrálna	216	204	3	12420	Solinky	3
Na chotári	Na chotári	56	61	-2	939	Zástranie	27
Priehrada	Vodná nádrž Hričov	5	6	-6	2286	Žilina	29
Poštová	Poštová	452	534	13	17998	Vlčince	20,5,6
Saleziánska	Saleziánska	75	68	8	10922	Hliny	22
Na lány	Na lány	121	171	0	1844	Budatín	27
Rakové	Rakové	99	98	0	699	Rosinky	24
Košická, TESCO hypermarket	Košická, TESCO hypermarket	469	504	4	2286	Žilina	1,14,24,26,4,50
Rajecká, Solinky	Stanica Solinky	173	206	3	12420	Solinky	20,22
Oceliarska	Oceliarska	56	53	0	2286	Žilina	24,29
Kortinská	Kortinská	27	34	-1	335	Žilinská Lehota	29
Zvolenská	Zvolenská	139	203	9	17998	Vlčince	20,26
...	...	...	...	...	...	...	...

Table 3 Dataset containing information about each stop

As can be see on Table 3, first column represents the name of the stop, second column stands for number of boarding passengers, third column stands for alighting passengers, fourth column represents quality of concrete stop (also negative numbers). Fifth column represents number of inhabitants living in traffic zone, which is the sixth column, and the last column represents the list of lines that the given stop is part of.

All these factors can be counted as positive ones. However, estimating the OD matrix and determining the approximate traffic flow, the possibilities of using urban public transport need to be know. This involves information about travel time and necessary transfers when traveling from one point in the city to another.

To obtain this information, it was necessary to implement a multi-layer Dijkstra's algorithm, which first create individual layers of the graph for every public transport line and than connect it using common stops. Calculated shortest path estimates travel time and number of transfers between individual nodes of the graph. This informations can be then used for calculating traffic impedance – a fundamental part of the relation for implementing gravity model.

The result of implemented Dijkstra's algorithm can be seen in Table 4.

Trip starting stop	Trip final stop	Minutes	Transfers
...	...	...	...
Žilinská Lehota	Vranie	50	1
Žilinská Lehota	Plaváreň	45	1
Žilinská Lehota	Zábrežná	46	1
Žilinská Lehota	Zástranie	54	1
Žilinská Lehota	Závodská	16	0
Žilinská Lehota	Závodského	32	1
Žilinská Lehota	Zvolenská	50	1
Žilinská Lehota	Železničná stanica	24	0
Žilinská Lehota	Žilinská univerzita	49	1
Žilinská Lehota	Žitná	33	1
Žilinská Lehota	Veľká okružná	20	0
Žilinská univerzita	Asanačný podnik	41	1
Žilinská univerzita	Železničná stanica	18	0
Žilinská univerzita	Bagarova	33	1
Žilinská univerzita	Bánová	38	1
Žilinská univerzita	Bánovská cesta	32	1
Žilinská univerzita	Antona Bernoláka	28	1
Žilinská univerzita	Bratislavská	36	1
Žilinská univerzita	Brodňanská	41	1
Žilinská univerzita	Brodno	43	1
Žilinská univerzita	Budatínska	38	1
Žilinská univerzita	Bytčianska	41	1
...	...	...	...

Table 4 The output of Dijkstra's algorithm for counting travel time and transfers

As can be seen in Table 4, the first column represents the starting stop, second column represents the final stop. Third column points to travel time, between these two stops and last column represents number of necessary transfers.

An integral part of creating gravity models is modeling traffic impedance. In our case, focus is on travel time, the number of transfers to reach the destination, and quality of life. If a stop is not attractive enough in terms of housing, employment, and culture, it may have a negative rating. Since gravity models traditionally do not work with negative values, this factor will be modeled in the transport impedance and its coefficients will subsequently be adjusted to correctly express the results. It can be assumed that, despite longer travel times or higher number of transfers, a higher quality of life can affect transport impedance. People are willing to travel longer under better conditions because they see a certain advantage in it, and thus transport impedance is reduced. This relationship also applies in reverse, a shorter journey to a place that does not have a high quality of life has a negative impact on the traveler and increases transport impedance.

In the experimental part, it is important to correctly determine the relation for the traffic impedance. Several options for its alculation were tried. Considering the calculation time of the gravity model, but also for the best demonstration of the expected behavior of passengers, relations (6), (7), (8) were chosen:

$$R_{ij} = 4 + 1.5\,T_{ij} + 6\,P_{ij} + 3\,\log(O_{ij}) - 1.5\,Q_{ij} \tag{6}$$

$$F_{ij} = \frac{1}{R_{ij}^2} \tag{7}$$

$$G_{ij} = N_i\,V_j\,F_{ij} \tag{8}$$

where R_{ij} is traffic impedance betwwen stops i and j, T_{ij} is travel time, P_{ij} is number of transfers, O_{ij} is number

of inhabitants, Q_{ij} is average quality of life, N_i is number of boarding passengers, V_j is number of alighting passengers, G_{ij} is value of calculated gravity model.

For visualization of OD matrix to facilitate interpretation of gravity model computations heatmaps ware choosen. Result can be seen on Figure 1.

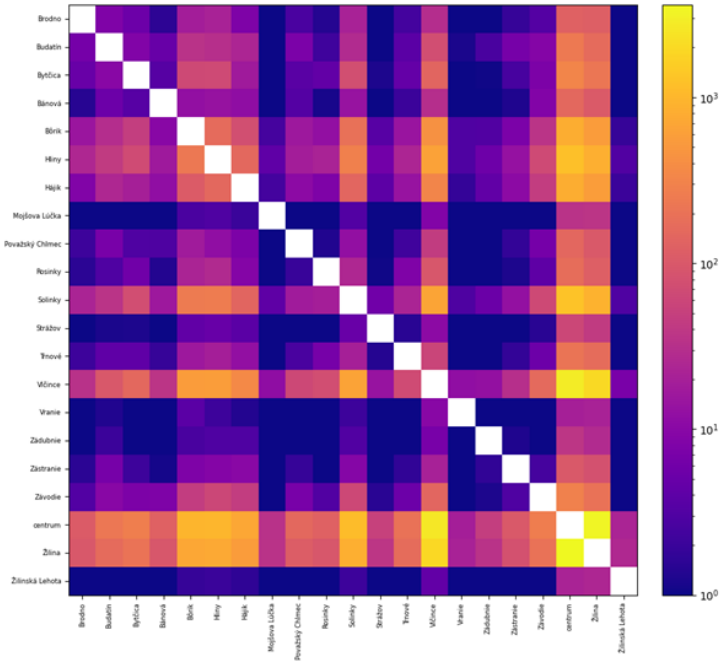


Figure 1 Final heatmap representation of OD matrix

5 Conclusion

In this paper, we discussed how data on the current timetable and the number of boarding and alighting passengers from optical sensors can be used to estimate the OD matrix. This matrix is most important input in public transport planning and depends on many technical, but also sociological and psychological factors.

Our next work will focus on creating an interactive application that, based on changing conditions, can estimate the number of passengers in the future and thus help to adapt the routes of public transport lines and the time interval of individual trips.

Acknowledgements

This paper was supported by VEGA 1/0220/25 - Regional equality in access to public rescue services provided from service centers on the transport network.

References

Majer T, Palúch S (2015) Estimation of municipal public transport demand using gravity model. In: AIESA Building Of Society Based On Knowledge. Conference Proceedings. Bratislava University of Economics and Business.

Abdulsalam R, Asmael N M (2025) Different methods for etimating the origin–destination matrix, ResearchGate, November 2025. `https://www.researchgate.net/publication/397175446_REVIEW_DIFFERENT_METHODS_FOR_ESTIMATING_THE_ORIGIN-DESTINATION_MATRIX`.

Wikipedia (2026) Maximum likelihood estimation, May 2026. `https://https://en.wikipedia.org/wiki/Maximum_likelihood_estimation`.

DPMŽ - Dopravný podnik mesta Žiliny s.r.o. (2026) About the Company, May 2026. https://www.dpmz.sk/o-spolocnosti/.

MobilityData (2026) General Transit Feed Specification (GTFS), May 2026. https://gtfs.org/documentation/overview/.

Structural Modelling of Sustainability Perception: Triangulation of Survey Data and Regional Indicators

Ladislava Míková[1], Pavlína Forstová Kuráňová[1], Václav Friedrich[1]

Abstract. This paper analyses regional and intergenerational differences in the perception of the three pillars of sustainability, environmental, social, and economic, within the context of the Czech Republic. The objective is to compare the subjective interpretations of these dimensions across various socio-demographic groups, with a specific focus on regional disparities. The methodology employs a triangulation of survey data (n = 500), analysis of regional ESG indicators, and policy documents. The structure and reliability of the proposed model are evaluated using quantitative methods of statistical modelling, specifically Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA).

The results indicate that perceptions of sustainability are heterogeneous and influenced by factors such as age, gender, and educational attainment. In regions with a traditional industrial orientation (e.g., Karlovy Vary, Ústí nad Labem, and Moravian-Silesian regions), a significant emphasis on the economic pillar was identified. Conversely, in economically advanced centres like Prague, a balanced approach toward all dimensions prevails. Comparative analysis indicated that the Venn diagram model provided a more nuanced representation of sustainability perception than the traditional Parthenon model.

Keywords: sustainability, three pillars, regional analysis, factor analysis, perception

JEL Classification: Q01, C38, R11

1 Introduction

The concept of sustainability has evolved from a purely environmental concern into a complex, multidimensional framework often referred to as the "Triple Bottom Line." According to the Brundtland Commission's definition, sustainable development is a process that meets the needs of the present without compromising the ability of future generations to meet their own needs. While this definition is globally accepted, its interpretation at the local and individual levels remains highly heterogeneous.

In the Czech Republic, historical industrial development has created regional differences in sustainability perception. This study analyses how socio-demographic and regional factors influence the prioritisation of the environmental, social, and economic pillars to support evidence-based regional policies.

2 Theoretical Framework

The evolution of sustainable development as both a scientific and policy-oriented concept has led to the establishment of a multidimensional framework that seeks to reconcile inherently conflicting objectives. The most widely recognised definition characterises sustainable development as a strategy that balances the needs of the present generation with the preservation of resources for future generations (WCED, 1987). In academic literature, this concept is frequently operationalised through the "Triple Bottom Line" approach, which posits that long-term sustainability is achievable only at the intersection of environmental integrity, social equity, and economic viability (Elkington, 1997; Milne and Gray, 2013).

Each of these dimensions carries a distinct weight in the perception of individuals. The environmental dimension focuses on the ecological carrying capacity of the planet, addressing systemic risks such as climate change and biodiversity loss. The social dimension emphasises the quality of human life, including equity, inclusion, and

[1] VSB – Technical University of Ostrava, Faculty of Economics, Department of Mathematical Methods in Economics, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic, ladislava.mikova.st@vsb.cz, pavlina.forstova.kuranova@vsb.cz, vaclav.friedrich@vsb.cz

wellbeing. The economic dimension evaluates the efficiency of resource allocation and the capacity to sustain long-term development (Purvis et al., 2019). Importantly, these dimensions are not independent; rather, they form an interconnected system in which trade-offs and synergies are continuously negotiated (Kates et al., 2005).

The subjective perception of sustainability is shaped by knowledge, personal values, and self-efficacy (Stern, 2000). Greater knowledge and awareness generally promote a more balanced understanding of sustainability, whereas personal values influence whether sustainability is perceived as a moral imperative or an economic constraint (Schwartz, 1992). Low self-efficacy often leads individuals to prioritise immediate economic security, particularly in regions undergoing structural economic transition (ČSÚ, 2024).

Furthermore, the conceptual modelling of sustainability influences its interpretation and practical application (Purvis et al., 2019). The Parthenon model depicts sustainability as three separate pillars, whereas the Venn diagram model emphasises their interdependencies and overlapping areas, providing a more comprehensive representation of sustainable development (Purvis et al., 2019; Kates et al., 2005).

In the specific context of the Czech Republic, these theoretical constructs are further shaped by regional socioeconomic characteristics (ČSÚ, 2024). Historical industrial development, particularly in coal-mining regions, has created distinct regional identities that influence how sustainability is perceived and prioritised. In such regions, the economic pillar often dominates due to concerns related to employment and economic stability. Understanding these regional and intergenerational differences is therefore essential for designing policies that are both analytically grounded and socially acceptable (ČSÚ, 2024).

2.1 Formulation of Research Hypotheses

Based on the synthesis of the theoretical frameworks and the specific socio-economic context of the Czech Republic, a set of seven research hypotheses was formulated to guide the empirical investigation. The first hypothesis (H1) posits that the majority of respondents identify with the Brundtland definition of sustainability as their primary conceptual framework, reflecting the global dominance of this paradigm. Furthermore, it is hypothesised (H2) that the Venn diagram model provides a higher resolution for distinguishing respondent attitudes and internal values than the more rigid Parthenon model.

Regarding regional distribution, the study assumes (H3) that in the majority of Czech regions, citizens perceive sustainability as a balanced equilibrium of all three pillars; however, a significant deviation is expected in specific areas. Specifically, the fourth hypothesis (H4) states that in regions with a long-standing tradition of heavy industry and mining, a significantly greater emphasis is placed on the economic pillar compared to environmental or social concerns.

The remaining hypotheses focus on socio-demographic determinants of perception. It is hypothesised (H5) that men place a greater relative emphasis on the economic dimension of sustainability, while women tend toward a more balanced perception of all three pillars. In terms of generational differences, the study assumes (H6) that younger and older generations differ significantly in their overall perception of sustainability. Finally, the seventh hypothesis (H7) suggests that individuals with lower levels of educational attainment prioritise the economic pillar, while those with higher education levels demonstrate a more holistic and balanced understanding of the sustainability triad.

3 Methodology

The research design is established as a comprehensive quantitative study utilising a multi-stage analytical framework to evaluate the perception of sustainability pillars in the Czech Republic. The primary data was gathered through a structured survey instrument administered in 2025, resulting in a dataset of 500 respondents. The sample included respondents from all fourteen administrative regions of the Czech Republic and reflected diversity in gender, age, and educational attainment. Women represented 55.4% of the sample and men 42.8%, while the remaining respondents identified otherwise or preferred not to disclose their gender. The largest age group consisted of Millennials (33.2%), followed by Xennials (28.4%), Baby Boomers (19.6%), and the Net Generation (18.8%). Regarding educational attainment, 42.4% of respondents reported university education, 33.8% secondary education with a graduation examination, 15.0% vocational education, and 8.8% primary education.

The analytical procedure is divided into three phases.

In the first phase, the suitability of the dataset for factor analysis was assessed. The Kaiser-Meyer-Olkin measure confirmed sampling adequacy (KMO = 0.882), while Bartlett's test of sphericity was statistically significant

($\chi^2(435) = 4725.554$, $p < 0.001$). The internal consistency of the measurement scales was verified using Cronbach's alpha ($\alpha = 0.826$), indicating a high level of reliability.

To reduce the dimensionality of the 30 attitudinal variables, Exploratory Factor Analysis (EFA) was conducted using Principal Component Analysis with Varimax rotation. The Kaiser-Meyer-Olkin measure of sampling adequacy reached 0.882 and Bartlett's test of sphericity was significant ($\chi^2 = 4725.554$, $p < 0.001$), confirming the suitability of the data for factor analysis. Although five factors exhibited eigenvalues greater than one according to the Kaiser criterion, inspection of the scree plot supported a three-factor solution. These three factors explained 41.72% of the total variance and were interpreted as "Knowledge and Awareness", "Personal Values", and "Self-Efficacy". The resulting factor structure was subsequently validated using Confirmatory Factor Analysis (CFA), which also enabled comparison of the Parthenon and Venn diagram models.

In the second phase, the formulated research hypotheses were tested using appropriate statistical methods. Differences between regions were analysed using Analysis of Variance (ANOVA) with post-hoc Tukey HSD tests. Socio-demographic differences were examined using independent samples t-tests and the Kruskal-Wallis test, depending on the distributional properties of the data. For the analysis of age differences, an overall sustainability score was calculated as the arithmetic mean of all 30 Likert-scale items (V12.1–V12.30). This composite score served as a summary indicator of respondents' overall sustainability perception and was subsequently used as the dependent variable in the Kruskal–Wallis analysis. In addition to statistical significance ($\alpha = 0.05$), effect sizes were calculated to assess the practical relevance of the observed differences.

In the final phase, subjective survey data were triangulated with objective regional ESG indicators obtained from the Czech Statistical Office. This integration of perception-based and structural data represents a key methodological contribution of the study, enhancing the robustness and explanatory power of the results.

All statistical analyses were conducted using standard statistical software, ensuring reproducibility and methodological transparency.

4 Results and Discussion

4.1 Validation of Sustainability Models and Conceptual Definitions

The empirical analysis began with the evaluation of the first two hypotheses concerning the conceptual grounding of the respondents. Based on the frequency analysis of preferred definitions, it was found that a significant majority of the participants (80.2 %) identified with the Brundtland Commission's definition of sustainability. This result provides strong empirical support for **H1**, confirming that the intergenerational and balanced approach remains the dominant paradigm in the Czech public consciousness.

The second stage of the analysis focused on the structural validity of the visualisation frameworks through Confirmatory Factor Analysis (CFA). When comparing the Parthenon model with the Venn diagram model, the statistical indices suggested a more suitable representation for the latter. Specifically, the Venn diagram model achieved a higher Comparative Fit Index (CFI = 0.94) and a lower Root Mean Square Error of Approximation (RMSEA = 0.052) compared to the Parthenon model (see Table 1). These findings provide support for **H2**, suggesting that respondents perceive sustainability more as a set of overlapping and interacting domains rather than as isolated pillars. This higher resolution in respondent attitudes indicates that the public is sensitive to the trade-offs between economic growth and environmental protection.

Model	CFI	RMSEA
Parthenon model	0.89	0.071
Venn diagram model	0.94	0.052

Table 1 Comparison of Model Fit Indices

According to Hair et al. (2018), CFI values above 0.90 indicate acceptable fit and values above 0.95 indicate excellent fit, while RMSEA values below 0.08 indicate acceptable fit and below 0.06 indicate good fit.

4.2 Regional Disparities and the Influence of Industrial Tradition

The core of the research focused on the geographical distribution of sustainability perceptions across the 14 regions of the Czech Republic. The results of the ANOVA tests revealed statistically significant differences ($p < 0.001$) in the prioritisation of the three pillars, thereby confirming the existence of regional "perception clusters."

In the majority of regions, particularly in the Central Bohemian and South Bohemian regions, the data showed a balanced perception where no single pillar significantly outweighed the others, which supports **H3**.

Figures 1 and 2 illustrate the regional distribution of sustainability perception under both conceptual models. The Venn diagram model provides a more differentiated classification than the Parthenon model.

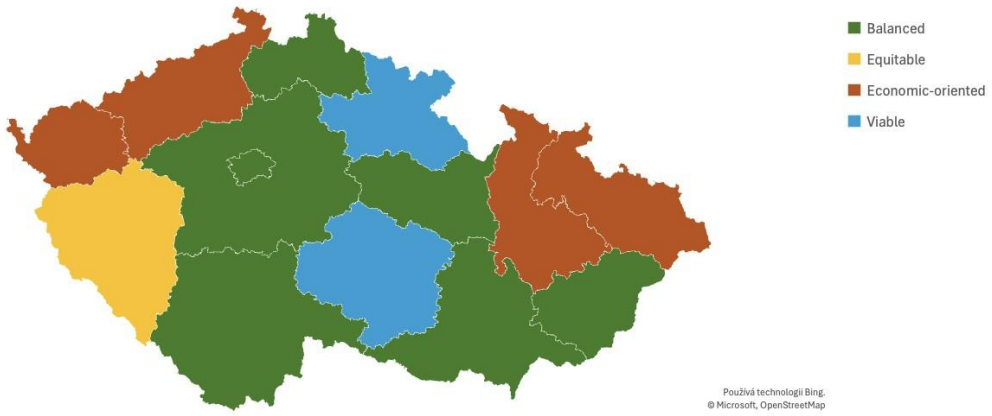


Figure 1 Regional distribution of sustainability perception based on the Venn diagram model

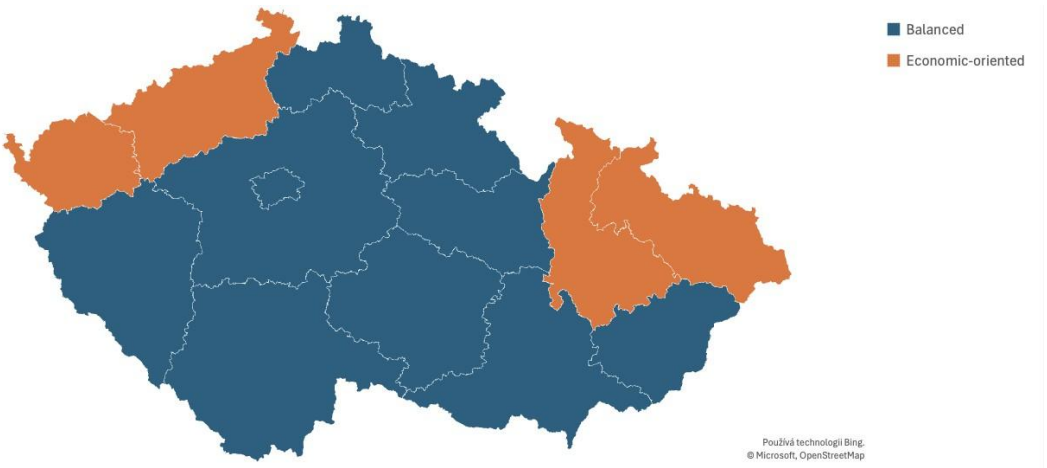


Figure 2 Regional distribution of sustainability perception based on the Parthenon model

Industrial regions such as the Moravian-Silesian and Ústí nad Labem regions are clearly associated with an economically oriented perception, whereas central and southern regions tend to exhibit a more balanced profile. However, a distinct deviation was identified in regions characterised by a high concentration of heavy industry and a history of coal mining, specifically the Moravian-Silesian, Ústí nad Labem, and Karlovy Vary regions. In these areas, the economic pillar was consistently rated as the highest priority, often at the expense of the environmental dimension. This trend was statistically confirmed by post-hoc Tukey HSD tests, which showed that respondents in these industrial hubs place a significantly higher value on "Economic Viability" and "Employment Stability" than respondents in Prague or the South Moravian region. This evidence provides robust support for **H4**, suggesting that the regional economic context and historical industrial pathways act as primary determinants of how sustainability is internalised.

4.3 Socio-demographic Determinants of Sustainability Perceptions

The final phase of the empirical analysis investigated the influence of selected socio-demographic characteristics, namely gender, age, and educational attainment, on the prioritisation of sustainability pillars. The results of the independent samples t-tests indicated a statistically significant difference between men and women regarding their approach to the three pillars, thereby supporting **H5**. Men consistently demonstrated a higher relative preference for the economic pillar, often associating sustainability with technological innovation and

market efficiency. In contrast, female respondents exhibited a more balanced perception, placing significantly higher value on the social and environmental dimensions. This aligns with international psychological research suggesting that women tend to prioritise communal and altruistic values, which are closely linked to the social pillar of sustainable development. The independent samples t-test revealed a statistically significant difference between men and women ($t = 3.52$, $p < 0.001$), with a small to moderate effect size (Cohen's $d = 0.31$).

Generational differences were analysed using the Kruskal–Wallis test based on an overall sustainability score, calculated as the mean of all 30 sustainability-related items. The analysis revealed statistically significant differences between age groups ($H = 33.90$, $p < 0.001$, $\varepsilon^2 = 0.062$), providing empirical support for **H6**. Post-hoc pairwise comparisons with Bonferroni correction indicated no significant differences between the Net Generation (15–25 years) and Millennials (26–40 years) (adjusted $p = 1.000$), nor between Xennials (41–60 years) and Baby Boomers (61+ years) (adjusted $p = 1.000$). In contrast, all comparisons between younger and older generations were statistically significant (adjusted $p \leq 0.004$). These findings indicate that younger respondents expressed significantly stronger overall sustainability perceptions than older generations.

Finally, the impact of educational attainment was examined to test **H7**. The Kruskal–Wallis test revealed statistically significant differences in the preference for the economic pillar across education levels ($H = 8.14$, $p = 0.043$, $\varepsilon^2 = 0.010$). Respondents with primary or vocational education showed a greater preference for the economic pillar, whereas respondents with higher levels of education demonstrated a more balanced perception of sustainability. These findings support **H7** and indicate that educational attainment is associated with differences in the prioritisation of sustainability pillars.

4.4 Discussion and Policy Implications

The identified perception gap has important implications for regional sustainability policy. The confirmation of regional disparities (H4) suggests that the transition to a carbon-neutral economy is also a psychological and social challenge. In industrial regions such as the Moravian-Silesian or Ústí nad Labem regions, the strong prioritisation of the economic pillar is a rational response to historical dependencies on heavy industry (ČSÚ, 2024). This interpretation is further supported by the triangulation of subjective survey responses with objective regional ESG indicators, which revealed a consistent correspondence between citizens' sustainability perceptions and the socio-economic characteristics of individual regions. In these regions, sustainability is often perceived as a threat to livelihoods rather than an opportunity. Therefore, communication strategies regarding the "Green Deal" must be differentiated; while urban centres like Prague may respond well to environmental incentives, industrial regions require a narrative focused on economic security, job creation, and "just transition" mechanisms.

Furthermore, the superiority of the Venn diagram model (H2) over the Parthenon model indicates that the Czech public is increasingly capable of recognising the complex trade-offs between the pillars (Purvis et al., 2019). This suggests that simplified, top-down communication that treats the three pillars as isolated silos is no longer effective. Educational programmes should emphasise the interconnections among sustainability pillars, for example through green jobs or socially responsible investment. The socio-demographic trends identified (H5, H6, H7) also point to generational and educational differences. Younger respondents expressed stronger overall sustainability perceptions, whereas respondents with higher educational attainment demonstrated a more balanced perception of sustainability (Stern, 2000). To prevent social polarisation, policies should also address the economic concerns of older and less-educated groups.

5 Conclusion

This study analysed sustainability perceptions in the Czech Republic using survey data from 500 respondents and statistical modelling. The findings supported all seven hypotheses, showing that while the Brundtland definition remains the dominant conceptual framework (H1), sustainability perceptions are significantly shaped by regional and socio-demographic factors. The Venn diagram model (H2) provided a more appropriate representation of sustainability perception than the Parthenon model.

The findings highlight significant regional disparities, particularly the economic prioritisation in industrial regions (H4), which poses a challenge for the implementation of uniform national sustainability strategies. Furthermore, the identified gender, age, and educational gaps (H5, H6, H7) suggest that communication regarding sustainability must be tailored to specific social groups to be effective. For policymakers, these results indicate that a "one-size-fits-all" approach to the Green Deal or regional transition plans is likely to face resistance in areas where economic stability is perceived to be at risk. Future research should focus on

longitudinal studies to track how these perceptions evolve as the Czech Republic continues its transition toward a more sustainable and carbon-neutral economy.

The triangulation with regional ESG indicators provided additional support for the survey findings. Regions characterised by a stronger industrial profile, particularly the Moravian-Silesian, Ústí nad Labem and Karlovy Vary regions, exhibited a stronger preference for the economic pillar of sustainability. Conversely, economically advanced regions with more diversified ESG profiles demonstrated a more balanced perception of environmental, social, and economic dimensions. These findings suggest that sustainability perceptions are not formed in isolation but are influenced by the broader socio-economic and environmental context of individual regions.

Despite its contributions, the study has several limitations. The analysis is based on self-reported perceptions, which may be influenced by subjective interpretation and social desirability bias. In addition, the cross-sectional design does not allow for the observation of changes in sustainability perception over time. Future research should therefore focus on longitudinal and comparative approaches involving additional national contexts.

Acknowledgements

This research was supported by the Student Grant Competition, Faculty of Economics, VSB–Technical University of Ostrava (project SP2026/064) and by the European Union through the project REFRESH – Research Excellence for Region Sustainability and High-tech Industries (No.CZ.10.03.01/00/22_003/0000048), funded under the Operational Programme Just Transition.

References

ČSÚ (2024) Udržitelný rozvoj v krajích ČR. Český statistický úřad, Prague

Elkington J (1997) Cannibals with Forks: The Triple Bottom Line of 21st Century Business. Capstone, Oxford

Hair J F, Black W C, Babin B J, Anderson R E (2018) Multivariate Data Analysis, 8th ed. Cengage, Boston

Kates R W, Parris T M, Leiserowitz A A (2005) What is sustainable development? Goals, indicators, values, and practice. Environment: Science and Policy for Sustainable Development 47(3):8–21

Milne M J, Gray R (2013) W(h)ither ecology? The triple bottom line, the global reporting initiative, and corporate sustainability reporting. Journal of Business Ethics 118(1):13–29

Purvis B, Mao Y, Robinson D (2019) Three pillars of sustainability: in search of conceptual origins. Sustainability Science 14(3):681–695

Schwartz SH (1992) Universals in the content and structure of values. Advances in Experimental Social Psychology 25:1–65

Stern P C (2000) Toward a coherent theory of environmentally significant behaviour. Journal of Social Issues 56(3):407-424

WCED (1987) Our Common Future. Oxford University Press, Oxford

Comparison of Bootstrap Methods in Robust SSD Portfolio Optimization

Jakub Neugebauer[1]

Abstract. This paper explores the use of alternative bootstrap approaches for calibrating robustness in a second-order stochastic dominance (SSD) portfolio optimization framework. Classical SSD models impose strict dominance constraints that are often overly restrictive in finite samples and sensitive to sample noise. To address this issue, two bootstrap methods for determining statistically insignificant levels of dominance violation are investigated. The first is random-weights approach, which approximates the distribution of the dominance statistic across randomly generated portfolios, and a benchmark-based approach, which derives critical values directly from the empirical benchmark distribution. These thresholds are incorporated into the SSD model to relax the dominance constraints in a controlled manner. The approaches are evaluated across multiple rolling windows and varying sample sizes, with performance assessed out-of-sample in terms of returns and violation frequency. Results show that the benchmark-based bootstrap yields critical values that track out-of-sample dominance violations more closely than the random-weights approach, at the cost of a less restrictive, but more realistically calibrated, dominance constraint.

Keywords: bootstrap methods, portfolio optimization, stochastic dominance, robust optimization

JEL Classification: C44, G11

1 Introduction

Portfolio optimization is often based on the selection of a specific risk measure. Classical approaches compare portfolios using variance, semivariance, Value at Risk, expected shortfall, or other numerical characteristics of the return distribution. Although these measures are useful and computationally convenient, they describe only one selected aspect of risk. As a result, the obtained portfolio is optimal only with respect to the chosen measure.

Stochastic dominance provides a more general framework for comparing investment alternatives. Instead of reducing the return distribution to one statistic, it compares the whole distribution. Second-order stochastic dominance (SSD) is especially important in portfolio optimization, because it corresponds to the preferences of all risk-averse investors with non-decreasing and concave utility functions. The use of SSD in portfolio optimization was made computationally tractable by the linear programming formulation of Dentcheva and Ruszczyński, 2006. Further developments include efficient solution methods and robust extensions of SSD-constrained models (see e.g. Fábián et al, 2011 and Hosseini-Nodeh et al, 2022).

A practical limitation of strict SSD constraints is their sensitivity to finite samples. In empirical applications, dominance relations are evaluated from historical data, which may contain noise, estimation errors, and extreme observations. Therefore, a small violation of the SSD constraint does not necessarily mean that the portfolio is truly dominated. This motivates the use of statistically calibrated relaxations of the dominance constraint.

Stochastic dominance testing provides a natural way to define such relaxations. Several tests for stochastic dominance have been proposed in the literature, including early tests by McFadden, 1989 and Kaur et al, 1994, Kolmogorov–Smirnov type tests by Barrett and Donald, 2003, and a testing framework for dependent time series by Linton et al, 2005. These tests make it possible to distinguish between economically relevant dominance violations and violations that may be caused only by sampling variability.

In Neugebauer, 2025, a robust stochastic dominance portfolio optimization framework called s-RSD was proposed for a general order of stochastic dominance. The framework relaxes the strict stochastic dominance constraint using a bootstrap critical value calibrated from a random-weights bootstrap procedure, where randomly generated portfolios are compared with the benchmark over rolling subsamples.

[1] Prague University of Economics and Business, Faculty of Informatics and Statistics, nám. Winstona Churchilla 1938/4, 120 00 Praha 3-Žižkov, jakub.neugebauer@vse.cz

Although the original formulation was developed for arbitrary dominance order s, this paper focuses only on the second-order case. Second-order stochastic dominance is computationally the least demanding formulation, since it can be represented as a linear programming problem Dentcheva and Ruszczyński, 2006. Furthermore, cardinality constraints are not included in this analysis in order to avoid mixed-integer formulations.

This paper extends the original random-weights bootstrap approach introduced by Neugebauer, 2025 by comparing it with an alternative benchmark-based bootstrap procedure. In contrast to the random-weights approach, the benchmark-based method constructs the bootstrap critical values directly from resampled benchmark returns.

The aim of this paper is to compare these two bootstrap calibration methods in the same second-order robust stochastic dominance framework. The comparison is based on the size of the obtained critical values, out-of-sample dominance violations, and portfolio performance across different significance levels and history lengths. The results show how the choice of bootstrap method affects the practical strictness of the SSD constraint and the resulting optimized portfolios.

2 2-RSD Model and Bootstrap Methods

Let $\hat{\mathbf{R}}$ denote the matrix of historical asset returns of size $t \times n$, where t is the number of observations and n is the number of assets, and let $\hat{\mathbf{r}}_i$ denote its i-th row, i.e. the vector of asset returns observed at time i. Let $\mathbf{w}$ be the vector of portfolio weights and let $\hat{\mathbf{b}} = (b_1, \dots, b_t)^\top$ denote the benchmark return series. The domain of the dominance functionals is fixed by the observed data as

$$\underline{x} = \min_{i,k=1,\dots,t;\, j=1,\dots,n} \{\hat{R}_{ij},\, b_k\}, \qquad \bar{x} = \max_{i,k=1,\dots,t;\, j=1,\dots,n} \{\hat{R}_{ij},\, b_k\}.$$

The robust second-order stochastic dominance (2-RSD) portfolio optimization problem is formulated as

$$\begin{aligned} \max \quad & \mathbb{E}\big[\hat{\mathbf{R}}\,\mathbf{w}\big] \\ \text{s.t.} \quad & t_{ij} \geq b_j - \hat{\mathbf{r}}_i \mathbf{w}, && i, j = 1, \dots, t, \\ & t_{ij} \geq 0, && i, j = 1, \dots, t, \\ & \frac{1}{t} \sum_{i=1}^{t} t_{ij} - \hat{D}_B^2(b_j) \leq \frac{S_\alpha^2}{\sqrt{t}}, && j = 1, \dots, t, \\ & w_i \geq 0, && i = 1, \dots, n, \end{aligned} \tag{1}$$

where $\hat{S}^2(\mathbf{w})$ is the SSD test statistic defined in equation (2) below, and S_α^2 is the bootstrap critical value at significance level α. Since $\hat{D}_B^2$ is affine between consecutive order statistics of the benchmark while $\hat{D}_P^2$ is convex in x, the supremum defining $\hat{S}^2(\mathbf{w})$ is always attained at $x = b_j$ for some $j \in \{1, \dots, t\}$. The per-j constraint in model (1) is therefore an exact discretization of the continuous supremum.

The SSD test statistic is based on the Kolmogorov–Smirnov type formulation of Linton et al, 2005 and is defined as

$$\hat{S}^2(\mathbf{w}) = \sqrt{t} \sup_{x \in [\underline{x}, \bar{x}]} \left[\hat{D}_P^2(x) - \hat{D}_B^2(x)\right], \tag{2}$$

where, for a return series $X = \{X_1, \dots, X_l\}$ of generic length l, the empirical second-order dominance functional is defined as

$$\hat{D}_X^2(x) = \frac{1}{l} \sum_{i=1}^{l} \max(x - X_i, 0).$$

Setting $l = t$ and $X_i = \hat{\mathbf{r}}_i \mathbf{w}$ (portfolio return at time i) yields $\hat{D}_P^2$, and setting $l = t$ and $X_i = b_i$ (benchmark return at time i) yields $\hat{D}_B^2$. The same functional, evaluated with $l = k$, is used in equations (3) and (4) below to define the bootstrap statistics on length-k subsamples.

If $S_\alpha^2 = 0$, the model reduces to the classical strict SSD formulation. Positive values of S_α^2 allow statistically insignificant violations of the SSD constraint and make the optimization problem less restrictive.

This paper compares two bootstrap approaches for determining the critical value S_α^2, following the general testing procedure of Linton et al, 2005: both construct overlapping rolling subsamples of length k, with starting index $j = 1, \dots, t - k + 1$, over which the test statistic is repeatedly evaluated.

The first approach, introduced in Neugebauer, 2025, is based on randomly generated portfolio weights. For each replication $i = 1, \ldots, n_b$, weights are drawn as $w_\ell \sim \mathcal{U}(0,1)$ i.i.d. across assets $\ell = 1, \ldots, n$ and normalized so that $\sum_\ell w_\ell = 1$; short-selling is precluded, since $w_\ell \geq 0$ by construction. The resulting random portfolio return series is $\Theta_i = \hat{\mathbf{R}}\mathbf{w}^{(i)}$, where $\mathbf{w}^{(i)}$ denotes the i-th draw of random weights; no resampling of the historical returns themselves takes place in this approach. For a given starting index $j \in \{1, \ldots, t-k+1\}$, define the length-$k$ subsamples

$$\Theta_{i,j} = \{\Theta_{i,j}, \Theta_{i,j+1}, \ldots, \Theta_{i,j+k-1}\}, \qquad B_j = \{b_j, b_{j+1}, \ldots, b_{j+k-1}\}.$$

The random portfolio subsample is compared with the corresponding benchmark subsample as:

$$\hat{S}^{rs}_{i,j} = \sqrt{k} \sup_{x \in [\underline{x}, \bar{x}]} \left[\hat{D}^2_{\Theta_{i,j}}(x) - \hat{D}^2_{B_j}(x) \right], \qquad i = 1, \ldots, n_b,\ j = 1, \ldots, t-k+1. \tag{3}$$

This produces $n_b(t-k+1)$ bootstrap values, and the critical value S^{rs}_α is determined as their $(1-\alpha)$-quantile.

The second approach proposed in this paper is based directly on benchmark resampling. For each replication $m = 1, \ldots, n_b$, an i.i.d. bootstrap sample $B^*_m = \{B^*_{m,1}, \ldots, B^*_{m,t}\}$ of length t is first drawn with replacement from $\hat{\mathbf{b}}$. For a given starting index $j \in \{1, \ldots, t-k+1\}$, the length-$k$ subsample of this resampled series is

$$B^*_{m,j} = \{B^*_{m,j}, B^*_{m,j+1}, \ldots, B^*_{m,j+k-1}\},$$

extracted in the same way as the benchmark subsample B_j defined above. The corresponding statistic is

$$\hat{S}^{bs}_{m,j} = \sqrt{k} \sup_{x \in [\underline{x}, \bar{x}]} \left[\hat{D}^2_{B^*_{m,j}}(x) - \hat{D}^2_{B_j}(x) \right], \qquad m = 1, \ldots, n_b,\ j = 1, \ldots, t-k+1. \tag{4}$$

The benchmark-based critical value S^{bs}_α is again determined as the empirical $(1-\alpha)$-quantile. In the empirical analysis of Section 3, $n_b = 10\,000$ bootstrap replications are used for both approaches.

The random-weights bootstrap approximates the distribution of SSD violations across feasible portfolios, while the benchmark-based bootstrap measures the variability of the benchmark distribution itself. The following section compares both approaches empirically for different significance levels and history lengths.

3 Results of Comparison

The empirical analysis is based on daily logarithmic returns of the S&P 500 index and its constituent assets from 1 January 2010 to 31 December 2025. The S&P 500 index is used as the benchmark portfolio, while the index constituents form the set of investable assets. The analysis uses the current set of S&P 500 constituents, which may introduce mild survivorship bias into the estimated critical values and portfolio characteristics. Constituents without any price history over the full period are excluded. Remaining missing values arising from staggered listing dates are filled with zero returns.

The comparison is performed using a rolling-window framework. For each evaluation window, the optimization model is estimated on an in-sample period of length $h \in \{60, 125, 250, 500\}$ trading days and evaluated on the immediately following out-of-sample period of the same length. For each history length, 100 rolling windows are generated uniformly across the dataset.

Three variants of the SSD optimization model are compared: the strict SSD formulation with $S^2_\alpha = 0$, the random-weights bootstrap calibration, and the benchmark-based bootstrap calibration. For the bootstrap approaches, significance levels $\alpha \in \{0.01, 0.05, 0.10, 0.25\}$ are considered. Bootstrap critical values are estimated using 10 000 bootstrap replications and rolling subsamples of length $k = 10$.

The comparison focuses on both economic and statistical properties of the obtained portfolios. The analyzed quantities include the average out-of-sample portfolio return r, portfolio volatility measured by standard deviance σ, the average bootstrap critical value S^2_α, the average out-of-sample SSD statistic T_B, the proportion of cases in which the out-of-sample SSD statistic exceeds the corresponding bootstrap critical value v_{S_α}, and the average ratio between the out-of-sample SSD statistic and the corresponding bootstrap critical value $\frac{T}{S^2_\alpha}$.

Table 1 additionally reports the proportions p_S and p_B. The quantity p_S represents the percentage of cases in which the out-of-sample SSD statistic of the strict SSD portfolio remains below the bootstrap critical value estimated by the corresponding bootstrap method. This evaluates how well the bootstrap threshold captures the out-of-sample behaviour of portfolios that satisfy the strict SSD constraint in-sample. The quantity p_B analogously represents the percentage of cases in which the out-of-sample SSD statistic of the bootstrap-optimized portfolio remains below

its corresponding bootstrap critical value. Together, these measures provide additional information about how closely the bootstrap calibration reflects the observed out-of-sample dominance behaviour under the two bootstrap approaches.

Since the strict SSD portfolio does not depend on α, p_S is a direct check of how well the bootstrap is calibrated. If the bootstrap works well and the in-sample and out-of-sample returns come from the same distribution, p_S should be close to $1 - \alpha$. The quantity p_B is different. The bootstrap-optimized portfolio is chosen so that its in-sample SSD statistic is pushed as close as possible to the constraint S^2_α. Because of this, p_B does not only measure the calibration of the bootstrap, but also a selection effect. The optimizer tends to pick portfolios whose in-sample statistic looks better than it really is, simply because these are the portfolios that satisfy the constraint. For this reason, p_B should be seen as a measure of how robust the optimized portfolio is out-of-sample, not as a calibration check in the same sense as p_S.

		random_weights				**benchmark**			
α	**History**	S^2_α	T_B	p_S (%)	p_B (%)	S^2_α	T_B	p_S (%)	p_B (%)
0.01	60	0.0046	0.0297	5	1	0.0310	0.0504	67	20
	125	0.0051	0.0336	5	3	0.0327	0.0616	63	9
	250	0.0056	0.0341	5	3	0.0345	0.0618	70	11
	500	0.0058	0.0428	10	4	0.0373	0.0730	59	9
0.05	60	0.0036	0.0286	4	1	0.0223	0.0448	49	10
	125	0.0037	0.0317	2	3	0.0235	0.0534	45	7
	250	0.0038	0.0320	3	2	0.0246	0.0531	39	9
	500	0.0037	0.0404	7	4	0.0263	0.0622	39	5
0.10	60	0.0029	0.0278	3	1	0.0178	0.0414	38	7
	125	0.0029	0.0306	2	3	0.0188	0.0484	26	5
	250	0.0029	0.0309	2	1	0.0196	0.0485	29	8
	500	0.0028	0.0394	7	4	0.0208	0.0567	27	5
0.25	60	0.0017	0.0264	2	1	0.0106	0.0351	11	3
	125	0.0016	0.0287	2	3	0.0113	0.0403	10	5
	250	0.0016	0.0291	1	0	0.0119	0.0406	17	5
	500	0.0014	0.0380	6	2	0.0126	0.0486	14	5

Table 1 Bootstrap critical values and out-of-sample SSD behaviour

Table 1 shows substantial differences between the two bootstrap calibration approaches. The benchmark-based bootstrap produces significantly larger critical values S^2_α across all significance levels and history lengths, often by nearly one order of magnitude compared to the random-weights bootstrap. Consequently, the corresponding out-of-sample SSD statistics T_B are also substantially larger for the benchmark bootstrap portfolios.

The most important difference is observed in the proportions p_S and p_B. For the random-weights bootstrap, both quantities remain very low across all configurations, especially for smaller significance levels. This indicates that the obtained bootstrap thresholds do not reflect well the observed out-of-sample behaviour, even for portfolios satisfying the strict SSD constraint in-sample. In contrast, the benchmark-based bootstrap achieves substantially higher values of both p_S and p_B, particularly for $\alpha = 0.01$, where p_S reaches values between 59% and 70%.

These results suggest that the benchmark-based bootstrap reflects the magnitude of out-of-sample SSD violations more realistically than the random-weights bootstrap. Of course, a perfect match cannot be expected since market conditions evolve over time and the out-of-sample returns do not follow exactly the same distribution as the in-sample data. However, the benchmark-based approach still provides a noticeably closer approximation of the observed out-of-sample dominance behaviour and therefore appears to be a more appropriate calibration procedure for the robust SSD framework.

Table 2 reports the descriptive out-of-sample characteristics of the optimized portfolios. The strict SSD formulation violates the out-of-sample SSD condition in almost all cases, with $v_{S^2_\alpha}$ typically between 96% and 100%. This confirms that strict in-sample SSD constraints do not transfer well to unseen financial data and are highly sensitive to changes in the return distribution.

The random-weights bootstrap slightly improves the economic performance relative to the strict SSD model, leading

to moderately higher average returns and volatility. However, the corresponding ratios $\frac{T}{S_\alpha^2}$ remain extremely large across all significance levels, often exceeding values of 10 and reaching almost 30 for $\alpha = 0.25$. Together with the very high violation frequencies, this indicates that the random-weights bootstrap produces critical values that are substantially too small relative to the observed out-of-sample SSD fluctuations.

The benchmark-based bootstrap produces noticeably different results. The optimized portfolios achieve consistently higher average returns than both the strict SSD and random-weights approaches. At the same time, portfolio volatility also increases, reflecting the less restrictive nature of the benchmark-based calibration. Most importantly, the values of $\frac{T}{S_\alpha^2}$ are lower and remain much closer to one, typically between 2 and 4. The corresponding violation frequencies are also consistently lower than for the other approaches.

		strict			**random_weights**				**benchmark**			
α	**History**	r (%)	σ (%)	$v_{S_\alpha^2}$ (%)	r (%)	σ (%)	$v_{S_\alpha^2}$ (%)	$\frac{T}{S_\alpha^2}$	r (%)	σ (%)	$v_{S_\alpha^2}$ (%)	$\frac{T}{S_\alpha^2}$
0.01	60	0.10	1.90	99	0.11	2.15	99	8.06	0.13	2.95	80	1.94
	125	0.11	1.77	98	0.12	1.99	97	7.83	0.13	2.72	91	2.14
	250	0.10	1.63	100	0.11	1.78	97	7.27	0.13	2.29	89	2.00
	500	0.09	1.56	96	0.09	1.67	96	7.74	0.10	2.05	91	2.00
0.05	60	0.10	1.90	99	0.11	2.10	99	10.78	0.12	2.72	90	2.42
	125	0.11	1.77	98	0.12	1.94	97	10.16	0.13	2.51	93	2.60
	250	0.10	1.63	100	0.10	1.74	98	9.81	0.13	2.13	91	2.40
	500	0.09	1.56	96	0.09	1.64	96	11.43	0.10	1.92	95	2.40
0.10	60	0.10	1.90	99	0.11	2.07	99	14.09	0.12	2.59	93	2.84
	125	0.11	1.77	98	0.11	1.91	97	12.63	0.13	2.38	95	2.96
	250	0.10	1.63	100	0.10	1.72	99	12.45	0.12	2.05	92	2.76
	500	0.09	1.56	96	0.09	1.62	96	14.75	0.10	1.85	95	2.76
0.25	60	0.10	1.90	99	0.11	2.01	99	28.17	0.12	2.35	97	4.20
	125	0.11	1.77	98	0.11	1.85	97	24.13	0.12	2.17	95	4.16
	250	0.10	1.63	100	0.10	1.68	100	22.77	0.11	1.90	95	3.82
	500	0.09	1.56	96	0.09	1.60	98	27.77	0.09	1.75	95	3.89

Table 2 Out-of-sample portfolio returns, volatility and SSD violations

Overall, the results indicate that the benchmark-based bootstrap provides a more realistic calibration of the robust SSD constraint. Although the out-of-sample SSD condition is still violated relatively often, which is expected under changing market conditions, the benchmark-based approach appears to capture the empirical scale of SSD fluctuations better than the random-weights bootstrap.

4 Conclusion

This paper compared two bootstrap approaches for calibrating robustness in second-order stochastic dominance portfolio optimization. Within this empirical setting, based on a single benchmark index (S&P 500) and a single historical period, the results indicate that the benchmark-based bootstrap produces larger critical values that track out-of-sample SSD violations more closely than the random-weights approach, and that the resulting portfolios achieved higher returns and lower relative violation ratios.

These findings should be interpreted with some caution. The comparison relies on a single market and estimation period, no theoretical reference value is currently available for the proportion p_B (see Section 3), and both bootstrap procedures still leave a substantial share of out-of-sample violations unexplained. Future research should analyze the effect of the subsample length k used in both bootstrap procedures, since this paper fixes $k = 10$. Overall, the benchmark-based bootstrap appears to provide a more suitable, though not fully satisfactory, calibration procedure for robust SSD optimization under changing market conditions.

Acknowledgements

The research project was supported by Grant No. IGA F4/18/2024 of the Internal Grant Agency, Faculty of Informatics and Statistics, Prague University of Economics and Business. Artificial intelligence tools, specifically OpenAI's GPT-5 OpenAI, 2026, were used in the preparation of this manuscript for improvement of grammar and language style, for editing and expansion of textual sections, and for support in the development and presentation of mathematical formulations. The author is fully responsible for the content and interpretation of the entire manuscript.

References

Barrett GF, Donald SG (2003) Consistent tests for stochastic dominance. Econometrica 71(1):71–104. https://doi.org/doi.org/10.1111/1468-0262.00390, https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00390, https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00390

Dentcheva D, Ruszczyński A (2006) Portfolio optimization with stochastic dominance constraints. Journal of Banking & Finance 30(2):433–451. https://doi.org/doi.org/10.1016/j.jbankfin.2005.04.024, https://www.sciencedirect.com/science/article/pii/S0378426605000725, risk Management and Optimization in Finance

Fábián CI, Mitra G, Roman D (2011) Processing second-order stochastic dominance models using cutting-plane representations. Mathematical Programming 130:33–57. https://doi.org/10.1007/s10107-009-0326-1, doi.org/10.1007/s10107-009-0326-1

Hosseini-Nodeh Z, Khanjani-Shiraz R, Pardalos PM (2022) Distributionally robust portfolio optimization with second-order stochastic dominance based on wasserstein metric. Information Sciences 613:828–852. https://doi.org/doi.org/10.1016/j.ins.2022.09.039, https://www.sciencedirect.com/science/article/pii/S0020025522010842

Kaur A, Prakasa Rao B, Singh H (1994) Testing for second-order stochastic dominance of two distributions. Econometric Theory 10(5):849–866. https://doi.org/10.1017/S0266466600008884

Linton O, Maasoumi E, Whang YJ (2005) Consistent testing for stochastic dominance under general sampling schemes. The Review of Economic Studies 72(3):735–765. http://www.jstor.org/stable/3700672

McFadden D (1989) Testing for stochastic dominance. In: Fomby TB, Seo TK (eds) Studies in the Economics of Uncertainty. Springer New York, New York, NY, pp 113–134

Neugebauer J (2025) Portfolio optimization with robust stochastic dominance testing: A genetic algorithm approach. European Journal of Operational Research https://doi.org/https://doi.org/10.1016/j.ejor.2025.12.037, https://www.sciencedirect.com/science/article/pii/S0377221725010045

OpenAI (2026) Gpt-5 language model. https://openai.com, accessed: 14-05-2026

Path Dependent Arbitrage Portfolio Strategies

Valentina Piantoni[1], Denise Mirabella[2]

Abstract. This study investigates portfolio strategies aimed at optimizing both the expected first passage time of wealth toward predefined benchmark levels and the expected current maximum and minimum of the wealth process. Assuming that returns follow Markov processes, approximated through homogeneous Markov chains, the research identifies potential arbitrage opportunities by exploiting momentum anomalies.

The empirical analysis is conducted on the components of the S&P 500 index over a 20-year period (2002–2021) using a rolling window approach, covering three major financial stress periods: the 2008 subprime mortgage crisis, the 2011 European sovereign debt crisis and the 2020 COVID-19 pandemic. The results demonstrate that the temporal frequency of data significantly impacts strategy effectiveness. Specifically, when using daily data, reverse arbitrage proves to be the most effective strategy for wealth generation. Conversely, with weekly returns, "winner" portfolios outperform, and reverse arbitrage loses its efficacy. For monthly data, the success of the strategy is strictly dependent on the chosen time horizon. Overall, arbitrage strategies based on winner-loser dynamics emerge as the most robust and effective approaches during periods of profound market crisis.

Keywords: Path-dependent strategies, Markov chains, Reverse arbitrage, Momentum, First passage time

JEL Classification: C44, G11

1 Introduction

Recent literature has increasingly focused on the severe losses experienced by momentum strategies during market rebounds, known as momentum crashes (Barroso and Santa-Clara, 2015; Daniel and Moskowitz, 2016), pushing the state-of-the-art towards dynamic risk-managed approaches and the re-evaluation of momentum factors over time (Huang et al., 2020; Ehsani and Linnainmaa, 2022). Simultaneously, from a methodological perspective, non-parametric approaches based on Markov Chains and first passage times have recently regained significant traction in portfolio optimization for their superior ability to manage intra-horizon tail risks without strict distributional constraints (Zsurkis et al., 2024). Our proposed Markovian path-dependent approach contributes to this recent strand of literature by offering a non-parametric timing solution that intrinsically adjusts to extreme market shifts without requiring complex volatility forecasting. Ortobelli et al. (2019) introduced an innovative timing portfolio selection strategy leveraging the Markovian properties of certain Lévy processes. Building on this theoretical framework, we propose alternative portfolio selection strategies that optimize the first passage times of wealth at predefined benchmark levels without relying on specific distributional assumptions. This approach enhances the flexibility of portfolio management and simplifies computational demands. The rejection of the Gaussian distribution for log-returns, due to observed skewness and excess kurtosis, has motivated the use of parametric Markov processes like Lévy processes in financial modeling. While these models capture market dynamics effectively, forecasting their future values often entails high computational complexity. To address this, Duan and Simonato (2001) proposed an approximation of the Markovian evolution of portfolio values using Markov chains. Our approach, however, builds on the adaptive node-based methodology of Iaquinta and Ortobelli (2006), significantly improving computational efficiency in portfolio selection.

Unlike Ortobelli et al. (2019), we introduce nonparametric timing strategies that maximize the first-time wealth decreases and minimize the first-time wealth increases, within a fixed temporal horizon as suggested by Kardaras and Platen (2010). We demonstrate that the optimal portfolio derived from this optimization tends to

[1] University of Bergamo, Department of Management, Via dei Caniana, 2, 24127, Bergamo, valentina.piantoni@unibg.it
[2]University of Bergamo, Department of Economics, Via dei Caniana, 2, 24127, Bergamo, denisee.mirabella@gmail.com

exhibit a negative correlation with portfolios aimed at minimizing the time to reach lower wealth thresholds. This insight underpins the development of arbitrage-based portfolio strategies as variations of traditional wealth optimization.

A key contribution of this paper is an empirical comparison of hedging and timing portfolio strategies. The primary scientific objective of this research is to address the vulnerability of traditional momentum strategies to sudden market crashes (momentum crashes) by introducing a flexible, path-dependent framework. Rather than relying on strict distributional assumptions, we aim to demonstrate how optimizing the first passage time and extreme wealth values can mitigate tail risk during systemic crises. Empirical comparison serves as a methodological tool to validate this theoretical framework, identifying under which temporal granularities these arbitrage opportunities remain robust and practically implementable.

2 Portfolio Returns Approximated with Markov Chains

This section introduces a method for approximating Markov processes, which is utilized in empirical analysis. Specifically, it employs the approach proposed by Duan and Simonato (2001) and Iaquinta et al. (2011) to construct the Markovian transition matrix. We consider n risky assets with gross returns $z_{t+1}=[z_{1,t+1},...,z_{n,t+1}]'$. Let $x=[x_1,...,x_n]'$represent the portfolio weights. The portfolio returns over the interval $[t,t+1]$ is defined as:

$$z_{(x),t+1} = x'z_{t+1} = \sum_{i=1}^{n} x_i z_{i,t+1} \tag{1}$$

To simplify the computation, the range of the portfolio's gross returns is divided into N intervals, each corresponding to a state $z_{(x)}^{(i)}$, where i=1,…,N, arranged in descending order ($z_{(x)}^{(i)} > z_{(x)}^{(i+1)}$for i=1,...,N−1). The return associated with each state is calculated as the geometric mean of the interval's endpoints:

$$z_{(x)}^{(i)} = z_{(x)}^{(1)} u^{1-i} \tag{2}$$

where u is a scaling factor based on the number of states N and historical portfolio returns.

According to Iaquinta et al. (2011) the possible wealth values W_k after k steps can be efficiently represented in a wealth-tree, whose number of nodes grows linearly with the time k, thanks to the recombining property of Markov chains. This property significantly reduces the memory requirements. The wealth W_k at the l-th node and time k is given by:

$$w^{(l,k)} = (z^{(1)})^k * u^{(1,l)}, l = 1, \ldots, (N-1)k+1 \tag{3}$$

The Markov transition matrix P, which is time-homogeneous, is estimated using maximum likelihood estimates. In particular, the entries of P are computed based on the observed frequency of transitions between states in historical data.

The recombining structure of the wealth-tree allows for efficient computation of both unconditional and conditional distributions of future wealth (e.g., conditional on an initial state $z^{(s_0)}$). The algorithm proposed by Iaquinta and Ortobelli has a computational complexity of $O(N^3k^2)$ and serves as the foundation for determining the distribution of first passage times.

A first passage refers to the earliest time at which wealth $W_k(X)$ reaches a specified set A, which is often a closed interval. This concept is particularly valuable in portfolio management. The first passage time is defined as:

$$\tau_A(x) = \inf\{k \geq 0 | k \in N_0;\ W_k(x) \in A \subseteq R\} \wedge T \tag{4}$$

where T is a predetermined time horizon.

By leveraging the Markov chain framework, the unconditional distribution of the first passage time τ_A can be calculated recursively and efficiently.

This recursive calculation is particularly powerful for evaluating portfolio strategies, as it allows for the estimation of when wealth will reach a target level under different market conditions. The ability to compute first passage times with high computational efficiency also makes this method ideal for real-time portfolio optimization in dynamic financial markets.

In summary, the use of Markov chains and first passage times provides a rigorous yet computationally efficient framework for portfolio management. The ability to approximate the distribution of portfolio returns and

determine the expected time to reach specific wealth thresholds offers valuable insights for investors aiming to optimize their portfolios in both stable and volatile market conditions.

3 Arbitrage Strategies Based on the Momentum

Momentum strategies are based on the idea of selecting stocks that have performed well in the past ("winners") and avoiding or selling those that have underperformed ("losers"). These strategies rely on the analysis of the spread between winner and loser portfolios, which can be evaluated using factor models to identify the underlying drivers of momentum profits. The core principle is that assets that have performed well in the past will continue to do so in the short term, while those that have underperformed are likely to continue declining.

The structure of momentum strategies requires making critical decisions regarding key parameters, which include:

- The ranking or formation period, during which portfolios of stocks are evaluated based on their historical returns, allowing the investor to determine which assets qualify as "winners" and "losers."
- The holding or investment period, during which the portfolios are left unchanged, and the performance of the selected assets is tracked.
- The ranking criterion, which identifies the winners and losers based on performance metrics, typically using cumulative returns over a specified period, such as six months or a year. This generates a zero-investment, self-financing strategy based on an arbitrage portfolio (whose the sum of weights is zero) that aims to capture the momentum profits from past trends.

The strategy entails selling loser portfolios and buying winner portfolios at the end of the ranking period, followed by performance evaluation during the holding period. Traditional methods for defining winners and losers often rely on cumulative returns over the ranking period. However, these methods overlook the risk characteristics of stocks and assume that portfolio of returns follow a normal distribution—a common assumption that is frequently contradicted by empirical evidence showing skewness and excess kurtosis in stock returns.

To address these shortcomings, we propose the application of a risk-return optimization framework. This framework ensures that the winner portfolios capture higher returns, while the loser portfolios help mitigate potential risks. By incorporating the statistical properties of non-Gaussian returns, the strategies account for the observed skewness and excess kurtosis, leading to more accurate and reliable performance metrics.

Moreover, to better manage the inherent risks, we propose constructing portfolios of winners and losers using risk-adjusted criteria during the ranking period. This method ensures that the allocation in each stock is based not only on past performance but also on the volatility and risk characteristics of the assets. At rebalancing points, two optimization problems are solved:

1. We identify a winner portfolio x that minimizes the expected first time $E(\tau_{\text{win}}(x,bar,T))$ the portfolio wealth $W_t(x)$ reach a given benchmark bar (large enough) in a given temporal horizon T, where
$$\tau_{\text{win}}(x,bar,T)=inf\{t>0:W_t(x)>bar\}\wedge T$$
ensuring that it captures the highest possible returns.
2. We identify the loser portfolio x that minimizes the expected first time $E(\tau_{lose}(x,1/bar,T))$ the portfolio wealth $W_t(x)$ reach the reciprocal benchmark 1/bar (small enough) in a given temporal horizon T, where
$$\tau_{lose}(x,1/bar,T)=inf\{t>0:W_t(x)<1/bar\}\wedge T.$$

These problems are expressed as:

1. for the winner portfolio:
$$min_x\, E(\tau_{\text{win}}(x,bar,T))$$
$$s.t.\sum_{i=1}^{n} x_i = 1;\ x_i\geq 0;\ i=1,\dots,n \tag{5}$$
2. for the loser portfolio:
$$min_y\, E(\tau_{lose}(y,1/bar,T))$$
$$s.t.\sum_{i=1}^{n} y_i = 1;\ y_i\geq 0;\ i=1,\dots,n \tag{6}$$

We generally observe that historical series of the portfolio of returns solution of problem (5) is generally negatively correlated with the historical series of the portfolio of returns solution of problem (6) with a strong negative correlation (in particular during the market crises periods). In order to get profit from this observed

return behavior we propose an arbitrage strategy obtained investing in the arbitrage portfolio of weights x-y, where x and y denote the portfolios weights respectively of the winner (solution of 5) and loser (solution of 6) portfolios.

These constraints enforce non-negativity, precluding short selling within both winner and loser portfolios. By optimizing these criteria, the methodology allows for the construction of portfolios that balance risk and return taking into account the returns behavior, accommodating the Markovian non-Gaussian nature of stock returns, which is often seen in financial markets, as highlighted in recent studies in financial literature. This aspect of non-parametric momentum strategies, which incorporate any distributional anomaly (such that, the fat tails and skewness) observed in the distribution of returns, provides more realistic insights compared to traditional models that assume normally distributed returns.

In addition to these risk-adjusted methods, the momentum strategy can also benefit from further enhancements by incorporating adaptive techniques that adjust portfolio weights dynamically based on market conditions. For instance, machine learning algorithms, such as reinforcement learning, can be integrated to continually optimize portfolio allocations by adapting to new data and market conditions in real time. This dynamic approach could lead to superior performance in rapidly changing market environments.

This method not only refines momentum strategies but also ensures computational efficiency, as emphasized in Ortobelli et al. (2019), by leveraging algorithms capable of handling complex distributions, such as Lévy processes. These improvements make the approach highly applicable in real-world portfolio management, especially when managing portfolios with non-normal return distributions. The integration of computational techniques like Monte Carlo simulations and importance sampling can further refine the optimization process, enabling faster and more accurate decision-making.

Moreover, as financial markets continue to evolve, the robustness of momentum strategies could be tested in emerging markets or applied to alternative assets, such as cryptocurrencies or commodities. These markets often exhibit different risk-return characteristics, which can be captured using adapted momentum-based strategies. By extending the approach to non-traditional assets, investors can explore new opportunities for diversification and enhanced returns.

4 Empirical Results

The empirical analysis is conducted on a dataset of daily, weekly, and monthly returns of the S&P 500 constituents over a 20-year period (2002–2021), sourced from Refinitv Datastream. To strictly avoid survivorship bias, the dataset comprises 982 assets, including both active equities and delisted companies that were part of the index during the examined timeframe. Missing data were handled using previous-close prices, and returns were adjusted for dividends and stock splits. A "rolling window" approach is employed, recalibrating the portfolio every two months. The algorithm assumes that returns follow a Markov chain to optimize the expected first passage times and the expected current maximum and minimum of the wealth process. The analyzed data covers three major financial crises: the 2008 subprime crisis, the 2011 European sovereign debt and credit risk crisis, and the 2020 COVID-19 pandemic.

4.1 Daily Returns

Daily returns are analyzed over three different time horizons: seven, fourteen, and twenty-one days. When considering daily data, the "winner" strategy generally presents lower levels of wealth compared to the loser strategy. This outcome indicates that applying a reverse arbitrage strategy is more appropriate. The 21-day horizon delivers the strongest long-term performance; however, all strategies experience a pronounced decline in performance during periods of market stress. Among the different horizons, the intermediate 14-day time horizon offers a favorable balance between short-term reactivity and long-term stability.

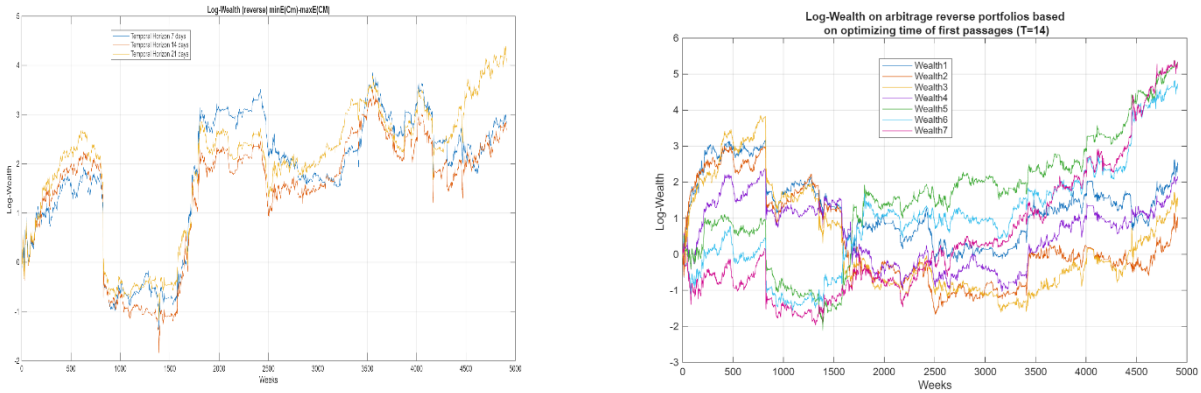


Figure 1(a) and Figure 1(b) Daily Frequency Results. (a) Reverse aarbitrage portfolio strategies based on the optimization of the current maximum and current minimum considering daily returns. (b) Reverse arbitrage portfolio strategies based on first passage times considering daily returns (T=14 days)

4.2 Weekly Returns

The analysis is also conducted using weekly returns across horizons of four, eight, and twelve weeks. Unlike the results obtained with daily data, both strategies show that the "winner" strategy presents higher wealth compared to the strategy of the loser. As a result, the use of reverse arbitrage is not appropriate in this context. Arbitrage portfolio strategies based on the optimization of the current maximum and current minimum show very positive results even during periods of market stress, with the four-week horizon presenting a sharp positive peak during the subprime crisis. Portfolio strategies based on first passage times show a greater sensitivity to market instability, but the twelve-week horizon demonstrates a more stable and better performance.

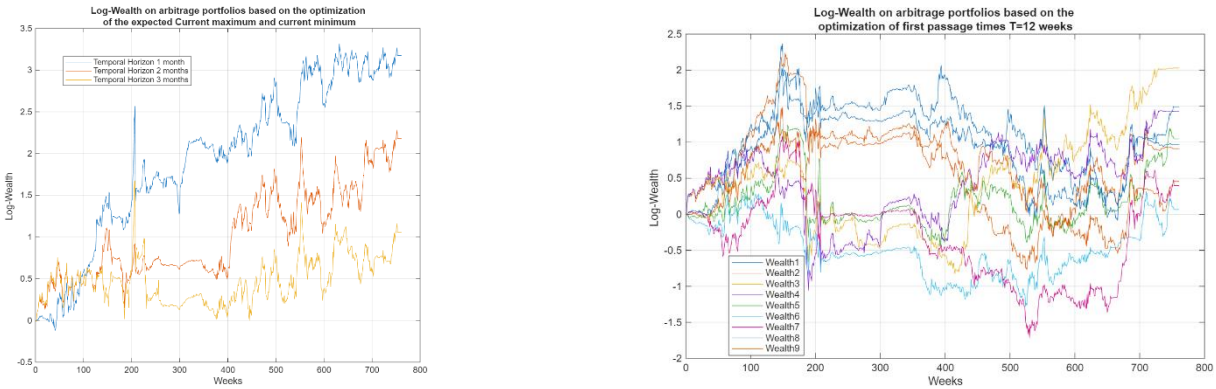


Figure 2(a) and Figure 2(b) Weekly Frequency Results. (a) Arbitrage portfolio strategies based on the optimization of the current maximum and current minimum considering weekly returns. (b) Arbitrage portfolio strategies based on first passage times considering weekly returns (T=12 weeks)

4.3 Monthly Returns

In the case of monthly returns, the effectiveness of arbitrage strategies depends on the time horizon considered (T = 2, 4, or 6 months). With monthly data, the "winner" strategy generally presents lower wealth than the strategy of the loser, making it appropriate to apply a reverse arbitrage strategy over short- and medium-term horizons. On the other hand, in the long run (six months), the strategy of the winner tends to outperform the strategies of the loser.

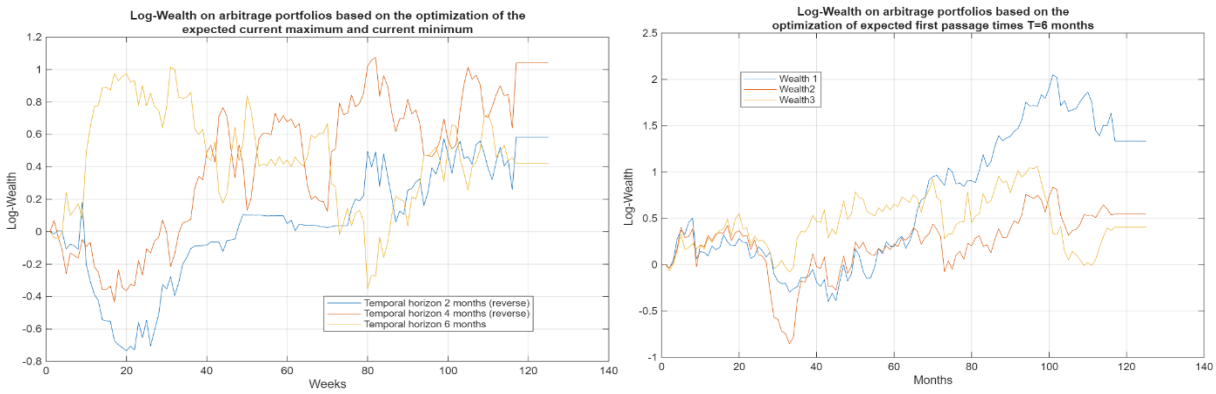


Figure 3(a) and Figure 3(b) Monthly Frequency Results. (a) Reverse arbitrage portfolio strategies based on the optimization of the current maximum and current minimum considering monthly returns. (b) Arbitrage portfolio strategies based on first passage times considering monthly returns (T=6 months)

5 Conclusions and Limitations

DAILY	WEEKLY	MONTHLY
Reverse arbitrage	LOSER: Lower future wealth WINNER: Higher future wealth	The effectiveness of the arbitrage strategy depends on the time horizon.

Table 1 Summary of empirical results by data frequency

The results of the arbitrage and momentum strategies show significant differences depending on data frequency. With daily returns, reverse arbitrage generated greater future wealth, whereas with weekly returns, "winner" portfolios outperformed. For monthly data, effectiveness depends on the time horizon. Within our sample, winner-loser arbitrage strategies demonstrated remarkable resilience during market crashes. However, these findings are bounded by certain limitations that suggest avenues for future research. First, the generalizability of these results is currently limited to the US large-cap equity market (S&P 500); future studies should test this framework across emerging markets or different asset classes. Second, from a practical perspective, high-frequency daily rebalancing implies significant transaction costs and liquidity constraints, which could partially erode the theoretical arbitrage profits observed ex-post. Finally, while the use of homogeneous Markov chains provides computational efficiency, future research could explore non-homogeneous or regime-switching models to better capture time-varying volatilities. Despite these limitations, this study highlights the practical applicability of path-dependent optimization for real-world risk management.

References

Applebaum D (2009) Lévy Processes and Stochastic Calculus (2nd ed.). Cambridge University Press. https://doi.org/10.1017/cbo9780511809781

Asness CS, Moskowitz TJ, Pedersen LH (2013) Value and momentum everywhere. Journal of Finance 68:929–985. https://doi.org/10.1111/jofi.12021

Barroso P, Santa-Clara P (2015) Momentum has its moments. Journal of Financial Economics 116:111–120. https://doi.org/10.1016/j.jfineco.2014.11.010

Blitz D C, Huij J, Martens M (2011) Residual momentum. Journal of Empirical Finance 18:506–521. https://doi.org/10.1016/j.jempfin.2011.01.003

Cont R, Tankov P (2003) Financial Modelling with Jump Processes. Chapman & Hall/CRC. https://doi.org/10.1201/9780203485217

D'Amico G (2006) Statistical inference for Markov chain European option: estimating the price, the bare risk and the theta by historical distributions of Markov chain. Journal of Statistics & Management Systems 9:737-754. https://doi.org/10.1080/09720510.2006.10701233

Daniel K, Moskowitz T J (2016) Momentum crashes. Journal of Financial Economics 122:221–247. https://doi.org/10.1016/j.jfineco.2015.12.002

Duan J C, Simonato J-G (2001) American option pricing under GARCH by a Markov chain approximation. Journal of Economic Dynamics and Control 25:1689–1718. https://doi.org/10.1016/s0165-1889(00)00003-8

Ehsani S, Linnainmaa JT (2022) Factor momentum and the momentum factor. Journal of Finance 77:1877–1919. https://doi.org/10.1111/jofi.13131

Huang D, Li J, Wang L, Zhou G (2020) Time-series momentum: Is it there? Journal of Financial Economics 135:774–794. https://doi.org/10.1016/j.jfineco.2019.08.004

Iaquinta G, Ortobelli S, Angelelli E (2011) GARCH type portfolio selection models with the Markovian approach. International Journal of Mathematical Models and Methods in Applied Sciences 5:308–315. https://hdl.handle.net/10446/25510

Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. Journal of Finance 48:65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Kardaras C, Platen E (2010) Minimizing the expected market time to reach a certain wealth level. SIAM Journal on Financial Mathematics 1:16–24. https://doi.org/10.1137/080741124

Kemeny J G, Snell J L (1976) Finite Markov Chains. Springer. https://doi.org/10.1007/978-1-4684-9455-6_4

Meyn S P, Tweedie R L (2009) Markov Chains and Stochastic Stability. Springer Science & Business Media.

https://doi.org/10.1017/cbo9780511626630

Norris JR (1998) Markov Chains. Cambridge University Press. https://doi.org/10.1017/cbo9780511810633

Ortobelli S, Angelelli E, Ndoci A (2019) Timing portfolio strategies with exponential Lévy processes. Computational Management Science 16:97–127. https://doi.org/10.1007/s10287-018-0332-y

Schoutens W (2003) Lévy Processes in Finance: Pricing Financial Derivatives. Wiley. https://doi.org/10.1002/0470870230

Zsurkis G, Nicolau J, Rodrigues P M M (2024) First passage times in portfolio optimization: A novel nonparametric approach. European Journal of Operational Research 312:1074-1085. https://doi.org/10.1016/j.ejor.2023.07.044

Choosing a Value-at-Risk UC Backtest across Standard Tolerance Levels: An Exact-Size Power Analysis

Krzysztof Piontek[1], Maria Mach-Król[2]

Abstract. We compare the finite-sample power of three unconditional coverage (UC) backtest configurations for Value-at-Risk: Kupiec, Pérignon–Smith, and a three-threshold Kratz–Lok–McNeil-type likelihood-ratio test. All use a unified exact Lehmann γ-randomization framework; size is analytical for Kupiec and calibrated to the stated enumeration tolerance for the multilevel tests, while powers are computed by finite-sample summation without Monte Carlo uncertainty. The data-generating process is a Student-t distribution calibrated to daily DAX log-returns over 2021–2026; misspecification shifts the degrees-of-freedom parameter at fixed location and scale, producing non-linear spillovers across $q \in \{0.01, 0.025, 0.05\}$ for $T \in \{100, \ldots, 1000\}$. Three findings emerge. First, the exact two-sided uniformly most powerful unbiased (UMPU) Kupiec construction removes the *power-less-than-size* pathology of discrete regulatory settings; the same pathology afflicts the deep-pair multilevel construction at $T = 100$. Second, multilevel dominance is not uniform: the three-threshold test wins in shallow tails at longer windows but is outperformed by single-threshold Kupiec configurations in deeper regimes. Third, under risk over-statement, power is non-monotone in q and often strongest at the FRTB-equivalent anchor $q = 0.025$ rather than at the deepest Basel level. Conditional on the examined fixed-scale deformation, the required windows for power 0.80 exceed the standard regulatory window several-fold.

Keywords: VaR backtesting, model risk, randomization, unconditional coverage, test power, single-level and multinomial tests

JEL Classification: C12, C15, C52, G21, G32

1 Introduction

Value-at-Risk (VaR) and Expected Shortfall (ES) remain the core market-risk metrics under the Basel III reforms and the Fundamental Review of the Trading Book (FRTB; BCBS 2016, 2019). Although the FRTB capital charge is computed from ES at the 97.5% level, daily desk-level backtests are still conducted on VaR at 99% and 97.5%, so the finite-sample behaviour of VaR-based tests continues to drive validation decisions. Because the cost of failing to reject a misspecified model may exceed the cost of rejecting a correct one, backtest power is a first-order concern for supervisors and model risk managers.

The object of a UC backtest is simple: over T trading days, one records whether each realized loss exceeded the model-implied VaR (a *violation*) and asks whether the observed frequency is consistent with the nominal tolerance q. We study three likelihood-ratio configurations retaining different amounts of tail information. Kupiec (1995) compresses violations at one level into a single count. Pérignon and Smith (2008, PS) monitor an ordered pair, splitting the tail into two nested zones. Our three-threshold test adapts the multinomial construction of Kratz, Lok and McNeil (2018, KLM), who study at least four VaR levels for implicit ES backtesting under the Emmer–Kratz–Tasche (2015) approximation. The practical question is which tail-partition dimensionality to use in each examined tolerance regime.

Three sources of finite-sample distortion remain insufficiently isolated in the published evidence. The first is *discreteness*: at regulatory tolerance levels and short windows, the discrete distribution of violation counts places a probability atom at the boundary of any candidate rejection region, so no deterministic critical value attains exact size. The second is *calibration heterogeneity*: published comparisons rely either on asymptotic critical values invalid when Tq is small (Kaszyński et al. 2020) or on heterogeneous Monte Carlo schemes, so reported

[1] Wroclaw University of Economics and Business, Department of Financial Investments and Risk Management, ul. Komandorska 118/120, 53-345 Wroclaw, Poland, krzysztof.piontek@ue.wroc.pl, ORCID 0000-0001-9197-561X.
[2] VSB – Technical University of Ostrava, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic, maria.antonina.mach-krol@vsb.cz, ORCID 0000-0002-7234-7340.

power differences may partly reflect implementation artefacts. The third is *deformation specification*: the convention of Pérignon and Smith (2008), Leccadito et al. (2014) and Kratz et al. (2018) deforms each tail probability multiplicatively and independently — mathematically convenient, but not necessarily corresponding to a common underlying distributional perturbation.

We address the three sources jointly. All tests use a single Lehmann γ-randomization scheme: an analytically exact two-sided uniformly most powerful unbiased (UMPU) construction for binomial Kupiec, and upper-statistic boundary randomization for multinomial PS and the three-threshold KLM-type test, whose near-null power we check numerically. The Student-t data-generating process (DGP) is calibrated to daily DAX log-returns (2021–2026); deviation from H_0 shifts $\nu \to \nu^*$ at fixed location and scale and is anchored at $q_{\text{anchor}} = 0.025$. Size equals α analytically for Kupiec and within the stated enumeration tolerance for PS and KLM; powers are obtained by finite-sample summation with zero Monte Carlo uncertainty.

The individual ingredients are well established: randomized exact tests belong to the classical repertoire (Lehmann and Romano 2022), Dufour (2006) develops general finite-sample Monte Carlo inference, and Ziggel et al. (2014) study improved VaR backtests; low power in short windows is documented in Kupiec (1995) and Piontek (2010). The contribution is threefold. First, joint exact-size calibration of the three configurations removes calibration heterogeneity as a confound. Second, deformation acts on a distributional parameter and generates endogenous non-linear spillover across levels. Third, the design establishes non-uniform multilevel dominance and a grid-optimal tolerance $q^*(T; \theta_{\text{ref}})$ that can lie at the anchor rather than the deepest regulatory level.

2 Methodology

2.1 Notation, hypotheses, and regulatory regimes

Let $\{r_t\}_{t=1}^T$ be daily log-returns and $\widehat{\text{VaR}}_t(q)$ the model-implied forecast at tolerance level $q \in (0,1)$. Violation indicators are $I_t(q) = \mathbb{1}\{r_t < -\widehat{\text{VaR}}_t(q)\}$. Under correct calibration and independent identically distributed violations, $N(q) = \sum_{t=1}^T I_t(q) \sim \text{Bin}(T, q)$. Departures from H_0 are parameterized multiplicatively at the anchor:

$$H_0\colon\ \text{E}[I_t(q)] = q, \qquad H_{1,\theta_{\text{ref}}}\colon\ p_{\text{eff}}(q_{\text{anchor}}) = \theta_{\text{ref}} \cdot q_{\text{anchor}}, \quad \theta_{\text{ref}} \neq 1. \qquad (1)$$

The reference multiplier θ_{ref} is tied to $q_{\text{anchor}} = 0.025$; effective probabilities at the other levels follow endogenously from the mechanism below. The scope is restricted to unconditional coverage, following Christoffersen's (1998) decomposition of UC, independence and conditional coverage: UC is the property implied directly by the definition of VaR as the q-quantile of the loss distribution; Section 4 discusses the implications. The levels $q \in \{0.01, 0.025, 0.05\}$ correspond to the Basel III internal models approach (IMA) VaR backtest at 99%, the FRTB-equivalent 97.5% ES tail probability, and a shallow level used in the post-Basel power literature. The pairs define $\text{PS}^{(0.01,0.025)}$ and $\text{PS}^{(0.025,0.05)}$. For compactness, the three-threshold KLM-type LR configuration is denoted KLM^{full}, where *full* refers only to the complete three-level grid examined here.

2.2 Three UC tests as tail-information projections

All three tests are likelihood-ratio statistics on an ordered vector of category counts and differ only in the dimensionality K of the tail partition. An increasing grid $0 < q_1 < \cdots < q_K < 1$ splits each day's outcome into $K+1$ ordered cells: cell K collects returns below the deepest VaR threshold, cells $K-1, \ldots, 1$ the bands between consecutive thresholds, and cell 0 the non-violation region; under H_0 the cell probabilities are $p_K = q_1$, $p_k = q_{K-k+1} - q_{K-k}$ for $k = 1, \ldots, K-1$, and $p_0 = 1 - q_K$. Retention of tail information grows with K, but cell sparsity worsens at regulatory tolerance levels. Table 1 summarizes the constructions.

Let $\hat{p} = N/T$. The Kupiec proportion-of-failures (POF) statistic at a single level q is

$$\text{LR}_{\text{POF}}(q) = 2\left[N \log\frac{\hat{p}}{q} + (T-N)\log\frac{1-\hat{p}}{1-q}\right], \qquad \text{LR}_{\text{POF}} \overset{a}{\sim} \chi_1^2. \qquad (2)$$

The PS and KLM tests are special cases of the multinomial likelihood-ratio (G^2) statistic on the $K+1$ ordered cells:

$$\text{LR}_{\text{multi}}^{(K)} = 2\sum_{k=0}^{K} O_k \log\left(\frac{O_k}{T\,p_k}\right), \qquad \text{LR}_{\text{multi}}^{(K)} \overset{a}{\sim} \chi_K^2, \qquad (3)$$

where O_k is the observed and Tp_k the expected count in cell k under H_0, with the convention $0\log 0 = 0$.

Test	K	Monitored levels	Statistic	Asympt. ref.	Randomization
Kupiec	1	single level q	(2)	chi-sq(1)	two-sided UMPU (γL, γU)
PS	2	ordered pair of levels	(3)	chi-sq(2)	upper-statistic boundary γ
KLM full	3	triple (0.01, 0.025, 0.05)	(3)	chi-sq(3)	upper-statistic boundary γ

Table 1 The three UC backtests as projections of the tail. Kupiec attains size α analytically; PS and KLM do so within the stated enumeration tolerance

2.3 Exact Lehmann γ-randomization, step by step

Because discreteness leaves no deterministic critical value attaining exact size (Section 1), a randomized test resolves the problem analytically through the test function

$$\varphi(x) = \begin{cases} 1, & S(x) > c^*, \\ \gamma, & S(x) = c^*, \\ 0, & S(x) < c^*, \end{cases} \qquad (4)$$

where S is the test statistic and $\varphi(x)$ is the probability of rejecting when x is observed; the pair (c^*, γ) is chosen so that $\mathrm{E}_{H_0}[\varphi] = \alpha$ holds exactly. For the Kupiec test, whose statistic is U-shaped in $N \sim \mathrm{Bin}(T, q)$, a two-sided construction is required: the UMPU test (Lehmann and Romano 2022, Ch. 4) randomizes on two boundary atoms c_L, c_U with probabilities (γ_L, γ_U) solving the two-equation system

$$\mathrm{P}_{H_0}(\text{reject}) = \alpha, \qquad \left.\frac{\partial \pi(\theta)}{\partial \theta}\right|_{\theta=1} = 0, \qquad (5)$$

where $\pi(\theta)$ is the power function. The second condition enforces local unbiasedness and, for the one-parameter exponential family, guarantees $\pi(\theta) \geq \alpha$ for all $\theta \neq 1$; this algebraic guarantee eliminates the *power-less-than-size* pathology of the discrete binomial likelihood ratio. As a worked example, for $T = 250$, $q = 0.01$, $\alpha = 0.05$ the analytical solution is $(c_L, c_U, \gamma_L, \gamma_U) = (0,6,0.399,0.144)$: the test rejects with probability 0.399 at zero observed violations and 0.144 at exactly six, raising power at $\theta_{\text{ref}} = 2/3$ from 0.003 under naive γ-randomization on the upper tail of N only to 0.115.

For the multilevel tests, only large values of the convex G^2 statistic indicate misfit, irrespective of the direction in the multinomial probability space; a single γ on the upper-statistic boundary therefore calibrates size. We do not derive a directly comparable UMPU multinomial rule and instead check near-null power numerically over a refined deformation grid (Section 3), with implications discussed in Section 4. PS and KLM sizes and powers are computed by finite-sample enumeration within a concentration window whose missed null mass is below 10^{-10}; no Monte Carlo simulation is used.

Market-style deformation of the Student-t DGP

The DGP is a location-scale Student-t distribution with degrees-of-freedom parameter ν calibrated to daily DAX log-returns from Yahoo Finance over 2021–2026 ($n = 1272$). Tukey (1977) IQR cleaning at $k = 3$ removes 8 observations (0.63%) and yields $\hat{\nu} = 5.21 \pm 0.80$ (maximum-likelihood estimate, MLE, $\pm$ standard error). Because trimming removes observations informative about tail thickness, it mechanically raises $\hat{\nu}$; the uncleaned sample gives $\hat{\nu} = 4.00 \pm 0.47$, while aggressive symmetric trimming of 1% gives $\hat{\nu} = 9.92$. The pipeline input is $\hat{\nu} = 5.0$, a rounding of the IQR-cleaned estimate; for readability, the same symbol $\hat{\nu}$ denotes this input below. Section 3 verifies the reported dominance relations over $\hat{\nu} \in \{4,5,6\}$, which contains the uncleaned estimate; the unused trim-1% calibration lies outside and is not covered by this check. An S&P 500 cross-check under the same protocol gives $\hat{\nu} = 5.21 \pm 0.85$ (7 observations removed). This is a calibration cross-check, not a robustness replication; the $\hat{\nu}$-sweep plays the latter role.

Departure from H_0 shifts $\nu \to \nu^*$ while holding location and scale fixed, rather than perturbing individual quantiles. The design reduces to five steps: (i) calibrate $\hat{\nu}$ to market data; (ii) for a given θ_{ref} in (1), solve

$$\theta_{\text{ref}} \cdot q_{\text{anchor}} = F_{\nu^*}(F_{\hat{\nu}}^{-1}(q_{\text{anchor}})) \qquad (6)$$

for ν^*, where F_ν is the Student-t CDF; (iii) propagate the deformation to the remaining levels through $p_{\text{eff}}(q_k) = F_{\nu^*}(F_{\hat{\nu}}^{-1}(q_k))$; (iv) obtain the exact discrete distribution of the cell counts under H_1 from the vector p_{eff}; (v) evaluate $\pi(\theta_{\text{ref}}) = \mathrm{E}_{H_1}[\varphi]$ by finite-sample summation. The mapping in step (iii) is non-linear: for $\theta_{\text{ref}} = 1.5$ with $\hat{\nu} = 5$ we obtain $\nu^* = 3.30$ and induced multipliers $\theta_{\text{eff}}(0.01) = 1.89$, $\theta_{\text{eff}}(0.025) = 1.50$, $\theta_{\text{eff}}(0.05) = 1.29$; for $\theta_{\text{ref}} = 2/3$, $\nu^* = 7.93$ and $\theta_{\text{eff}} = (0.499,0.667,0.790)$. The spillover concentrates deformation in the deepest tail and dilutes it in the shallowest, representing a plausible degrees-of-freedom misspecification at fixed

location and scale. Because $\mathrm{Var}(t_\nu) = \nu/(\nu - 2)$, it changes both tail shape and unconditional variance rather than only variance-standardized kurtosis. The DGP omits volatility clustering by design: the scope is unconditional coverage, and admitting conditional heteroscedasticity would conflate UC with conditional coverage (Christoffersen 1998); the conditional extension is left to future work.

Deformation grid and window sizes

The deformation grid is multiplicative and log-mirror symmetric: $\theta_{\mathrm{ref}} \in \{1/m, m\}_{m\in\{1.0,1.1,1.2,1.3,1.4,1.5\}}$, giving 11 grid points including $\theta_{\mathrm{ref}} = 1$ for the size sanity check. Each mirror pair $(1/m, m)$ is log-equidistant from H_0, so any asymmetry $\pi(1/m) \neq \pi(m)$ is a structural property of the Student-t spillover rather than a grid artefact. The window sizes $T \in \{100,250,500,750,1000\}$ cover the Basel III IMA window ($T = 250$), a longer-window benchmark ($T = 750$), and an asymptotic reference. The nominal level is $\alpha = 0.05$ throughout; a sensitivity check at $\alpha = 0.01$ is summarized in Section 3. We write $\pi(\theta_{\mathrm{ref}}, T)$ and drop arguments where clear.

3 Results

3.1 Exact sizes and near-null power checks

All six configurations are calibrated to size $\alpha = 0.05$: analytically for Kupiec UMPU and within the enumeration tolerance (missed null mass below 10^{-10}) for PS and KLM. The central column of Table 2 and middle row of Table 3, both at $\theta_{\mathrm{ref}} = 1$, document this property.

A 31-point refinement of the deformation grid checks near-null power for the upper-statistic multilevel calibrations. For $T \geq 250$, all three are approximately size-balanced on the examined grid: $\min_\theta \pi(\theta) \geq \alpha - 0.0003$, with the small deficit nearest $\theta = 1$. For $T = 100$, the deep pair $\mathrm{PS}^{(0.01,0.025)}$ has $\pi(2/3) = 0.012$, one-quarter of nominal size, a material discreteness pathology absent from Kupiec UMPU. Cross-test comparisons therefore start at $T = 250$; this numerical balance improves comparability but does not establish theoretical equivalence of the rejection rules.

3.2 Anchor regime in detail

Table 2 focuses on the FRTB-equivalent anchor regime $q = 0.025$, reporting the exact Kupiec UMPU power on the full multiplicative log-mirror grid. The convex power profile is structurally asymmetric in the mirror sense: $\pi(2/3) < \pi(3/2)$ at every $T \geq 250$, with the gap widening from 0.073 at $T = 250$ to 0.194 at $T = 1000$, largest for the extreme pair and shrinking monotonically toward the null (only 0.027 for the inner pair $(5/6, 6/5)$ at $T = 1000$).

T	2/3	5/7	10/13	5/6	10/11	1	11/10	6/5	13/10	7/5	3/2
100	0.0777	0.0702	0.0630	0.0567	0.0520	**0.0500**	0.0524	0.0594	0.0711	0.0873	0.1077
250	0.1375	0.1124	0.0896	0.0701	0.0558	**0.0500**	0.0568	0.0769	0.1099	0.1551	0.2110
500	0.2413	0.1856	0.1349	0.0925	0.0622	**0.0500**	0.0641	0.1059	0.1742	0.2654	0.3722
750	0.3422	0.2584	0.1805	0.1149	0.0684	**0.0500**	0.0714	0.1352	0.2386	0.3709	0.5147
1000	0.4396	0.3314	0.2273	0.1380	0.0748	**0.0500**	0.0787	0.1648	0.3016	0.4680	0.6333

Table 2 Finite-sample power of Kupiec UMPU at the FRTB-equivalent anchor $q = 0.025$ on the full multiplicative log-mirror grid. The central column $\theta_{\mathrm{ref}} = 1$ contains the analytically exact size $\alpha = 0.050$

3.3 Power across configurations

Table 3 reports the finite-sample power $\pi(\theta_{\mathrm{ref}}, T)$ for all six configurations and four window sizes at the two extreme mirror points $\theta_{\mathrm{ref}} \in \{2/3, 3/2\}$. Bold entries mark the maximum within each $(T, \theta_{\mathrm{ref}})$ block. Three structural patterns govern the results.

Mirror-pair asymmetry. For $\mathrm{Kupiec}^{\mathrm{UMPU}}_{q=0.01}$ the asymmetry $\pi(3/2) - \pi(2/3)$ grows from 0.106 at $T = 250$ to 0.238 at $T = 1000$, with the same sign across all six configurations. The asymmetry is structural — the log-mirror grid guarantees equal log-distance from H_0 within each pair (Section 2).

Non-uniform multilevel dominance. The ranking at $T = 1000$, $\theta_{\mathrm{ref}} = 3/2$ is

$$\text{Kupiec}_{q=0.01}^{\text{UMPU}}\ (0.673) > \text{PS}^{(0.01,0.025)}\ (0.649) > \text{Kupiec}_{q=0.025}^{\text{UMPU}}\ (0.633) >$$

$$> \text{KLM}^{\text{full}}\ (0.598) > \text{PS}^{(0.025,0.05)}\ (0.554) > \text{Kupiec}_{q=0.05}^{\text{UMPU}}\ (0.513).$$

KLM^{full} trails $\text{Kupiec}_{q=0.01}$ by 7.5 p.p. but leads $\text{Kupiec}_{q=0.05}$ by 8.5 p.p.; the advantage of the full multinomial test depends on the regime against which it is compared.

Non-monotonicity in q. Within the Kupiec family for $\theta_{\text{ref}} = 2/3$ and $T = 250$ the ranking by power is $\pi_{q=0.01} = 0.115 < \pi_{q=0.025} = 0.138 > \pi_{q=0.05} = 0.122$; the maximum lies at the anchor rather than at either grid endpoint. The same non-monotonicity persists for all T on at least one mirror side.

θ_ref	Configuration	T = 250	T = 500	T = 750	T = 1000
	Kupiec q=0.01 (UMPU)	0.115	0.223	0.326	0.434
	Kupiec q=0.025 (UMPU)	**0.138**	**0.241**	**0.342**	**0.440**
2/3	Kupiec q=0.05 (UMPU)	0.122	0.198	0.278	0.354
(risk	PS (0.01, 0.025)	0.107	0.191	0.318	0.402
overstated)	PS (0.025, 0.05)	0.114	0.194	0.279	0.365
	KLM full	0.099	0.197	0.281	0.370
1	**axis of symmetry**	**0.050**	**0.050**	**0.050**	**0.050**
	Kupiec q=0.01 (UMPU)	**0.221**	**0.400**	**0.557**	**0.673**
	Kupiec q=0.025 (UMPU)	0.211	0.372	0.515	0.633
3/2	Kupiec q=0.05 (UMPU)	0.170	0.293	0.409	0.513
(risk	PS (0.01, 0.025)	0.180	0.377	0.497	0.649
understated)	PS (0.025, 0.05)	0.168	0.310	0.438	0.554
	KLM full	0.167	0.317	0.469	0.598

Table 3 Exact power $\pi(\theta_{\text{ref}}, T)$ across six test configurations and four window sizes, under exact Lehmann γ-randomization with $\alpha = 0.05$. Market-style DGP with $\hat{\nu} = 5$ and $q_{\text{anchor}} = 0.025$. The middle row marks the axis of symmetry $\theta_{\text{ref}} = 1$ at which $\pi = \alpha$ exactly. Bold values are the maximum within each (T, θ_{ref}) block

3.4 Window-size requirements

Table 4 reports the *exact* required window size T^* for power at least 0.80, computed within the same randomized UMPU framework as the power tables — consistent with the finite-sample argument of this paper, no asymptotic approximation is involved. (Under the proportional deformation, the one-sided Wald approximation $T^* = (z_{1-\alpha} + z_{0.8})^2\, p(1-p)/(p-q)^2$ understates the exact values by 4–6%: 3654 against 3818 at $q = 0.01$.) Under the proportional convention $p = 1.5q$, T^* decreases with q: from 15.3 times the Basel III window in the deep regime to 2.9 times in the shallow one. Under the market-DGP spillover with $\theta_{\text{ref}} = 1.5$ the ordering *reverses*: the deep regime concentrates the strongest effective deformation ($\theta_{\text{eff}}(0.01) = 1.89$) and requires the least data ($T^* = 1361$), while the shallow regime, where the spillover dilutes the signal, requires the most ($T^* = 1994$). The operational conclusion is robust to either reading: a single non-rejection in the regulatory window does not constitute strong evidence of correct VaR calibration.

Tolerance regime	q	T^*: $p = 1.5q$	T^*/T Basel	T^*: spillover	T^*/T Basel
Basel III deep	0.010	3818	15.3	1361	5.4
FRTB-equivalent mid	0.025	1499	6.0	1499	6.0
Academic shallow	0.050	725	2.9	1994	8.0

Table 4 Exact required window size T^* for $\pi \geq 0.80$ (Kupiec UMPU, $\alpha = 0.05$), under the proportional deformation $p = 1.5q$ and under the market-DGP spillover anchored at $\theta_{\text{ref}} = 1.5$; $T_{\text{Basel}} = 250$. The two entries at $q = 0.025$ coincide by construction, since $\theta_{\text{eff}}(q_{\text{anchor}}) = \theta_{\text{ref}}$

3.5 Robustness of the ranking

Two checks address the stability of the central ranking. First, a sweep over $\hat{\nu} \in \{4,5,6\}$ — containing the uncleaned DAX estimate (4.00) — recomputes the full pipeline, including spillover, at each value. Table 5 shows that $\text{Kupiec}_{q=0.01}$ dominates KLM^{full} throughout (gap 6.6 to 8.0 p.p.), while KLM's advantage over shallow Kupiec widens from 6.8 to 9.7 p.p.; the PS configurations (not shown) remain intermediate. Both gaps grow monotonically in $\hat{\nu}$, so the preprocessing choice moves the calibrated input rather than the conclusions; thinner-tailed calibrations are unlikely to reverse either relation. Second, at $\alpha = 0.01$ the complete six-test ordering changes in six of eight (θ_{ref}, T) blocks. Nevertheless, under risk understatement $\text{Kupiec}_{q=0.01}$ remains above KLM^{full} at every T, and KLM remains above $\text{Kupiec}_{q=0.05}$ for $T \geq 500$; under risk overstatement, $q = 0.025$ remains the strongest Kupiec configuration at every reported T. The complete ordering coincides with Table 3 only for $\theta_{\text{ref}} = 3/2$ at $T = 750$ and 1000; detailed tables are available on request.

$\hat{\nu}$	Kup 0.01	Kup 0.025	Kup 0.05	KLM full	Kup 0.01 − KLM	KLM − Kup 0.05
4	0.658	0.633	0.524	0.592	0.066	0.068
5	0.673	0.633	0.513	0.598	0.075	0.085
6	0.683	0.633	0.506	0.603	0.080	0.097

Table 5 Sensitivity of the exact power ranking to the calibrated tail parameter $\hat{\nu}$, at $\theta_{\text{ref}} = 3/2$, $T = 1000$, $\alpha = 0.05$; full exact pipeline, including the $\hat{\nu}$-specific spillover

4 Discussion

4.1 Two sources of asymmetry: spillover and discreteness

Two mechanisms make the power curve asymmetric around $\theta_{\text{ref}} = 1$. The first is the structural spillover of the Student-t distribution: the curvature of the inverse CDF makes the effective deformation in the deep tail stronger for $\theta_{\text{ref}} > 1$ than for the mirrored $\theta_{\text{ref}} < 1$. At $T = 1000$ the mirror-pair gap $\pi(3/2) - \pi(2/3)$ reaches 0.238 for $\text{Kupiec}_{q=0.01}^{\text{UMPU}}$ and 0.228 for KLM^{full}.

The second mechanism is test discreteness, concentrated at low q. For $q = 0.01$ and $T = 250$ the atom $P(N = 0 \mid H_0) = 0.081$ exceeds α, so a deterministic one-sided level-α Kupiec test cannot include $\{N = 0\}$ in its rejection region, producing $\inf_{\theta \neq 1} \pi(\theta) < \alpha$ — the *power-less-than-size* pathology, which the UMPU construction resolves at no additional sample cost. Section 3 shows the same pathology for the upper-statistic multilevel calibration at $T = 100$ in the deep pair: in short regulatory windows, discreteness is a first-order determinant of what a backtest can detect.

4.2 How comparable are the test constructions?

Kupiec uses a two-sided unbiased rejection region, whereas the multilevel tests use the upper tail of G^2; rankings may therefore reflect both tail information and rejection-rule construction. For $T \geq 250$ the near-null check shows approximate size balance on the examined grid, but it does not establish theoretical equivalence. Under over-deformation ($\theta_{\text{ref}} > 1$), where the headline reversal is observed, all procedures respond to excess violations, making the comparison operationally informative. In the under-deformation region ($\theta_{\text{ref}} < 1$), part of the single-threshold advantage can stem from Kupiec's two-sided UMPU rule; results should be read as comparisons of the implemented tests, not equal constructions. We do not derive a directly comparable exact unbiased two-sided multinomial rule for the present deformation, so this remains a limitation rather than a non-existence claim.

4.3 Non-uniform multilevel dominance

The classical argument for multilevel backtests (Pérignon and Smith 2008; Leccadito et al. 2014; Kratz et al. 2018) is that pooling information across levels improves discrimination in the data-sparse deep tail. Under the examined fixed-scale deformation the advantage is regime-dependent, because two mechanisms operate against the multilevel construction. As an asymptotic heuristic, the degrees-of-freedom penalty raises the reference threshold from $\chi^2_{1;0.95} = 3.84$ for Kupiec to $\chi^2_{3;0.95} = 7.81$ for KLM, while the Kullback–Leibler signal $2TD_{\text{KL}}(p_{H_1} \parallel p_{H_0})$ at $\theta_{\text{ref}} = 3/2$ grows only modestly across tests. Signal dilution compounds the effect: under (6), two of the four cells aggregated by KLM^{full} carry weaker deformation than the deepest cell viewed alone.

When one tolerance level concentrates the strongest local deformation, a single-threshold test extracts the signal more efficiently than a multilevel test paying a degrees-of-freedom penalty for access to weaker cells; when the local signal is weak, the multilevel test borrows strength from deeper cells and wins. The $\hat{\nu}$-sweep of Table 5 confirms both dominance relations are stable over the empirically relevant calibration range. The reversal is conditional on the deformation mechanism: a proportional deformation acting identically on all levels would restore the more uniform classical ranking. For the model validator, the choice of test cannot be separated from the assumed source of model error.

4.4 Window-size requirements and model risk

Across the entire deformation grid and every window size up to $T = 1000$, the conventional benchmark $\pi \geq 0.80$ (Cohen 1988) is attained in no cell of Table 2 — a regularity already noted in earlier simulation evidence (Piontek 2010). Table 4 shows that the required window exceeds the Basel III IMA window by a factor of 2.9 to 15.3 under proportional deformation, and 5.4 to 8.0 under the market spillover. The Basel III traffic-light system (BCBS 2016) mitigates the issue only partially, by translating violation counts into capital multipliers rather than by direct power control. The implication for model risk assessment is structural: backtesting decisions in Basel III windows are made under low test resolution, and the supervisor effectively trades Type II error for capital conservatism. A Basel III non-rejection is not sufficient evidence of adequate calibration and should be supplemented with diagnostics whose finite-sample properties have been independently characterized.

4.5 Grid-optimal tolerance level $q^*(T; \theta_{\text{ref}})$

Under risk overstatement ($\theta_{\text{ref}} = 2/3$), Kupiec power is non-monotone in q. At $T = 750$, the three configurations give $\pi_{q=0.01} = 0.326$, $\pi_{q=0.025} = 0.342$ and $\pi_{q=0.05} = 0.278$; the optimum lies at the anchor rather than either endpoint. By contrast, under risk understatement ($\theta_{\text{ref}} = 3/2$), $q = 0.01$ is the strongest Kupiec configuration at every reported T (Table 3). The interior optimum under risk overstatement balances two forces: relative signal strength rises with depth ($|\theta_{\text{eff}}(0.01) - 1| = 0.501$ against 0.210 at $q = 0.05$), while statistical resolution rises with breadth (the expected count Tq is five-fold higher at $q = 0.05$). A signal-to-noise proxy $z(q) = |T(q - p_{\text{eff}})|/\sqrt{Tq(1-q)}$ — used descriptively, not for calibration — tracks the maximum: $z(0.01) = 1.38$, $z(0.025) = 1.46$, $z(0.05) = 1.32$. Anchoring only normalizes deformation size; power uses the full induced vector p_{eff}. The optimum therefore occurs at $q = 0.025$ because resolution, not relative signal, binds at regulatory window sizes.

The conditional optimum $q^*(T; \theta_{\text{ref}})$ drifts toward deeper tails as T grows and resolution ceases to bind. The regulator currently fixes the validation level *a priori* — Basel III at $q = 0.01$, FRTB at $q = 0.025$ — rather than conditioning it on the available sample. Against risk overestimation ($\theta_{\text{ref}} < 1$) the optimum at $T = 250$ is the anchor $q = 0.025$; a multi-anchor reporting protocol complementing the deepest regulatory level with the anchor raises attainable single-test detection power in that direction by 2.3 p.p. at $T = 250$ (0.138 against 0.115), leaving each individual test and its size unchanged.

4.6 Scope, randomization in practice, and an operational summary

Three limitations bound the operational reach of these results, and the recommendations below are conditional on them. First, the design isolates unconditional coverage under a static DGP: violation clustering, which drives many real-world failures, is outside scope, so the results do not rank tests against dependence alternatives. Second, one degrees-of-freedom mechanism, moving tail shape and variance together, is studied; pure scale misspecification at fixed ν, skewness or contamination may deform the level profile differently. Third, the analysis establishes statistical, not economic, significance: translating power differences into capital terms requires a loss function we do not model. Within these bounds, randomization is a measurement instrument, not a policy proposal: exact γ-randomization calibrates discrete finite-sample comparisons, whereas a mid-p rule avoids operational randomization but does not generally preserve exact size. This trade-off is left for follow-up work.

With these caveats, the rankings imply a conditional choice rule for fixed-scale degrees-of-freedom misspecification. In the Basel III regime ($q = 0.01$), Kupiec UMPU delivers the highest power against strong calibration error, with $\text{PS}^{(0.01,0.025)}$ the closest competitor when a multilevel report is required. In the FRTB-equivalent regime ($q = 0.025$), Kupiec at the anchor is often grid-optimal under risk overestimation. In the shallow regime ($q = 0.05$), for $T \geq 500$ against risk understatement, KLM^{full} outperforms single-threshold Kupiec by aggregating deeper-cell information. The dominant criteria are q and the expected direction of misspecification, not test dimensionality alone.

5 Conclusions

The paper compares the finite-sample power of Kupiec, Pérignon–Smith and a three-threshold KLM-type UC backtest within a unified Lehmann γ-randomization framework under a fixed-scale Student-t degrees-of-freedom deformation. Kupiec UMPU resolves the *power-less-than-size* pathology and restores two-sided detection in the Basel III regime; the deep-pair construction remains subject to it in short windows. Multilevel dominance is non-uniform: $\mathrm{KLM}^{\mathrm{full}}$ trails Kupiec in the middle and deep regimes but wins in the shallow tail at longer windows. The principal $T = 1000$ dominance relations are stable over $\hat{\nu} \in \{4,5,6\}$ and at $\alpha = 0.01$. Under risk overstatement, Kupiec power is non-monotone in q and the grid optimum lies at $q_{\mathrm{anchor}} = 0.025$. No configuration attains power 0.80 for the examined calibration error at $T \leq 1000$, and the required windows exceed the Basel III window several-fold.

The implications are conditional on the fixed-scale degrees-of-freedom mechanism studied here. For supervisors, the low power under this DGP supports multi-anchor reporting or explicit communication of test power in the validation cycle. For the internal model validator, the choice of test cannot be made independently of the expected direction of misspecification, and the UMPU correction is a natural default in the Basel III regime. For risk management, a single non-rejection in the regulatory window remains weak evidence of correct calibration regardless of the test selected.

Natural extensions include conditional coverage under GARCH-t processes; the size–implementation trade-off of mid-p rules; derivation of a directly comparable exact unbiased two-sided multinomial rule for the present deformation; joint $(\mathrm{VaR}, \mathrm{ES})$ backtests under elicitability constraints; and comparison with emerging e-process variants (Wang et al. 2025).

Acknowledgements

The authors thank two anonymous reviewers for comments that improved the paper.

References

Basel Committee on Banking Supervision (BCBS) (2016) Minimum capital requirements for market risk. BCBS d352

Basel Committee on Banking Supervision (BCBS) (2019) Minimum capital requirements for market risk. BCBS d457

Christoffersen P (1998) Evaluating interval forecasts. Int Econ Rev 39(4):841–862

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, NJ

Dufour J-M (2006) Monte Carlo tests with nuisance parameters: a general approach to finite-sample inference and nonstandard asymptotics. J Econom 133(2):443–477

Emmer S, Kratz M, Tasche D (2015) What is the best risk measure in practice? A comparison of standard measures. J Risk 18(2):31–60

Kaszyński D, Kamiński B, Pankratz B (2020) Assessment of the size of VaR backtests for small samples. Przegląd Statystyczny 67(2):114–151

Kratz M, Lok Y H, McNeil A J (2018) Multinomial VaR backtests: a simple implicit approach to backtesting expected shortfall. J Bank Finance 88:393–407

Kupiec P H (1995) Techniques for verifying the accuracy of risk measurement models. J Derivatives 3(2):73–84

Leccadito A, Boffelli S, Urga G (2014) Evaluating the accuracy of value-at-risk forecasts: new multilevel tests. Int J Forecast 30(2):206–216

Lehmann E L, Romano J P (2022) Testing statistical hypotheses, 4th edn. Springer, Cham

Pérignon C, Smith D R (2008) A new approach to comparing VaR estimation methods. J Derivatives 16(2):54–66

Piontek K (2010) The analysis of power for some chosen VaR backtesting procedures: simulation approach. In: Fink A, Lausen B, Seidel W, Ultsch A (eds) Advances in data analysis, data handling and business intelligence. Springer, Heidelberg, pp 481–490

Tukey J W (1977) Exploratory data analysis. Addison-Wesley, Reading MA

Wang Q, Wang R, Ziegel J F (2025) E-backtesting. Manag Sci 72(6):4952–4973. https://doi.org/10.1287/mnsc.2023.01659

Ziggel D, Berens T, Weiß G N F, Wied D (2014) A new set of improved value-at-risk backtests. J Bank Finance 48:29–41

Roughness and Jump Risk in Cryptocurrency Option Smiles: An Econometric Comparison

Michal Polak[1]

Abstract. We calibrate four parsimonious option-pricing models — Heston, Bates, rough Heston, and SVCJ — to 1,989 fifteen-minute Deribit BTC and ETH implied-volatility snapshots over a 21-day window, using a Fourier–cosine pricer, and compare their performance under three evaluation protocols: in-sample RMSE, chronological time-series out-of-sample, and within-snapshot cross-maturity out-of-sample. The model ranking is not invariant. Jumps dominate in sample on BTC while roughness dominates on ETH; the time-series out-of-sample leader differs from the cross-maturity leader in the XM subset. We find that roughness and jumps appear complementary across assets but act as competing representations within an asset, and that model selection on crypto option surfaces must be protocol-aware.

Keywords: option pricing, rough volatility, stochastic volatility with jumps, calibration, Fourier–cosine method, cryptocurrency options

JEL Classification: C58, G13, C13

1 Introduction

The short-maturity smile admits two competing explanations: rough volatility, in which the variance process is less regular than Brownian motion (Gatheral et al, 2018; El Euch and Rosenbaum, 2019), and jumps in price and variance (Bates, 1996; Duffie et al, 2000). Both reproduce the steep near-dated curvature, so on any single surface they are hard to tell apart. Cryptocurrency markets sharpen the question: option liquidity on Deribit is now deep enough to support the surface calibration long applied to equity indices (Alexander et al, 2023), and stochastic-volatility and jump models have already been fitted to Bitcoin options (Hou et al, 2020). Yet the rough and jump mechanisms have not been benchmarked head-to-head on crypto surfaces, let alone under multiple evaluation protocols.

We organise the empirical analysis around three model-comparison questions. First, does rough Heston improve on Heston in sample? Second, do jump-augmented specifications (Bates, SVCJ) improve on Heston in sample? Third, does the richer SVCJ specification provide the most stable out-of-sample performance across protocols? On $N = 1{,}989$ fifteen-minute snapshots per asset spanning a 21-day window, the answers to the first two questions are affirmative on both BTC and ETH, but the asset-level *best* model differs; the third is rejected as a universal ranking — SVCJ is not the consistent out-of-sample winner across assets and protocols. The ranking is not invariant: model selection on crypto surfaces must be protocol-aware. Our contribution is methodological: (i) a model-selection framework that distinguishes calibration stability from surface extrapolation through three complementary evaluation protocols (in-sample RMSE; chronological time-series OOS; cross-maturity OOS); (ii) to our knowledge the first head-to-head benchmark of rough-volatility and jump-diffusion specifications on cryptocurrency option surfaces; and (iii) econometric identification diagnostics for the Hurst exponent and jump intensity under bound-constrained calibration.

2 Models

We write all four models under the forward measure for the log-forward $X_t = \log F_t$ with stochastic variance V_t; the risk-free rate and any cost-of-carry are absorbed into the forward (§3).

Heston (1993).

$$dX_t = -\tfrac{1}{2}V_t\,dt + \sqrt{V_t}\,dW_t^X, \qquad dV_t = \kappa(\theta - V_t)\,dt + \sigma\sqrt{V_t}\,dW_t^V, \tag{1}$$

with $d\langle W^X, W^V\rangle_t = \rho\,dt$ and parameters $(v_0, \kappa, \theta, \sigma, \rho)$ $(K = 5)$.

[1] VSB – Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic, michal.polak@vsb.cz

Bates (1996). Heston plus log-normal price jumps, driven by a Poisson counter N_t of intensity λ:

$$dX_t = (-\tfrac{1}{2}V_t - \lambda m)\,dt + \sqrt{V_t}\,dW_t^X + J\,dN_t, \tag{2}$$

with $\log(1+J) \sim \mathcal{N}(\nu_J, \delta^2)$ and $m = \mathbb{E}[J] = e^{\nu_J + \delta^2/2} - 1$ ($K = 8$).

Rough Heston (El Euch and Rosenbaum, 2019). Replaces the variance dynamics with a Volterra equation driven by a fractional kernel,

$$V_t = V_0 + \int_0^t K_H(t-s)\,\lambda_{\mathrm{RH}}(\theta - V_s)\,ds + \int_0^t K_H(t-s)\,\nu\sqrt{V_s}\,dW_s^V, \tag{3}$$

where $K_H(\tau) = \tau^{H-1/2}/\Gamma(H+1/2)$ and $H \in (0, 1/2)$ is the Hurst exponent. Parameters $(v_0, \lambda_{\mathrm{RH}}, \theta, \nu, \rho, H)$ ($K = 6$).

SVCJ (Duffie et al, 2000). Heston plus correlated jumps in returns and variance, driven by a common Poisson process N_t of intensity λ:

$$dX_t = (-\tfrac{1}{2}V_t - \lambda\bar{m})\,dt + \sqrt{V_t}\,dW_t^X + J^X\,dN_t, \qquad dV_t = \kappa(\theta - V_t)\,dt + \sigma\sqrt{V_t}\,dW_t^V + J^V\,dN_t, \tag{4}$$

with variance jumps $J^V \sim \mathrm{Exp}(\mu_V)$, conditional return jumps $J^X \mid J^V \sim \mathcal{N}(\mu_s + \rho_J J^V, \sigma_s^2)$, and $\bar{m} = \mathbb{E}[e^{J^X} - 1]$ the risk-neutral jump compensator ($K = 10$).

Model	K	Rough vol	Return jumps	Variance jumps
Heston	5	–	–	–
Bates	8	–	✓	–
Rough Heston	6	✓	–	–
SVCJ	10	–	✓	✓

Table 1 Model taxonomy. K is the number of free parameters; rough vol indicates a fractional kernel with Hurst exponent $H < 1/2$; jump columns mark Poisson jumps in the respective component

3 Data, Calibration, and Pricing

Data and forward construction. We use $N = 1{,}989$ Deribit option-chain snapshots per asset (BTC and ETH), recorded at a 15-minute cadence in UTC between 10 and 31 March 2026, retrieved via Deribit's public REST API.[1] Deribit options are European and quoted in the inverse premium convention (premia denominated in the underlying); the calibration target is Deribit's mid-mark implied volatility for each (strike, maturity) cell. Deribit does not publish a risk-free rate; we therefore extract the at-maturity forward F_T per snapshot from put-call parity, $F_T = K/(1 - C^{\mathrm{inv}} + P^{\mathrm{inv}})$, taking the cross-strike median per maturity (Alexander et al, 2021). Snapshots are filtered to $T \in [7, 180]$ days, $|K/F - 1| < 0.30$, mid IV $\in [0.01, 5.00]$, open interest ≥ 1 (BTC) or ≥ 5 (ETH), and a minimum of 15 options across ≥ 3 distinct maturities. After filtering, BTC snapshots carry on average 265 options across 5–6 maturities; ETH carries 178 options across the same maturity grid (Table 2). We return to this chain-breadth gap when interpreting the asset-level differences in §4 and §6.

Asset	Snapshots	Options (mean)	Options (median)	Maturities (mean)	Maturities (range)
BTC	1,989	265	270	5.2	5–6
ETH	1,989	178	173	5.2	5–6

Table 2 Filtered Deribit option-chain sample. Filters: $T \in [7, 180]$ days, $|K/F - 1| < 0.30$, mid IV $\in [0.01, 5.00]$, OI ≥ 1 (BTC) / ≥ 5 (ETH); ≥ 15 options and ≥ 3 distinct maturities per snapshot

[1] Per-snapshot calibration outputs, generation scripts, and processed CSVs are available from the author on request.

Calibration. Each model is calibrated independently per snapshot by minimising the root-mean-square error in implied-volatility space,

$$\mathrm{RMSE}(\boldsymbol{\theta}) = \sqrt{\tfrac{1}{n}\sum_{i=1}^{n}\left(\sigma_i^{\mathrm{model}}(\boldsymbol{\theta}) - \sigma_i^{\mathrm{mkt}}\right)^2}, \tag{5}$$

where $\sigma_i^{\mathrm{model}}$ is obtained by Black–Scholes inversion of model prices in the inverse premium convention, all options weighted uniformly. Heston and Bates use Levenberg–Marquardt (50 iterations) followed by Nelder–Mead (200 iterations) with three random restarts. Rough Heston and SVCJ add a differential-evolution warm-start (Storn and Price, 1997); each is nested against its predecessor (rough Heston against Heston; SVCJ against Bates), with up to twelve seeded retries if the richer model under-fits its nested counterpart on a snapshot. Parameter bounds are widened for the high-volatility crypto regime: $\kappa \le 200$, $\sigma \le 40$, $\lambda \le 250$, $\mu_V \le 1.5$, $\rho_J \in [-1, 1]$ for SVCJ, and $H \in [0.001, 0.49]$ for rough Heston (enforcing $H < 1/2$).

Pricing. Vanilla prices and implied volatilities are computed via the Fourier–cosine expansion of Fang and Oosterlee, 2009 with $N_{\cos} = 128$ terms, using the analytic characteristic function for Heston, Bates, and SVCJ, and a Padé-accelerated fractional Adams–Bashforth scheme with $n_{\mathrm{quad}} = 250$ nodes to integrate the characteristic-function ODE of El Euch and Rosenbaum, 2019 for rough Heston.

Calibration cost. Calibrations ran on a Scaleway POP2-32C-128G cloud instance (32 AMD EPYC vCPUs, 128 GB RAM), with snapshots calibrated independently and distributed across cores and the C++ pricing core invoked from Python. Mean wall-time per snapshot is 0.3 s (Heston), 0.7 s (Bates), 21 s (rough Heston), and 37 s (SVCJ); the rough-Heston ODE solve and SVCJ's wider parameter space dominate compute. The full 15,912-fit grid (4 models × 2 assets × 1,989 snapshots) costs roughly 66 CPU-hours of aggregate single-core time. Outlier fits with mean-IV RMSE above 0.05 account for 0.96% of BTC Bates snapshots and 0% everywhere else; the multi-restart and nested-fit protocol delivers near-uniform calibration quality across both assets.

4 In-Sample Fit

Table 3 reports mean and median per-snapshot RMSE. Rough Heston and the jump-augmented models improve on Heston in mean RMSE on both assets, answering the first two questions affirmatively. Yet the asset-level winner differs sharply. On BTC the ordering (worst to best RMSE) is Heston > rough Heston > Bates > SVCJ, with SVCJ improving on Heston by roughly 40% in mean RMSE (0.0053 vs 0.0089). On ETH the ordering is Heston > Bates > SVCJ > rough Heston, with rough Heston the clear best fit (0.0060 vs 0.0090). The Bates and SVCJ mean RMSEs on ETH are within 0.0001, effectively tied; median RMSE preserves all rankings reported above. Two readings

Model	RMSE (mean)	RMSE (median)
BTC		
Heston	0.0089	0.0081
Bates	0.0070	**0.0045**
Rough Heston	0.0080	0.0071
SVCJ	**0.0053**	0.0048
ETH		
Heston	0.0090	0.0089
Bates	0.0076	0.0070
Rough Heston	**0.0060**	**0.0059**
SVCJ	0.0075	0.0074

Table 3 In-sample fit, $N = 1{,}989$ snapshots per asset. RMSE is on mid implied volatility, averaged across snapshots. Best per column per asset in bold

are consistent with the asymmetry. The first is specification: BTC's surface carries persistent short-dated skew that the jump kernel captures more parsimoniously than a fractional diffusion, while ETH's short-dated dynamics are closer to rough without identifiable discrete jump events. The second is sample size: BTC chains are roughly 50% wider than ETH chains (Table 2), so a $K = 10$ model can be supported on BTC where the same model is starved of identifying information on ETH (see §6).

5 Out-of-Sample Performance

We evaluate two out-of-sample protocols. The *time-series* (TS) protocol splits the snapshot stream 67/33 chronologically, calibrates each training snapshot independently, takes the per-parameter median across the training calibrations, and scores the test half with that median parameter vector — a stress test of parameter stability rather than a recommended forecasting recipe (practitioners typically recalibrate at every snapshot). The *cross-maturity* (XM) protocol holds out 25% of each snapshot's maturities and refits on the remainder, isolating cross-maturity extrapolation. The XM panel is drawn from a separate live-snapshot subset of the same Deribit data (97 BTC and 54 ETH snapshots), capped to 30 and 23 respectively for compute: per-snapshot XM refits with rough Heston and SVCJ cost 21–37 s each, making a full XM run on the 1,989-snapshot panel infeasible within the compute budget. Table 4 and Figure 1 report the resulting test RMSE.

Model	TS test RMSE	XM train RMSE	XM test RMSE
BTC (TS test n = 657, XM n = 30)			
Heston	0.0224	0.0130	0.0156
Bates	**0.0141**	0.0073	0.0163
Rough Heston	0.0217	0.0075	**0.0101**
SVCJ	0.0146	0.0076	0.0122
ETH (TS test n = 657, XM n = 23)			
Heston	0.0226	0.0094	0.0267
Bates	0.0367	0.0068	0.0215
Rough Heston	0.0252	0.0075	0.0169
SVCJ	**0.0174**	0.0080	**0.0162**

Table 4 Out-of-sample fit on mid IV. TS = chronological 67/33 split, test scored with the median train parameter vector. XM = within-snapshot maturity split, 25% of maturities held out, capped at 30 snapshots per asset. Best test column per asset in bold

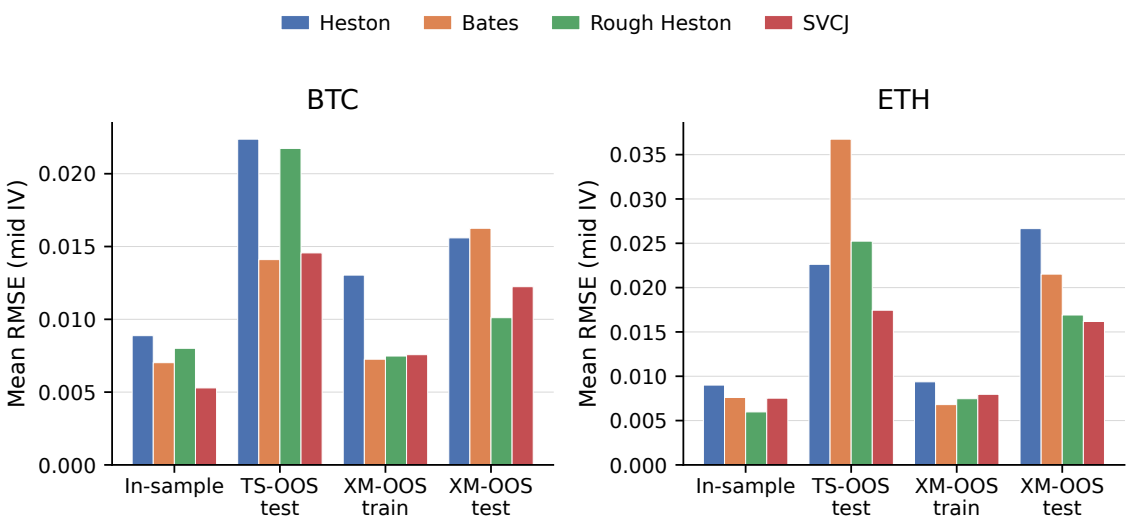


Figure 1 Protocol-dependent ranking on BTC and ETH. Mean per-snapshot RMSE on mid implied volatility, by model and evaluation protocol. The leader changes across protocols and across assets

The ranking is not invariant. On BTC, Bates wins the TS test (0.0141) but rough Heston wins the XM test (0.0101); SVCJ is second-best on both. On ETH, SVCJ is the clear leader under both protocols (0.0174 TS, 0.0162 XM); SVCJ is therefore the consistent OOS winner on ETH but not on BTC. The XM train RMSE is uniformly small (0.007–0.013) because each snapshot is re-fitted, isolating extrapolation across the maturity dimension; the gap between XM test and TS test RMSE therefore measures the cost of parameter *staleness* over time. The wide TS–XM-test gap on ETH Bates (0.0367 vs 0.0215) is a particularly clear example: Bates fits each snapshot's surface well, but its parameters drift fast enough within the 21-day window that a median over the train half

generalises poorly.

6 Identification of λ and H

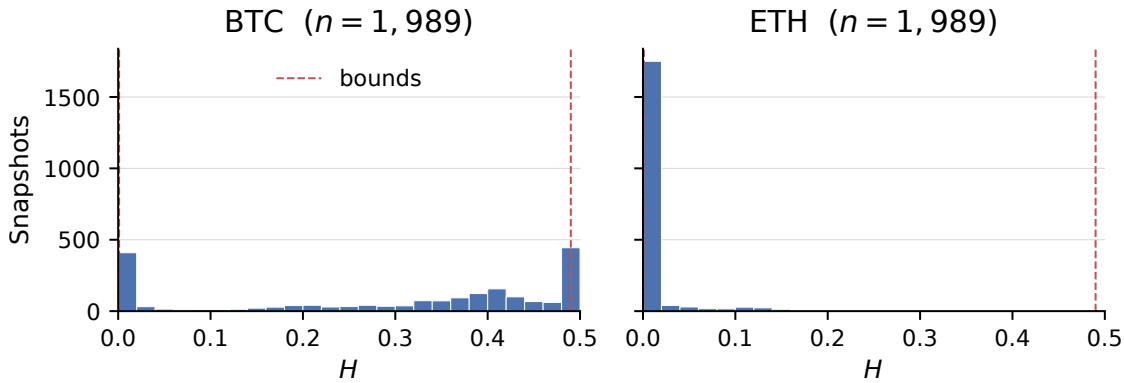


Figure 2 Distribution of the rough-Heston Hurst exponent H across the $N = 1{,}989$ snapshots per asset. Dashed lines mark the calibration bounds $H \in [0.001, 0.49]$

Asset	H low %	H int %	H up %	RH>H %	$\bar{\lambda}$ Bates	$\bar{\lambda}$ SVCJ
BTC	19.9	58.7	21.5	0.2	0.62	2.44
ETH	86.5	13.5	0.0	0.0	2.12	3.89

Table 5 Identification diagnostics. H pinning shares classify calibrated Hurst exponents against the bounds [0.001, 0.49] with a 0.005 tolerance. RH>H % is the share of snapshots where rough Heston's RMSE exceeds Heston's despite the nested-fit guarantee.$\bar{\lambda}$ is the mean fitted jump intensity (per year) under Bates and SVCJ respectively

The rough-Heston Hurst exponent behaves very differently across the two assets (Figure 2). On ETH, 86.5% of snapshots pin the lower bound and none pin the upper bound: the calibration repeatedly asks for a kernel *rougher* than the imposed limit of $H = 0.001$. On BTC the distribution is bimodal, with 19.9% at the lower bound, 21.5% at the upper, and the remainder spread across the interior. The two surfaces do not share a common roughness regime, and H should not be treated as a structural constant across crypto assets. An alternative reading is that the ETH pinning reflects misspecification rather than deep roughness: the same short-maturity excess curvature can be captured by the fractional kernel or by jumps, and the lower bound on H is doing identification work that a wider bound or a different parameterisation might relieve.

Jump intensity reveals matching asymmetry. Mean fitted λ under Bates is $0.62\,\mathrm{yr}^{-1}$ on BTC and $2.12\,\mathrm{yr}^{-1}$ on ETH; under SVCJ it rises to 2.44 and $3.89\,\mathrm{yr}^{-1}$ respectively. ETH's higher fitted intensity coexists with rough Heston's decisive in-sample win on the same asset: roughness and jumps are not substitutes that trade off cleanly, but *competing* representations of the same short-maturity excess curvature, with the fractional kernel parameterising it more efficiently on ETH and the jump kernel doing so on BTC. Rough Heston exceeds Heston's RMSE on only 0.2% (BTC) and 0% (ETH) of snapshots, so the nested-fit guarantee against Heston is rarely binding.

7 Conclusion

Rough Heston and the jump-augmented models improve on Heston in-sample for both assets, but SVCJ is not a universal out-of-sample winner: Bates leads the BTC time-series test, rough Heston leads the BTC cross-maturity test, and SVCJ leads ETH under both protocols. Rough volatility and jumps appear complementary across assets, but within each asset they compete as reduced-form representations of short-maturity curvature; their relative explanatory power depends on the asset and the evaluation protocol. This protocol dependence is a model-risk statement as much as an econometric one: a desk selecting SVCJ on its in-sample fit to BTC would hold the weaker time-series performer out of sample, so the model choice—and any margin, reserve, or hedge calibrated through it—should be validated under the protocol that matches its production use, with recalibration cadence treated as a first-order decision rather than an implementation detail. The BTC–ETH asymmetry mirrors structural differences

in the option chains themselves — BTC carries ∼ 50% more options across the same maturity grid — which is consistent with stronger SVCJ identification on BTC and with ETH's rough-Heston Hurst exponent being pushed against its lower bound. The evidence is conditioned on a single venue (Deribit), a 21-day window, uniform option weighting, and mid-mark implied volatility as the calibration target; longer panels, vega- or liquidity-weighted objectives, and cross-venue replication are natural extensions. Within these conditions, reporting one ranking under one protocol misrepresents the evidence.

References

Alexander C, Chen D, Imeraj A (2021) Inverse and quanto inverse options in a Black-Scholes world. arXiv preprint https://doi.org/10.48550/arXiv.2107.12041

Alexander C, Deng J, Feng J, et al (2023) Net buying pressure and the information in bitcoin option trades. Journal of Financial Markets 63:100764. https://doi.org/10.1016/j.finmar.2022.100764

Bates DS (1996) Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. The Review of Financial Studies 9(1):69–107. https://doi.org/10.1093/rfs/9.1.69

Duffie D, Pan J, Singleton KJ (2000) Transform analysis and asset pricing for affine jump-diffusions. Econometrica 68(6):1343–1376. https://doi.org/10.1111/1468-0262.00164

El Euch O, Rosenbaum M (2019) The characteristic function of rough Heston models. Mathematical Finance 29(1):3–38. https://doi.org/10.1111/mafi.12173

Fang F, Oosterlee CW (2009) A novel pricing method for European options based on Fourier-Cosine series expansions. SIAM Journal on Scientific Computing 31(2):826–848. https://doi.org/10.1137/080718061

Gatheral J, Jaisson T, Rosenbaum M (2018) Volatility is rough. Quantitative Finance 18(6):933–949. https://doi.org/10.1080/14697688.2017.1393551

Heston SL (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. The Review of Financial Studies 6(2):327–343. https://doi.org/10.1093/rfs/6.2.327

Hou AJ, Wang W, Chen CYH, et al (2020) Pricing cryptocurrency options. Journal of Financial Econometrics 18(2):250–279. https://doi.org/10.1093/jjfinec/nbaa006

Storn R, Price K (1997) Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization 11(4):341–359. https://doi.org/10.1023/A:1008202821328

Regional Apartment Price Disparities in the Czech Republic: Clustering, Labour Market Analysis with a Focus on the Moravian-Silesian Region

Alena Pozdílková[1], Martin Pozdílek[2]

Abstract. This paper examines spatial heterogeneity in apartment prices across Czech districts using a longitudinal dataset of daily advertised listing prices covering 77 districts from August 2018 to March 2026. The central research question is whether classical bid-rent theory adequately explains regional apartment price disparities in the Czech Republic, and where it breaks down in structurally depressed regions. Two analytical dimensions are investigated. First, K-Means clustering based on three features — long-run mean price per m², total price growth over the observation period, and average monthly listing volume — identifies five structurally distinct market segments. The resulting clusters capture not only static price levels but also market dynamics, revealing that historically depressed districts (Ostrava, Most, Karviná) form a distinct group characterized by low absolute prices but the fastest price growth (+261% on average), consistent with a process of price convergence. Second, the analysis is applied to the Moravian-Silesian Region as an exploratory case study, integrating commuting distances to Ostrava, labor market indicators (unemployment rate, median wage), and housing affordability measured via price-to-income ratios. Findings demonstrate that proximity to the regional employment center explains apartment prices only partially — structural factors, including post-industrial stigmatization, unemployment persistence, and amenity premiums, emerging as dominant determinants. The results challenge the classical monocentric bid-rent framework and highlight the complexity of housing market dynamics in post-industrial regions.

Keywords: apartment prices, K-Means clustering, housing affordability, Moravian-Silesian Region, Czech Republic, post-industrial regions

JEL Classification: D81, C69

1 Introduction

Regional housing markets in the Czech Republic have undergone a profound transformation since the post-communist transition of the 1990s, evolving from a centrally planned system characterized by uniform pricing toward a market-driven environment defined by growing spatial disparities. Over the past decade, and particularly since 2018, apartment prices have risen sharply across the country. Yet, the pace and magnitude of this growth have varied considerably across districts, reflecting deep-seated differences in economic structure, labor market conditions, and geographic accessibility.

Understanding the determinants of regional apartment price heterogeneity is of considerable importance for both academic research and housing policy. From a theoretical perspective, the spatial distribution of housing prices is shaped by a complex interplay of forces, including bid-rent gradients - Alonso (1964), spatial equilibrium mechanisms linking wages, rents, and amenities – Roback (1982), and the responsiveness of local housing supply to demand signals. Empirically, however, the extent to which these frameworks apply to post-transition economies – and particularly to structurally depressed regions – remains an open question.

This paper contributes to the growing literature on Czech and Central European housing markets by providing a comprehensive district-level analysis of apartment prices over the period 2018–2026. Using K-Means clustering and labour market indicators, we examine spatial patterns across all 77 Czech districts and offer a detailed exploratory case study of the Moravian-Silesian Region – a historically industrial area undergoing significant economic restructuring. The region provides a compelling setting in which the classical relationships between

[1] University of Pardubice, Department of Automatization and Mathematics, Studentská 95, Pardubice, alena.pozdilkova@upce.cz.
[2] University of Pardubice, Department of Information Technologies, Studentská 95, Pardubice, martin.pozdilek@upce.cz.

proximity to employment centers, wages, unemployment, and housing prices are tested against the reality of post-industrial transition. Specifically, the paper addresses the following research question: to what extent do Czech apartment markets exhibit spatially differentiated price clusters, and how do labour market conditions and commuting accessibility shape housing affordability in post-industrial regions such as the Moravian-Silesian Region?

The complex interdependence of housing and labour markets has been systematically reviewed by Haas and Osland (2014), who highlight that commuting, migration, and residential choice are jointly determined rather than sequential decisions. Complementing this, Eliasson and Westerlund (2024) provide empirical evidence from Sweden that high relative housing prices in productive urban regions reduce geographical labour mobility, particularly among young and highly skilled workers — a mechanism with direct relevance to the Czech context, where Prague's detachment from national price levels may constrain workforce reallocation.

In the context of Central and Eastern European housing markets, Kostelecký (2000) documented the emergence of dramatic regional price inequalities following post-communist transition, with the Ostrava and Ústí nad Labem regions forming the lowest price segment due to industrial decline and skilled labour emigration. More recently, Hlavacek and Kalabiska (2020) identified income levels, unemployment, mortgage availability, and spatial autocorrelation as the principal determinants of district-level apartment prices in the Czech Republic, confirming that price shocks in one district propagate to neighbouring areas. Hromada, Čermáková, and Piecha (2022) further demonstrated that structurally depressed cities such as Ostrava and Most experienced disproportionate price growth after 2018, attributable to speculative investment and the transfer of investor demand from Prague to cheaper regional markets. At the comparative level, Slavata, Ardielli, and Maciejasz (2024) showed that housing affordability in the Moravian-Silesian Region lags behind Polish Silesia despite broadly comparable wage levels, while Matějková and Tichá (2025) confirmed that affordability conditions in Czech cities rank among the most strained in Central Europe, with Prague exhibiting price-to-income ratios exceeding those of Vienna, Warsawa, and Bratislava.

2 Methodology

2.1 Data Description

The data reserved for our research comes from sreality.cz. These are some web pages containing ads. Each advert was recorded separately using a Python script. At 2:30 am every day, the script tries to download all ads from each district from the website sreality.cz. The MongoDB database is used to store ads. New ads are saved directly to the database. Daily and monthly summaries for all used filters are created in the database. The filtered data can then be displayed or exported to a file after access via the web interface. Pozdílková et al. (2019)

The dataset consists of daily observations of apartment listing statistics aggregated at the district level for the Czech Republic. The raw file contains 2,709 daily rows spanning from 1 August 2018 to 31 March 2026, covering 87 geographic entries (10 Praha districts, Praha-město, Praha-východ, Praha-západ, and 74 other districts).

For each district and each day, three variables were recorded:

- **Mean listing price** (avg): average advertised price per m² in CZK
- **Standard deviation** (stddev): price dispersion across active listings
- **Number of listings** (totnum): total count of active apartment listings

The original dataset contains **11** separate entries for Prague (Praha 1–10 and Praha-město). For the district-level analysis, all Prague entries were merged into a single "Praha" record by computing the *simple arithmetic mean* of mean prices across all 11 entries for each time point and then averaging over the full period. This yields a single representative price series for the capital (a merged mean of approximately **89,868 CZK/m²**). Praha-východ and Praha-západ are retained as separate districts. After consolidation, the dataset covers **77 districts**.

2.2 Clustering Results (k=5, Three Features)

To capture both the structural price level and the market dynamics of each district, clustering is performed on three standardized features. The first is the **long-run mean price per m²**, computed as the arithmetic mean of daily average prices over 2018–2026. The second is **total price growth (%)**, defined as the percentage change between the mean price in the first and final 90 days of the observation period. This dynamic feature distinguishes markets that are expensive but stagnating from those that are cheap but rapidly converging. The

third is **average monthly listing volume**, the mean number of active listings per month, which captures market liquidity.

All three features were standardized using **z-score normalization (StandardScaler)**, with each feature transformed to zero mean and unit variance. This ensures that differences in scale between features do not distort the Euclidean distance calculations in K-Means.

Two complementary methods were used to select the optimal number of clusters k:

Elbow Method (Inertia / WCSS)

The Within-Cluster Sum of Squares (WCSS) was computed for $k \in \{2, \ldots, 11\}$. This metric measures the total squared distance between each data point and its assigned cluster centroid. As k increases, WCSS decreases monotonically; the optimal k is identified at the **"elbow" point** where the marginal gain in compactness diminishes sharply.

Silhouette Score

The Silhouette Score measures how similar each point is to its own cluster relative to its nearest neighbouring cluster, ranging from −1 to +1. A higher silhouette score indicates better-defined, well-separated clusters. The score was computed for the same range of k values as a complementary criterion.

K-Means Algorithm

Clustering was performed using the K-Means algorithm with the following configuration:

- **Initialization:** k-means++ (via 20 independent restarts to avoid local optima)
- **Distance metric:** Euclidean distance on standardized features
- **Convergence:** Standard tolerance with max 300 iterations
- **Random seed:** 42 (for reproducibility)
- **Implementation:** scikit-learn KMeans (Python)

3 Analysis and Results

3.1 Clustering by Mean Apartment Price

The elbow and silhouette analysis were performed for $k \in \{2, \ldots, 11\}$ on the three-feature standardized matrix. The elbow curve shows a notable inflection at k = 5, and k = 3 yields the highest silhouette score; k = 5 was selected as the final solution, balancing compactness with interpretability.

Cluster	Label	Count	Min (CZK/m²)	Max (CZK/m²)
0	High-price, slow growth	4	134,446	133,908
1	Mid-price, moderate growth	18	63,414	91,151
2	Mid-price, slow growth	22	44,904	59,565
3	Mid-price, fast growth	26	41,539	64,997
4	Low-price, fastest growth	7	23,922	67,793

Table 1 K-Means Clustering Results (k=5, features: mean price, growth, listing volume). Source: own processing

The three-feature clustering yields five economically interpretable segments. Cluster 0 groups the four highest-priced districts (Praha, Brno, Olomouc, Karlovy Vary) characterized by high absolute prices but the slowest relative growth (+110%), consistent with market maturity. Cluster 4 captures seven post-industrial districts (Ostrava, Most, Karviná, Chomutov, Teplice, Ústí nad Labem, Jablonec nad Nisou) with the lowest prices but by far the fastest growth (+261%), reflecting price convergence from a depressed baseline. Cluster 1 captures the *"second league"* comprising Brno, Praha-západ, Praha-východ, and two mountain recreation districts (Trutnov, Semily), where tourism premiums push up prices. Cluster 4 (periphery) is geographically concentrated in the structurally depressed north-west Bohemia (Most, Sokolov, Teplice, Chomutov) and north-east Moravia (Karvinná, Ostrava area).

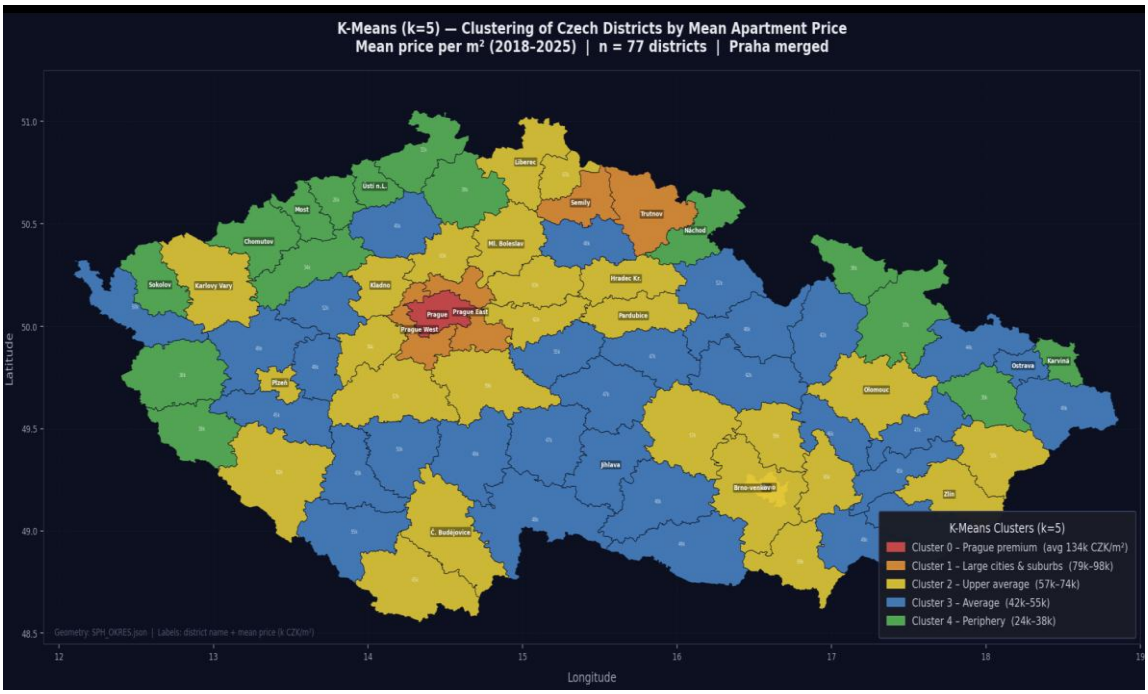


Figure 1 K-Means Clustering Map (k=5, three features). Source: own processing

3.2 Analysis of Moravian-Silesian Region

Table 3 presents a comparative overview of all six Moravian-Silesian districts across key dimensions: mean price level, price growth over the observation period, and average monthly listing volume. This section presents exploratory evidence and is intended to complement the national-level clustering analysis rather than to establish causal relationships.

District	Price CZK/m²	Growth 2018→2026	Avg. listings/month	National cluster	PTI (years)
Frýdek-Místek	48 687	+165 %	110	C3 – Fast growth	5.5
Ostrava-město	46 681	**+267 %**	**501**	C4 – Fastest growth	5.2
Opava	44 429	+177 %	103	C3 – Fast growth	5.3
Nový Jičín	37 985	+201 %	78	C3 – Fast growth	4.3
Bruntál	37 262	+173 %	84	C3 – Fast growth	5.0
Karviná	26 735	**+236 %**	219	C4 – Fastest growth	3.3

Table 2 Results for Moravian-Silesian Region. Source: Own processing

Key findings:

- **Prices and growth:** Frýdek-Místek is the most expensive district in the region (48,687 CZK/m²), but the largest relative growth was recorded in Ostrava (+267%) and Karviná (+236%) – the two historically cheapest locations are now catching up with the rest of the fastest.
- **Market liquidity:** Ostrava dominates in terms of ad volume (500/month), with Karviná second (219) – despite low prices, it is a very active market.
- **Price dispersion:** The volatility graph shows significant peaks in Frýdek-Místek and Bruntál – a heterogeneous offer (large dispersion in the quality of apartments).

The heatmap in Figure 4 provides a multi-metric overview of all six MSK districts. Colour intensity encodes relative performance within each metric on a normalized scale (red = high / unfavourable, green = low / favourable). Karviá stands out as the most internally contradictory district: it offers the lowest prices and best housing affordability (PTI = 3.3 years), yet records the highest unemployment and, despite being only 29 minutes from Ostrava, remains the most structurally distressed market. Frýek-Místek presents the opposite profile — the highest prices and worst affordability (PTI = 5.5 years), but the lowest unemployment in the region.

Bruntál is uniquely disadvantaged across two structural dimensions simultaneously: high unemployment and the greatest distance from the regional employment center.

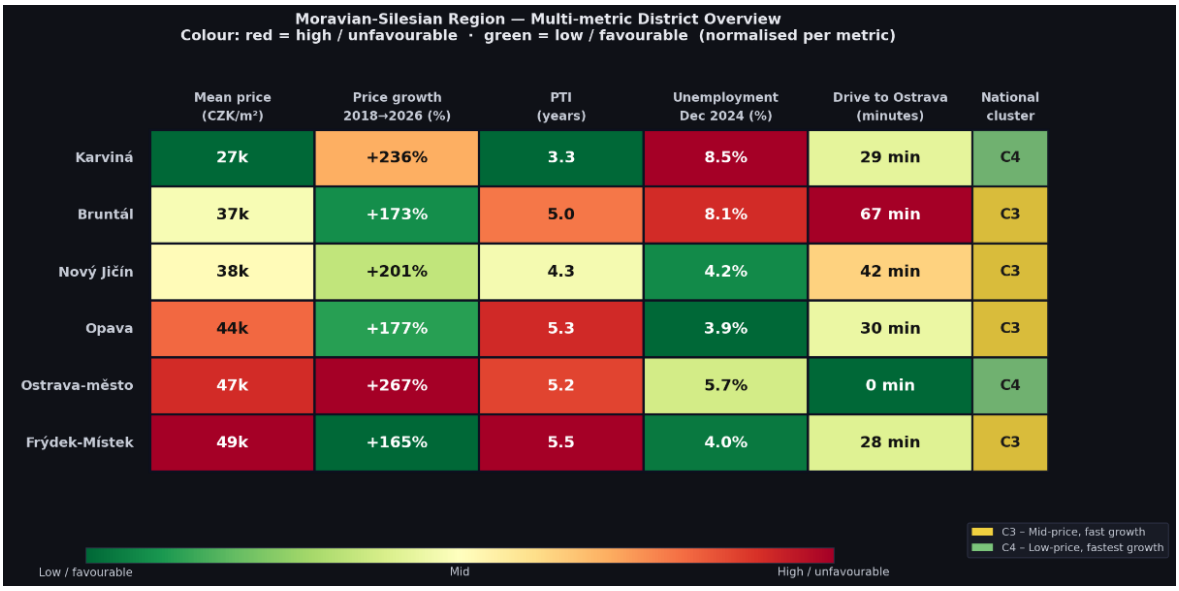


Figure 2 Multi-metric heatmap of Moravian-Silesian Region districts. Source: own processing

National cluster context and housing affordability. Contextualized within the national clustering results, four MSK districts — Frýdek-Místek, Opava, Nový Jičín, and Bruntál — belong to Cluster 3 (mid-price, fast growth). In contrast, Ostrava-město and Karviná belong to Cluster 4 (low-price, fastest growth). This classification underscores that the entire MSK region is characterized by strong convergence dynamics, yet from very different starting points.

Housing affordability, measured as the price-to-income ratio (PTI) for a 50 m² apartment relative to local median gross wages (ISPV, 2023), ranges from 3.3 years in Karviná to 5.5 years in Frýdek-Místek. The relatively low PTI in Karviná reflects its depressed price level rather than favourable wage conditions. In contrast, the high PTI in Frýdek-Místek reflects price inflation driven by amenity premiums (the Beskydy Mountains, the automotive industry) relative to wages, comparable to those in Ostrava. The PTI deteriorated across all districts between 2020 and 2026, as price growth substantially outpaced wage growth, except in Karviná, where prices grew rapidly but from a very low base.

The Moravian-Silesian Region was further analyzed based on accessibility to Ostrava, the center and place of employment for a large part of the population. Geographic calculations were based on the exact coordinates of all cities from the GeoJSON, and real distances were calculated, to which estimated travel times were assigned based on the nature of the region's road network.

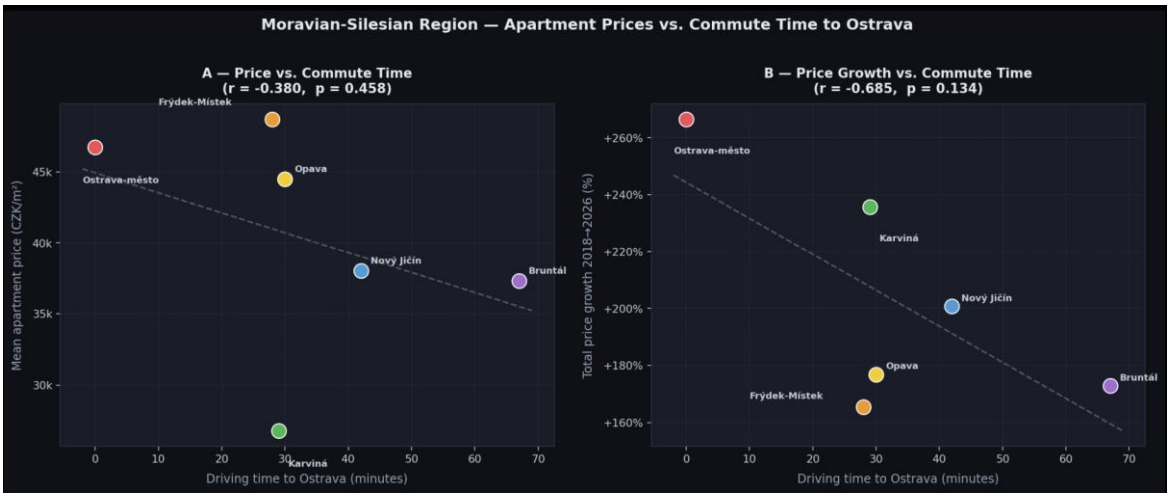


Figure 3 Results for Moravian-Silesian Region. Source own processing

Unemployment was added as an additional parameter to the model (from the Czech Statistical Office/Ministry of Labor and Social Affairs; it is accurate at the district level). Unfortunately, wages are only available at the regional level – ISPV data for the MSK region were used, and relative estimates were added based on the economic structure of individual districts.

4 Discussions

As an exploratory case study, the Moravian-Silesian analysis reveals that unemployment and apartment prices are negatively correlated across districts – districts with high unemployment (Karviná, Bruntál) have low prices. This is economically logical, but surprising in the context of commuting: Karviná is only 29 minutes from Ostrava, yet prices remain the lowest. These findings suggest that unemployment rates and the perceived image of a locality suppress prices more than geographical proximity to the employment center, directly challenging the predictions of the classical bid-rent model in this post-industrial context.

Housing affordability (PTI): Paradoxically, Karviná is the most affordable – buying a 50m² apartment costs 3.3 annual gross wages. Frýdek-Místek is the worst off (5.5 years), where high prices (driven by the Beskydy Mountains and the automotive industry) meet wages comparable to Ostrava.

Availability graph over time: In practically all districts, housing affordability has deteriorated significantly since 2020 – prices grew significantly faster than wages. The only exception is Karviná, where affordability has remained relatively stable.

5 Conclusion

The national clustering analysis, based on three features capturing price level, price dynamics, and market liquidity, confirms that Czech apartment markets are highly spatially segmented. A key analytical contribution of the multi-feature approach is the identification of a distinct convergence cluster: historically depressed districts in north-west Bohemia and the Moravian-Silesian Region record the lowest absolute prices yet the fastest growth (+261%), while high-price districts such as Praha and Brno exhibit market maturity with the slowest appreciation (+110%). The classical bid-rent model (Alonso, 1964) provides a useful theoretical baseline, but its explanatory power is limited in post-industrial contexts. In the Moravian-Silesian Region, Karviná most strikingly challenges the model: located only 29 minutes from Ostrava, it nonetheless records the lowest prices in the region. As Hromada et al. (2022) document, such anomalies reflect the so-called poverty trade and persistent local stigmatization – phenomena well documented in other post-industrial regions. The rapid price growth recorded in historically depressed areas (+236% Karviná, +267% Ostrava) corresponds to the pattern described by Agnew and Lyons (2017): the arrival of investment and stabilization of employment are quickly reflected in prices, particularly from a low baseline. Taken together, the findings suggest that any adequate explanation of housing price disparities in the Czech Republic must account not only for accessibility and income, but also for structural labour market conditions, historical industrial legacies, and the persistence of local reputation effects.

Acknowledgements

The paper was supported by institutional support of the University of Pardubice.

References

Alonso W (1964) Location and Land Use: Toward a General Theory of Land Rent. Harvard University Press

Agnew J, Lyons R (2018) The Impact of Employment on Housing Prices: Detailed Evidence from FDI in Ireland. Regional Science and Urban Economics 70: 174-189

Eliasson K, Westerlund O (2024) Housing Markets and Geographical Labour Mobility to High-Productivity Regions: The Case of Stockholm. European Urban and Regional Studies 2024 31(3): 259–280

Haas A, Osland L (2014) Editorial: Commuting, migration, housing and labour markets. Complex interactions. Urban studies 51(3): 463-476. DOI:10.1177/0042098013498285

Hlavacek M, Kalabiska R (2020) Regional Determinants of Housing Prices in the Czech Republic. Czech Journal of Economics and Finance 2022 72(1): 2-29. Charles University Prague

Hromada E, Čermáková K, Piecha M (2022) Determinants of House Prices and Housing Affordability Dynamics in the Czech Republic. December 2022. European Journal of Interdisciplinary Studies 14(2): 119-132, DOI:10.24818/ejis.2022.24

Kostelecký T (2000) Housing and Its Influence on the Development of Social Inequalities in the Post-Communist Czech Republic. Czech Sociological Review 36(2): 177-194

Matějková L, Tichá A (2025) Housing Market Trends and Affordability in Central Europe: Insights from the Czech Republic, Slovakia, Austria, and Poland. Buildings 15(10)

Oswald A J (1999) A Conjecture on the Explanation for High Unemployment in the Industrialized Nations. Part I. University of Warwick Economic Research Papers 475.

Pozdílková A, Nedvědová M, Marek J (2019) An Approach to Web Advertising Server parsing - Apartment Price Analysis in the Czech Republic. In: 15th International Scientific Conference on Informatics INFORMATICS 2019: proceedings, pp 333-337

Roback J (1982) Wages, Rents, and the Quality of Life. Journal of Political Economy 90(6): 1257–1278

Slavata D, Ardielli E, Maciejasz M (2024) Comparison of Housing Affordability in Czech and Polish Regions. Review of Economic Perspectives 24(1): 17–36

Analysis of Productivity in an Automated Production Line and Modeling Its Dynamics

Pavel Pražák[1], Josef Toman[2], Andrea Ševčíková[3]

Abstract. This paper presents a data-driven analysis of an automated production line aimed at quantifying productivity and modeling its stochastic dynamics. Shift-level data are preprocessed to ensure temporal consistency. Productivity per shift is analyzed using nonparametric methods, and differences across working days confirm a non-uniform weekly pattern. The typical production level of a particular Czech firm is estimated. The temporal dynamics are modeled as a discrete-time stochastic process with weekly seasonality. The time series is described by a seasonal ARIMA model applied directly to production levels, capturing the recurring dependence structure associated with the five-working-day cycle. The results indicate a stable production level and a weekly dependence pattern, where production on a given working day is related to production on the corresponding working day of the previous week. The estimated typical production per shift is further used to derive component consumption rates and replenishment frequencies, supporting production planning and logistics optimization.

Keywords: automated production line, productivity, SARIMA, inventory and supply planning

JEL Classification: C22, C14

1 Introduction

To analyze the production process on an automated production line and to prepare inputs for modeling, it is necessary to collect data that can describe the production process. This includes identifying, describing, and measuring material flows and the supply of components for the final product. The data should also include information about the layout of the production line and material flows so that their efficiency can be evaluated. In general, such data can be obtained from the company's information system. However, these data need to be further discussed in order to understand them properly, complete missing information, or reorganize them if needed. It is often assumed that an *automated production line operates in a stable and constant way over time*. However, *this assumption may not reflect reality*. The real production (per day, per shift, per hour) can be variable due to seasonal effects, human factors, or unexpected random events. Therefore, this variability should be carefully described and analyzed to support effective production planning and to develop suitable models of the production process.

Statistical analysis plays a fundamental role in modeling and optimizing automated production systems, where manufacturing processes are characterized by variability, uncertainty, and temporal dependence. In industrial practice, statistical process control (SPC) is widely used to monitor production performance and detect deviations from stable operating conditions. Classical approaches focus on the analysis of variability, distribution characteristics, and process stability using statistical tools such as control charts and hypothesis testing (Montgomery, 2019). These methods enable the identification of systematic deviations and support decision-making in real manufacturing environments.

Time-series analysis represents another key approach for analyzing production data obtained from automated lines. Models such as ARIMA and SARIMA, based on the Box–Jenkins methodology, are commonly applied to model production output, identify trends and seasonality, and support production planning (Box et al., 2015; Cowpertwait & Metcalfe, 2009). In manufacturing applications, these methods are typically applied to production indicators such as output per shift, machine utilization, or failure occurrence. In particular, production per shift represents a

[1] University of Hradec Kralove/Faculty of Informatics and Management, Department of Informatics and Quantitative Methods, Rokitanskeho 62, Hradec Kralove 3, 500 03, pavel.prazak@uhk.cz.
[2] University of Hradec Kralove/Faculty of Informatics and Management, Department of Informatics and Quantitative Methods, Rokitanskeho 62, Hradec Kralove 3, 500 03, josef.toman@uhk.cz.
[3] University of Hradec Kralove/Faculty of Informatics and Management, Department of Informatics and Quantitative Methods, Rokitanskeho 62, Hradec Kralove 3, 500 03, andrea.sevcikova@uhk.cz.

key performance indicator used for evaluating productivity and variability of automated production lines. In real industrial applications, time-series methods are often combined with statistical evaluation of variability and performance, (Lee et al., 2015) or (Kumar and Jain, 2010). For example, (Sen and Alkan, 2025) analyze failure data from an industrial assembly line using time-series modeling and statistical error metrics, demonstrating the importance of statistical validation in predictive modeling. Similarly, (Javier et al., 2024) apply statistical analysis of time-series data for monitoring production processes and detecting anomalies in manufacturing systems. In recent years, research has increasingly integrated statistical methods with modern data-driven approaches. While machine learning techniques improve prediction accuracy in complex and nonlinear environments, statistical methods remain essential for data preprocessing, hypothesis testing, and interpretation of results (Kusiak, 2018). The combination of statistical modeling, time-series analysis, and modern analytical approaches thus provides a robust framework for understanding and optimizing automated production lines within Industry 4.0 environments. Despite these advances, most studies focus on individual machines or subsystems rather than entire production systems. This study addresses this gap by analyzing real production data at the level of a complete automated production line.

2 Data and Methods

The actual production of an automated production line, taking into account its possible variability, can be described based on an analysis of production data, which is outlined in this section. Assume that production data for a given year are available, containing information on the number of units N_t produced on day t and the number of shifts s_t completed on that day. The average productivity per shift on day t is then defined as $p_t = N_t/s_t$. Some shifts may start on the last day of a non-working period (typically on Sunday or at the end of public holidays). In such cases, the corresponding production records should be assigned to the following working days when the shifts are completed. This allows the preparation of production records at the level of individual shifts. The data may also include exceptional or atypical events, such as extraordinary production on Saturdays, which should be removed during data cleaning. To determine the typical production per shift, two approaches can be applied. All analyses were conducted in R (version 4.5.2; R Core Team, 2025), using the *tseries, forecast, urca, moments, BSDA*, and *FSA* packages for statistical testing, descriptive analysis, and time-series modelling.

2.1 Data Analysis

The recorded values p_t, $t \in \{1, 2, \dots, n\}$, where n is the number of working days per year, can be treated as independent observations of the productivity per shift. A statistical description can then be performed. Given the nature of the data, nonparametric methods are more appropriate for hypothesis testing, particularly the sign test for the median or the Wilcoxon signed-rank test, (Groebner et al., 2018). Differences in production across working days can be assessed using the Kruskal–Wallis test for multiple independent samples.

2.2 Time Series Modeling

The calculated productivity values p_t per shift on days t, where $t \in \{1, 2, \dots, n\}$, can be regarded as a time series with a period of five working days. A suitable model can be sought for this time series, and its parameters estimated. Before the actual analysis, it is advisable to adjust the time series so that it satisfies the assumption of a regular time index. It is necessary to identify missing observations and subsequently impute suitable values. Missing values for individual days of the week will be replaced by the average value of the corresponding day of the week. This step is essential for applying ARIMA-type models, which assume uniformly distributed data over time. Unresolved gaps in the time series would lead to a distortion of the autocorrelation structure and incorrect model identification. It is therefore advisable to impute missing values in the data caused by production line downtime on public holidays or during company-wide vacations. As a time-series model, one can use SARIMA model with a seasonality of five days based on the Box-Jenkinson methodology, (Cowpertwait and Metcalfe, 2009), (Box et al., 2015) and (Franses et al., 2014). The general SARIMA$(p, d, q)(P, D, Q)_s$ model can be written as follows

$$\Phi_P(\boldsymbol{B}^s)\phi_p(\boldsymbol{B})(1-\boldsymbol{B})^d(1-\boldsymbol{B}^s)^D p_t = \Theta_Q(\boldsymbol{B}^s)\theta_q(\boldsymbol{B})\varepsilon_t \tag{1}$$

where **B** denotes the lag operator, p_t denotes the observed value of the time series at time t (production per shift), ε_t is a white noise error term, and $\phi_p, \theta_q, \Phi_\mathrm{P}, \Theta_\mathrm{Q}$ represent non-seasonal and seasonal autoregressive and moving average polynomials of orders p, P, q and Q, respectively. More specifically, p, d, q denote the orders of the non-seasonal autoregressive, differencing, and moving average components, while P, D, Q represent the corresponding

seasonal components, and s is the seasonal period (here we consider that $s = 5$). To select an appropriate model, we rely on AIC and BIC criteria and perform residual diagnostics. In particular, the Ljung–Box test is used to assess the absence of autocorrelation, while normality of residuals is also examined.

2.3 Data preparation for the consumption of subcomponents and for the operation of intermediate warehouses

The final product of an automated production line consists of components that must be delivered to the line. These components are consumed by the automated production line. Typically, components are not delivered to the line individually, but are prepared in batches, such as on pallets or in packages. Let us assume that the production of the final product requires m components $M_1, M_2, \dots, M_m$, which are delivered to the production line in batches of $q_1, q_2, \dots, q_m$ pieces. Suppose it has been demonstrated that the typical production output per shift is Q final products and that, as mentioned above, we know the number of deliveries of all subcomponents of this product to the intermediate warehouses associated with the production line. Then it is possible to determine the number of deliveries to these intermediate warehouses and the consumption rate of subcomponents - for the i-th subcomponent of the product, $i \in \{1, 2, \dots, m\}$, the number of deliveries is the ratio Q/q_i, where q_i is the number of units in a single delivery. Since the number of deliveries to the intermediate warehouses can only be an integer, we must consider the upper integer part of the given fraction, i.e., $n_i = \lceil Q/q_i \rceil$. Based on these values, one can plan the supply of subcomponents of the final product to the assembly line and also consider an optimal supply scenario.

3 Results and Discussion

This section presents a description and analysis of data from a company operating in the Czech Republic, which uses an automated production line with 12 inputs and 12 semi-finished products. The dataset covers the full year 2024. The data were further prepared as described in Section 2. The calculated productivity of the production line per shift can be used to determine the number of material deliveries to the production line and, consequently, the consumption of components. In addition, the uniformity of production across the workdays of the week will be assessed.

3.1 Productivity of the automated line

The analysis first considers the following working hypothesis derived from the firm's management statement

V: The typical production output of the production line per shift is 3 500 products.

The distribution and summary description of production per shift are shown in Figure 1 and Table 1. These results were obtained based on shifts held over 234 working days (Saturdays were not considered, and partial production on Sundays was reported as part of the Monday shift).

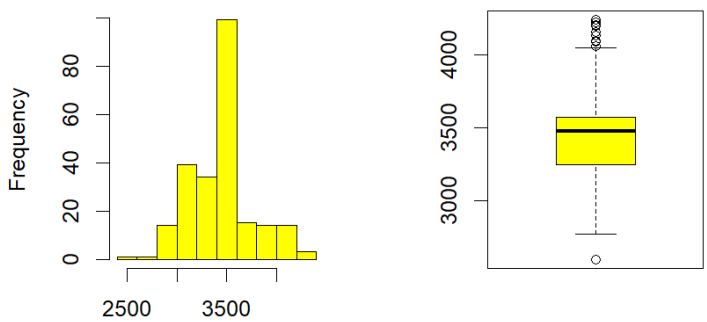


Figure 1 Distribution of production per shift

The distribution of the production line productivity per shift shows only a minor deviation from symmetry, as indicated by the low skewness value (SK = 0.24), suggesting slight positive skewness. The kurtosis value (K = 3.04) is very close to that of a normal distribution, indicating a slightly leptokurtic shape. The Shapiro-Wilk test indicated a statistically significant deviation from normality, W = 0.97, $p < 0.001$. Given the sensitivity of the test to sample size, this result should be interpreted with caution. In combination with descriptive characteristics, the distribution shows rather minor deviations from normality. Nevertheless, non-parametric methods were considered for subsequent analyses.

	n	Min	Q1	Median	Average	Std. Dev.	Q3	Max
Output per shift	234	2595	3246	3473	3450	312.28	3570	4242

Table 1 Descriptive statistics of production per shift

Based on the nonparametric one-sample Sign test, it cannot be confirmed that the median productivity of the production line differs from 3 500 units per shift ($s = 109,\ p = 0.39$; 95% CI [3424.98, 3510.00]). Therefore, this finding does not allow for the rejection of statement *V*, and the specified value of production line productivity per shift can be retained for further analysis.

Figure 2 and Table 2 show the distribution of production line productivity across the individual workdays of the week. The results in Table 2 indicate differences in production levels and variability across the week, with higher and more stable production typically observed at the beginning of the week and lower, more variable production towards the end. Using the Kruskal-Wallis test, it can be demonstrated that production line productivity differs across the individual workdays, $\chi^2 = 12.05, d.f. = 4,\ p = 0.02$. The results confirm that production is not uniform across the workdays of the week.

	n	Min	Q1	Median	Average	Std. Dev.	Q3	Max
Monday	44	3116	3399	3551	3566	229.07	3626	4224
Tuesday	48	2769	3279	3470	3448	283.60	3570	4053
Wednesday	48	2962	3297	3492	3452	246.98	3570	4158
Thursday	48	2912	3167	3456	3426	292.78	3570	4098
Friday	46	2595	2990	3234	3365	443.14	3565	4242

Table 2 Descriptive statistics of production per shift by day of the week

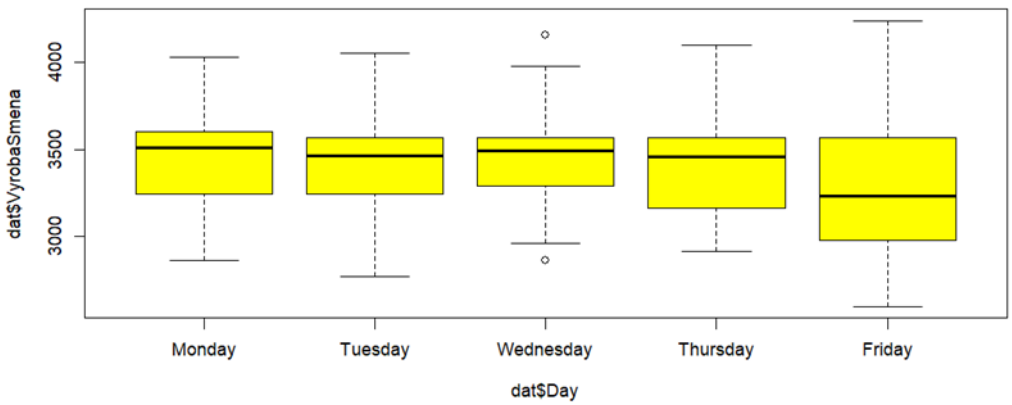


Figure 2 Box plot of production per shift by day of the week

Using Dunn's post-hoc test (Benjamini-Hochberg), we can demonstrate that production on Monday significantly differs from that on Friday, $g = -3.4, p = 0.007$. The difference in production between the other pairs of days is not significant. This corresponds to the weekly work rhythm, where Monday is above average and Friday is weaker in terms of production and also has higher variability (standard deviation), cf. Table 2.

3.2 Consumption of the automatic production line

Based on the distribution values obtained, we can determine the number of runs needed to replenish the production line so that smooth operation is possible. If typical production per shift is $Q = 3\,500$ units, as shown in the Section 3.1, we can use the procedure described in Section 2.3 to determine how many times per shift it is necessary to replenish the intermediate storage areas with components for the final product. In the company whose data we are using, the procedure described above yielded the values for the number of runs shown in Table 3. Given that

the number of replenishments is constrained to natural numbers, the integer part of the resulting fraction is considered, see also Section 2.3. The results shown in Table 3 indicate, for example, that at least 9 pallets deliveries of semi-finished product/component IP1 can be expected during a single shift. Similarly, it can be expected that at least 15 pallets of component IP2 will need to be delivered. We could continue in this manner for all 12 semi-finished products, but since these results will not be used further in this article, only a selection is shown in Table 3. Nevertheless, results of this nature are of significant importance for the effective planning of components deliveries to the production line. In the event that individual deliveries of semi-finished products can be accommodated within a single replenishment cycle, the formulation of an optimization model for this process can be considered. It means that the data from Table 3 can be further used as input for an optimization problem involving the replenishment of components and their consumption by an automated production line.

Intermediate Warehouse	**QTY,** q_i	**Unit**	**Median runs/shift**	**Number of pallets/shift**	**Intermediate product – IP**
M1	400	Pallet	8.8	**9**	IP 1
M2	240	Pallet	14.6	**15**	IP 2
M3	288	Pallet	12.2	**13**	IP 3
⋮	⋮	⋮	⋮	⋮	⋮
M12	8820	Pallet	0.4	**1**	IP 12

Table 3 Component consumption and number of replenishments per shift

3.3 Time Series Analysis of Productivity

The production per shift series was also modelled as a time series. The time series of productivity for individual working days is presented in Figure 3, illustrating the evolution of production levels and recurring five-working-day patterns. The Augmented Dickey–Fuller (ADF) test was used to examine the stationarity of the time series. The test was estimated both with an intercept and with an intercept plus a linear trend. In both specifications, the null hypothesis of a unit root was rejected: $\tau_2 = -11.259$ compared with the 5% critical value of -2.87, and $\tau_3 = -12.486$ compared with the 5% critical value of -3.42. These results indicate that the series is rather stationary. Since the deterministic trend was weak and the objective was to retain a parsimonious specification, the productivity series was finally modelled in levels using $\mathrm{ARIMA}(0,0,0)(1,0,1)_5$. The estimated model yielded $AIC = 3709$ and $BIC = 3723$. The specification captures the weekly dependence structure associated with the five working days, rather than short-term day-to-day dynamics. Although the residuals of the selected model deviate from normality, the Ljung–Box test did not indicate significant autocorrelation ($Q^* = 4{,}647, d.f. = 8, p = 0.79$), suggesting that they can be regarded as approximately white noise.

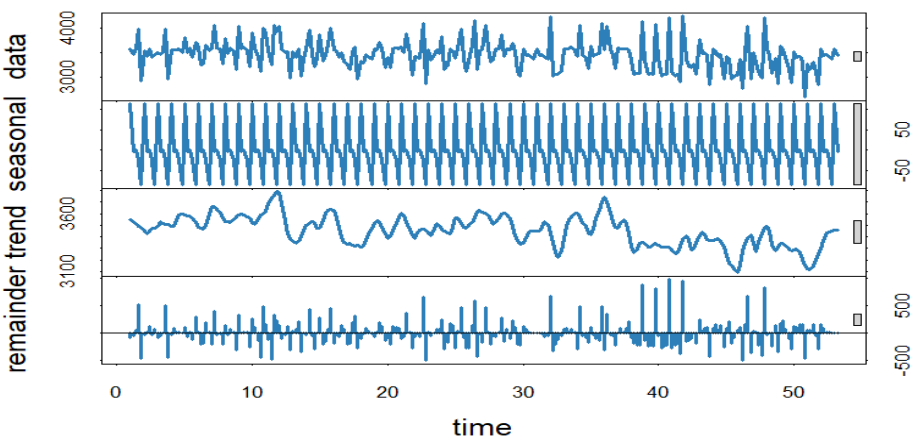


Figure 3 STL decomposition of production per shift with a five-observation seasonal period; one time-axis unit corresponds to five working days or one working week

The estimated model can be written as

$$p_t - 3450.3704 = 0.6885(p_{t-5} - 3450.3704) + \varepsilon_t - 0.4624\,\varepsilon_{t-5}, \quad (2)$$

where p_t represents production per shift at the day t and ε_t denotes a white noise error term. The estimated mean production ($\mu = 3450.3704;\ s.e. = 29.3767$) units per shift, represents the long-run average level around which the productivity series fluctuates. The seasonal autoregressive coefficient ($\Phi_1 = 0.6885\ s.e. = 0.1631$) indicates a positive dependence between production on a given working day and production on the corresponding working day of the previous week. In other words, above-average production in a given position of the working week tends to be followed by above-average production in the same position of the following week. The seasonal moving average coefficient is ($\Theta_1 = -0.4624;\ s.e. = 0.2013$) suggests that unexpected shocks occurring one working week earlier are partially corrected in the current period. Thus, the model captures a stable five-working-day dependence structure while allowing temporary disturbances to be dampened over time. The model achieved $RMSE = 282.08$ and $MAPE = 6.16\%$ indicating that the typical prediction error is approximately 282 units per shift and that the average relative absolute error is about 6.16%. Overall, the model provides a parsimonious representation of productivity dynamics in levels and captures the recurring weekly dependence without requiring first-order differencing.

4 Conclusions

This study analyzed productivity in an automated production line using statistical methods and time-series models. The results showed that the typical production per shift is about 3 500 units, and this value can be used for planning and decision-making. Although statistical tests suggested that the data are not perfectly normal, the deviations are small and do not affect the practical conclusions. The analysis also confirmed that production differs across working days. Production is usually higher at the beginning of the week and lower, and more variable, at the end. These results reflect real production conditions, where weekly work patterns and organization influence performance. Based on the estimated production, it was possible to calculate component consumption and plan deliveries to intermediate warehouses. The time-series analysis confirmed that productivity fluctuates around a stable level and follows a clear five-working-day dependence structure. Production on a given working day is related to production on the corresponding working day of the previous week, while weekly shocks are partially corrected. This indicates a stable production process operating under regular schedules and controlled conditions. The results show that combining statistical analysis with seasonal time-series modeling helps to better understand production and supports more effective planning and management. In future work, a SARIMAX model is planned, which will include the number of shifts per day as an additional explanatory variable.

Acknowledgements

The authors acknowledge funding from the Excellence project run at the Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic.

References

Box G E P, Jenkins G M, Reinsel G C, Ljung G M (2015) Time series analysis: Forecasting and control, 5th ed. Wiley

Cowpertwait P S P, Metcalfe A V (2009) Introductory time series with R. Springer

Frances P H, van Dijk D, Opschoor A (2014) Time Series Models for Business and Economic Forecasting, 2nd ed. Cambridge

Groebner D F, Shannon P W, Fry P C, Smith K D (2018) Business statistics: A decision-making approach, 10th ed. Pearson

Javier G M N, Evans R, Trupke T, Hameiri Z (2024) Enhancing solar cell production line monitoring through advanced statistical analysis. Solar Energy Materials and Solar Cells 274

Kumar U, Jain V K (2010) Time series models (Grey-Markov, Grey model with rolling mechanism and SARIMA) to forecast energy consumption in India. Energy 35(4): 1709–1716

Kusiak A (2017) Smart manufacturing must embrace big data. Nature 544: 23–25

Lee J, Bagheri B, Kao H-A (2015) A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters 3: 18–23

Montgomery D C (2020) Introduction to statistical quality control, 8th ed. Wiley

R Core Team (2025) R: A language and environment for statistical computing (Version 4.5.2). R Foundation for Statistical Computing. https://www.r-project.org/

Sen M C, Alkan M (2025) Time-series prediction of failures in an industrial assembly line using artificial learning. Applied Sciences 15(11)

Evaluating Strikers and Forwards in the Bundesliga with a Two-Stage Decision-Making Framework

José Rodea-Zamudio[1], Martin Flegl[2], Enrique Delahoz-Domínguez[3], Eusebio Vargas Estrada[4]

Abstract. Player performance evaluation in elite football is a central challenge for scouting, as transfer decisions involve high uncertainty and substantial sporting and financial consequences. This paper introduces a two-stage framework that integrates machine learning with Data Envelopment Analysis to evaluate strikers and central forwards in the German Bundesliga between the 2020/2021 and 2024/2025 seasons. In the first stage, a Random Forest regression identifies the SoFIFA attributes most predictive of market value; since market value reflects the observed price of a player's attribute endowment, the selected attributes constitute an economically coherent input set. In the second stage, these inputs enter an output-oriented BCC model with per-90 on-field production as outputs, yielding efficiency scores that measure output obtained per unit of purchased endowment. We report bias-corrected efficiency estimates with confidence intervals obtained via the homogeneous bootstrap, assess the stability of first-stage selection across 100 replicates, examine sensitivity to weight restrictions, and validate scores against market valuations and independent FBref production data. Across 248 player-seasons, the mean efficiency is 0.797, falling to 0.666 after bias correction; no player is distinguishable from the frontier at 95% confidence, so the frontier is best interpreted as an ordering rather than a set of attained optima. Efficiency shows no significant association with market value ($\rho = 0.093, p = 0.47$) but is moderately correlated with on-field production ($\rho = 0.512,\ p < 0.001$), underscoring that it captures performance dimensions not reflected in price alone. Finally, we demonstrate that per-90 outputs require a minimum-participation threshold: a single player with only eight minutes of play attains the frontier and halves the 2023/2024 season mean.

Keywords: Bundesliga, Bootstrap, Data Envelopment Analysis, Machine learning, Player performance evaluation, Random Forest, Sports analytics.

JEL Classification: C44, C61, L83

1 Introduction

Transfer decisions in elite football are made under high uncertainty, with substantial costs and sporting as well as financial consequences that extend over several seasons (Gaurav & Chakraborty, 2019; Villa & Lozano, 2025). As van Arem et al. (2025) observe, the risk lies not only in the magnitude of investment but also in the possibility of acquiring players whose future output justifies neither their price nor their tactical role. Traditionally, player assessment has relied on the judgment of a few experts and scores derived from limited match samples. Such evaluations are prone to bias and provide only a partial view of a player's contribution. Two issues therefore arise: first, the need to predict how a player is expected to perform in future games; second, the need to benchmark efficiency against alternatives to ensure optimal resource allocation (Baker et al., 2026; Essien et al., 2024; Oukil & Govindaluri, 2017).

Machine learning approaches have been widely applied to classify players, compare profiles, and estimate market values using large attribute datasets (Gaurav & Chakraborty, 2019; McHale & Holmes, 2023). These models

[1] Tecnologico de Monterrey, School of Engineering and Sciences, Calle del Puente 222, 14380 Mexico City, Mexico. A01668386@tec.mx. ORCID: 0009-0008-9697-2555.
[2] Tecnologico de Monterrey, School of Engineering and Sciences, Calle del Puente 222, 14380 Mexico City, Mexico. martin.flegl@tec.mx. ORCID: 0000-0002-9944-8475.
[3] Universidad del Magdalena, Department of Industrial Engineering, Carrera 32 No. 22-08, Santa Marta, Magdalena, Colombia. enriquedelahoz@unimagdalena.edu.co. ORCID: 0000-0003-2531-6389.
[4] Tecnologico de Monterrey, School of Engineering and Sciences, Carlos Lazo 100, 01389 Mexico City, Mexico. eusebio.vargas@tec.mx. ORCID: 0000-0001-6533-398X.

generate predictive insights into on-field behavior based on retrospective statistics. Yet, even accurate prediction of market value does not guarantee that a player represents an efficient use of transfer expenditure.

A complementary stream of research employs Frontier models to evaluate efficiency across alternatives, or Decision Making Units (DMUs). Data Envelopment Analysis (DEA), one of the most prominent Frontier methods, measures relative efficiency and enables benchmarking for selection, remuneration, and substitution. Applications in football have extended DEA to multiple outputs, undesirable outputs, cross-efficiency evaluation, and weight restrictions. In club contexts, DEA has been used to examine the relationship between spending and sporting efficiency, often showing that higher expenditure does not automatically yield better performance (Arsu, 2021; Villa & Lozano, 2025). However, DEA's limitation mirrors that of machine learning: while DEA compares efficiency, it does not exploit non-linear structure or provide prospective signals.

This methodological separation leaves a decisional gap. Prediction without efficiency evaluation yields incomplete selection, while efficiency comparison without a principled input set is equally incomplete. We address this gap by coupling machine learning and DEA — and by making the coupling itself the object of justification and testing rather than treating it as a simple pipeline. Specifically, we evaluate strikers and central forwards in the Bundesliga across five seasons, subjecting the resulting scores to three forms of scrutiny often omitted in similar applications: precision estimation via bootstrap, sensitivity analysis with respect to modeling parameters, and validation against external references.

2 Materials and methods

2.1 Two-Stage Design and Methodological Framework

We adopt a two-stage design that links prediction to efficiency evaluation. The rationale is economic: in the transfer market, market value represents the price of a player's attribute endowment, so attributes that predict market value are those clubs effectively purchase. Efficiency is then defined as converting this priced endowment into on-field production. This measure is not a quality ranking, since extreme endowments face proportionally higher frontiers, and attributes not priced by the market are excluded.

In the first stage, Random Forest (Breiman, 2001) aggregates regression trees grown on bootstrap samples with random feature subsets at each split, which decorrelates the ensemble and controls variance. We use it for its ability to capture non-linear attribute–price relationships and to provide attribute-level importance measures (Bertoletti et al., 2022; Cheng et al., 2024), not for prediction per se: the stage serves as a selection device. Market value was regressed on 29 SoFIFA performance attributes using 200 trees, default hyperparameters, an 80/20 train–test split, and 5-fold cross-validation with shuffling. Importance was assessed by two measures. Mean decrease in impurity, the default in most implementations, is biased toward high-variance and high-cardinality predictors — a concern here since all 29 attributes are continuous ratings with differing dispersion. We therefore also computed permutation importance on held-out data (30 repetitions), which avoids this bias. The six attributes ranked highest under permutation importance formed the input set for the second stage, chosen to respect DEA discrimination requirements given the sample size per season (Dyson et al., 2001).

In the second stage, we employ an output-oriented BCC model (Banker et al., 1984) to evaluate how much per-90 on-field production a player achieves from a given attribute endowment. This variable-returns-to-scale specification asks directly how outputs can be expanded relative to purchased inputs, distinguishing it from the constant-returns CCR model. For the unit under evaluation, denoted o, the envelopment form is:

$$\max \varphi + \varepsilon \left(\sum_i s_i^- + \sum_r s_r^+ \right) \tag{1}$$

subjected to

$$\begin{aligned} \sum_j \lambda_j \, x_{ij} + s_i^- &= x_{io}, \qquad i = 1, \dots, m \\ \sum_j \lambda_j \, y_{rj} - s_r^+ &= \varphi y_{ro}, \qquad r = 1, \dots, s \\ \sum_j \lambda_j &= 1 \, , \lambda_j \, , s_i^-, s_r^+ \geq 0, \qquad j = 1, \dots, n \end{aligned} \tag{2}$$

where x_{ij} is the quantity of input i for unit j and y_{rj} the quantity of output r, λ_j are intensity variables, s_i^- and s_r^+ are slacks, and ε is the non-Archimedean infinitesimal preventing zero weights. The constraint $\sum_j \lambda_j = 1$ imposes

variable returns to scale and is what distinguishes this model from the constant-returns CCR specification. Scores are reported as the reciprocal $1/\varphi$, which maps them onto (0,1] with unity denoting efficiency; this is the transformation applied by default in MaxDEA and is stated here to avoid ambiguity. For example, a score of 0.60 means the player attains 60% of the frontier output, equivalently that output could be expanded by a factor of 1.67.

To prevent efficiency being attained by neglecting weak dimensions, we impose assurance-region bounds on virtual weights (Thompson et al., 1990; Allen et al., 1997). Each variable must contribute at least a proportion α of its virtual quantity, with feasibility requiring $\alpha < 1/max(m, s)$. Sensitivity analysis was conducted over $\alpha \in \{0, 0.02, 0.05, 0.08, 0.12, 0.15\}$, reporting the number of efficient units, mean scores, and rank correlations relative to the unrestricted baseline.

2.2 Data, variables and robustness protocol

The analysis covers all strikers and central forwards appearing in the Bundesliga between 2020/2021 and 2024/2025, the two roles treated as one position following standard categorization (Flegl et al., 2018; Ju et al., 2023). Attribute data and market values are taken from SoFIFA, a widely used proxy dataset in sports analytics (McHale & Holmes, 2023), while on-field production is sourced from FBref. Goals, assists, and cards are expressed per 90 minutes. The raw panel comprises 319 player-seasons. Records were matched across sources using normalized identifiers rather than names, since name-based matching split one player into two orthographic variants, inflating the count from 179 to 180.

Three data issues required adjustment. First, cards are treated as inputs in the output-oriented specification. Second, 17 player-seasons with no goals or assists ("bad units") were excluded, since the radial output-oriented measure is undefined for zero outputs and solvers incorrectly assign them full efficiency. Third, per-90 rates based on very small denominators are not comparable: eleven player-seasons recorded under 90 minutes. We therefore restrict the sample to players with at least 450 league minutes (five full matches) and report the consequences in section 3.5. After these filters, 248 player-seasons and 140 distinct players remain. Table 1 defines all variables and reports their ranges.

Two robustness checks accompany the main results. Stability: the first stage was repeated across 100 random seeds, reporting each attribute's mean importance, dispersion, frequency of selection, and the mean pairwise Jaccard index of selected sets. Precision: efficiency estimates were bias-corrected via the homogeneous bootstrap of Simar & Wilson (1998, 2007) with 2,000 replications, yielding scores and 95% confidence intervals. The bootstrap was applied to the unrestricted specification, consistent with its asymptotic theory; this limitation is noted in section 5.

The first stage was computed in Python (scikit-learn). The second stage was implemented independently in both envelopment and multiplier forms and solved with the HiGHS simplex routine; agreement to solver tolerance verified the specification against its dual.

Variable	Role	Definition	Scale	Observed range
Attacking Positioning	Input	Ability to find space in the opponent defensive structure	SoFIFA 0–100	56–95
Ball Control	Input	Quality of first touch on receiving the ball	SoFIFA 0–100	56–88
Reactions	Input	Speed of response to loose balls and rebounds	SoFIFA 0–100	49–94
Finishing	Input	Conversion of shots taken inside the box	SoFIFA 0–100	59–95
Volleys	Input	Execution of shots struck before the ball bounces	SoFIFA 0–100	45–89
Shot Power	Input	Force generated when striking the ball	SoFIFA 0–100	58–94
Cards per 90	Input	Yellow and red cards summed, unweighted	Events / 90 min	0.00–2.00
Goals per 90	Output	League goals scored	Events / 90 min	0.00–1.50
Assists per 90	Output	League assists provided	Events / 90 min	0.00–0.71

Table 1 Variables entering the second stage: definitions and observed ranges

3 Results

3.1 Stage one: attributes priced by the market

The Random Forest explains market value well. Cross-validated R^2 is 0.926 (SD 0.030 across folds), with RMSE of €5,388,469 and MAE of €2,644,986 against a mean market value of €12,251,332 and a median of €4,800,000. Since the error metrics are in euros, the gap between mean and median highlights the strongly right-skewed distribution against which they should be read. Predictive accuracy is stable across replicates ($R^2 = 0.927 \pm 0.044$ over 100 seeds). Attribute importance, however, is less stable, and the two measures diverge materially (Figure 1). Under impurity-based importance the leading attributes are Attacking Positioning (0.381), Ball Control (0.187), Reactions (0.159), and Finishing (0.095). Under permutation importance on held-out data, Ball Control dominates (0.489), Reactions falls to 0.009, and Volleys — fifth by impurity at 0.081 — has importance indistinguishable from zero. Only three of the six selected attributes are common to both rankings.

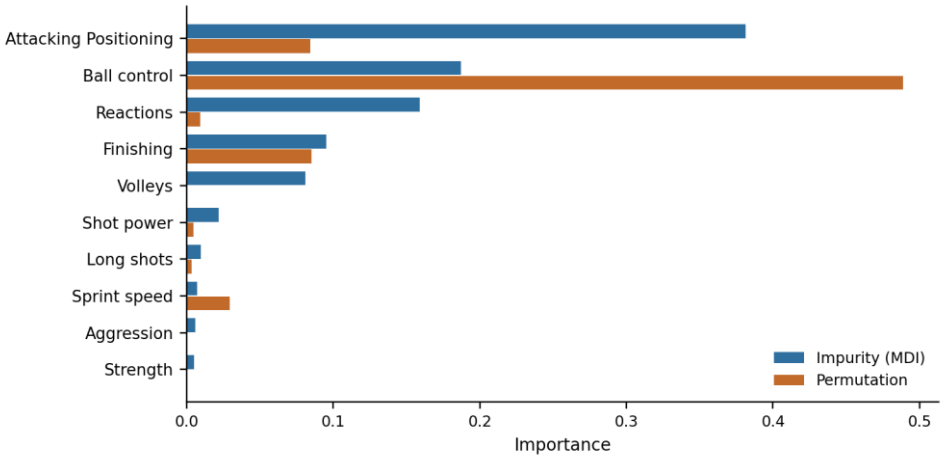


Figure 1 Attribute importance under impurity-based and permutation measures

Table 2 quantifies selection stability. Three attributes enter the selected six in every replicate; ten different attributes appear at least once. The mean pairwise Jaccard index of selected sets is 0.777 across 100 seeds for impurity-based importance and 0.713 across 20 seeds for permutation importance. The input set is therefore not an artefact of a single seed, but neither is it uniquely determined: roughly one attribute in six is interchangeable across replicates. We retain the impurity-based selection for the main analysis to preserve comparability with the earlier specification and examine the consequences of this choice in section 3.3.

Attribute	MDI	MDI SD	Permutation	Top-6 freq. (MDI)	Top-6 freq. (perm.)
Attacking Positioning	0.381	0.082	0.084	1.00	1.00
Ball Control	0.187	0.059	0.489	1.00	1.00
Reactions	0.159	0.049	0.009	0.98	0.95
Finishing	0.095	0.047	0.085	1.00	1.00
Volleys	0.081	0.035	−0.000	0.89	0.50
Shot Power	0.022	0.019	0.005	0.56	0.75
Long Shots	0.010	0.017	0.004	0.32	0.25
Sprint Speed	0.007	0.003	0.030	0.02	0.45

Note: MDI SD is the standard deviation of impurity-based importance across 100 replicates. Mean pairwise Jaccard index of the selected six: 0.777 (MDI, 100 seeds) and 0.713 (permutation, 20 seeds).

Table 2 Variables entering the second stage: definitions and observed ranges

3.2 Stage two: efficiency and its precision

Across the 248 retained player-seasons, the mean score is 0.797 (SD 0.231), with 99 units attaining unity on the deterministic frontier. Season means are similar, ranging from 0.762 to 0.833, with no marked departures. Table 3 reports the distribution by season together with bootstrap results. The bootstrap alters the interpretation substantially. The mean bias is −0.250 on the Farrell scale, reducing the bias-corrected mean to 0.666. More importantly, none of the 248 units retains a lower confidence bound at unity: at 95% confidence no player is distinguishable from the frontier. The median interval width is 0.192. Thus, the deterministic count of efficient units does not estimate how many players attained best practice; it reflects the artefact of estimating a frontier from the sample it envelops. The defensible outputs of the second stage are the player ordering and the confidence intervals around it, not the identity of an efficient set.

Season	*n*	Mean	SD	Efficient (det.)	Mean (bias-corr.)	Mean bias
2020/2021	53	0.762	0.262	19	0.620	−0.308
2021/2022	47	0.833	0.217	23	0.708	−0.211
2022/2023	47	0.819	0.220	18	0.693	−0.230
2023/2024	53	0.797	0.226	22	0.663	−0.257
2024/2025	46	0.780	0.226	17	0.655	−0.247
All	248	0.797	0.231	99	0.666	−0.250

Note: bias is reported on the Farrell scale (φ≥1); scores are its reciprocal. No unit has a lower 95% confidence bound at unity in any season.

Table 3 Variables entering the second stage: definitions and observed ranges

Player-level scores are aggregated across seasons by arithmetic mean. The geometric mean yields an almost identical ordering ($\rho = 0.998$), confirming that aggregation choice does not affect conclusions. The panel is unbalanced — 63 of the 140 players appear in two or more seasons — so we report the number of seasons alongside each score and restrict comparative statements to players observed at least twice. The full player table is provided as supplementary material.

3.3 Robustness, Validation, and Sensitivity Analyses

Table 4 reports the effect of α on discrimination and rank stability. The restriction reduces the number of efficient units from 99 to 31 as α rises from 0 to 0.12, but rank correlation against the unrestricted baseline declines steadily, reaching 0.613 at $\alpha = 0.12$. Thus, beyond modest values the restriction imposes a different ordering rather than sharpening a stable one. We therefore retain the unrestricted specification as the main result and treat the restriction as a sensitivity dimension rather than a modelling choice.

α	Efficient units	Mean score	Spearman vs. $\alpha = 0$
0 (unrestricted)	99	0.798	1.000
0.02	77	0.756	0.874
0.05	62	0.714	0.838
0.08	44	0.661	0.737
0.12	31	0.596	0.613

Table 4 Sensitivity of results to the virtual weight restriction α

External validation is shown in Table 5, relating efficiency scores (bias-corrected) to three independent variables: average market value, goals-plus-assists per 90 minutes, and average playing time. Scores are essentially uncorrelated with market value ($\rho = 0.093, p = 0.47$) among players with more than two seasons, which is consistent with efficiency being distinct from endowment. Clubs already observe market value; a strong correlation would render efficiency redundant. Players with fewer than two seasons are excluded since their averages are undefined.

Reference	Spearman ρ	p-value	n	Sample
SoFIFA market value	−0.056	0.51	140	All
SoFIFA market value	+0.093	0.47	63	≥ 2 seasons
Goals + assists per 90	+0.344	<0.001	140	All
Goals + assists per 90	+0.512	<0.001	63	≥ 2 seasons
Minutes played	−0.090	0.29	140	All
Minutes played	+0.016	0.90	63	≥ 2 seasons

Note: Market value here is the SoFIFA valuation, which also generated the first-stage selection. This makes the near-zero correlation a conservative finding rather than an optimistic one: any contamination would bias it upward, and none is observed. Validation against a fully independent valuation such as Transfermarkt remains desirable and is left to future work.

Table 5 Rank correlation between bias-corrected efficiency and external references

The participation threshold is a specification requirement. Without it, a player with only eight minutes in 2023/2024 and one assist attains the frontier, reducing the season mean from 0.616 to 0.316. Thresholds also stabilize season comparisons: without restriction, the SD of season means is 0.125; at 90 minutes it falls to 0.064, and at 450 minutes to 0.026. We adopt 450 minutes (five full matches); a 900-minute threshold yields a similar picture (mean 0.839, SD 0.025) but discards a third of the sample.

Finally, sensitivity to input selection was tested by repeating the second stage with three alternative six-variable sets: impurity-based Random Forest, permutation-based, and a tactical selection. Rank correlations between resulting orderings are 0.768 (impurity vs. permutation), 0.840 (impurity vs. tactical), and 0.690 (permutation vs. tactical). These values show that input choice is a genuine source of variation in efficiency measurement. Taken together with the Jaccard results of section 3.1, this indicates that any two-stage design inherits the instability of its selection step, and studies reporting a single input set without alternatives risk overstating precision.

4 Discussion

The corrected radial output-oriented scores are reported in Table 5 for the subsample of 49 players observed in at least two seasons. The highest scores are attained by Alario (0.835), Ekitike (0.832), Lewandowski (0.831), Kane (0.822), Behrens (0.819), and Haaland (0.808). This list is illustrative rather than exhaustive. The most revealing cases are those where efficiency diverges markedly from market valuation. Behrens (€2.9M), Ingvartsen (€3.8M), and Polter (€1.9M) achieve conversion rates comparable to Lewandowski (€10.5M). Although Lewandowski's per-90 output (0.73 goals+assists) is far superior, the three lower-valued players operate at the limits of what their attributes predict. For clubs unable to sign elite talent, such players represent extreme value, as they maximize the conversion of endowment into output. By contrast, low efficiency scores indicate that a contracted player does not perform as expected given his endowment. This may result from poor integration, injuries, or other external circumstances that reduce on-field output. For transfer targets, a low score suggests that the asking price is not justified by current production. In such cases, additional contextual arguments must complement the efficiency measure to explain underperformance.

The near-zero correlation with market value is the result that most directly demonstrates the framework's contribution. If efficiency were strongly correlated with valuation, it would be redundant with information clubs already possess. That it is not, while still correlating with production, shows that the ordering captures conversion efficiency — information valuation does not encode. Where efficiency is high but valuation low, the framework highlights potential underpriced candidates, which is precisely its intended use.

Two cautions bound these readings. First, no player is statistically distinguishable from the frontier, so differences between adjacent ranks should not be interpreted; only broad positions in the ordering are defensible. Second, the ordering shifts appreciably under alternative input sets and stronger weight restrictions, meaning the framework is best used to identify candidates for closer examination rather than to produce definitive rankings.

5 Conclusion

We have proposed and tested a two-stage framework combining Random Forest attribute selection with output-oriented DEA to evaluate strikers and central forwards in the Bundesliga across five seasons. The contribution lies not in the novelty of combining methods, which is by now unremarkable, but in articulating why the combination is coherent: market value represents the price of an attribute endowment, while efficiency measures output per unit of that endowment. Our assessment of whether the resulting scores can be believed is qualified. The ordering is defensible and provides information not contained in market valuation; the identification of an efficient set, however, is not.

Three methodological findings extend beyond this application. First, radial output-oriented models assign unity to units with zero output, rating as fully efficient players who produced nothing; such units must be excluded rather than scored. Second, per-90 rates require a minimum participation threshold, without which a single brief appearance can distort season means — for example, an eight-minute cameo with one assist can define a frontier and halve a season average. Third, virtual weight restrictions are infeasible for units with zero output and must be applied only where outputs are strictly positive. Each of these issues is easy to overlook, yet each materially distorts results.

Several limitations bound the conclusions. The bootstrap was applied to the unrestricted specification, since asymptotic theory for weight-restricted estimators is not established; the reported intervals therefore describe the precision of the unrestricted frontier. Market value and attributes both come from SoFIFA, so external validation is not fully independent and should be repeated against Transfermarkt. SoFIFA ratings are constructed assessments rather than direct measurements, and any bias in their construction propagates through both stages. The output set is confined to goals and assists, omitting defensive contribution, pressing, and link play, which understates the production of players in withdrawn roles. Finally, input selection is only moderately stable, and resulting orderings differ appreciably across selection routes. Extension to other leagues and positions would allow the stability of the input set to be examined across competitive contexts, which is the natural next question raised by this work. A larger pooled sample would also narrow the confidence intervals, which here remain wide enough that the frontier itself is statistically unattainable.

References

Allen R, Athanassopoulos A, Dyson R G, Thanassoulis E (1997) Weights restrictions and value judgements in Data Envelopment Analysis. Annals of Operations Research 73:13-34. https://doi.org/10.1023/A:1018968909638

Arsu T (2021) Investigation into the efficiencies of European football clubs with bi-objective multi-criteria data envelopment analysis. Decision Making: Applications in Management and Engineering 4(2):106-125. https://doi.org/10.31181/dmame210402106a

Baker J, Cattle A, McAuley A, Kelly A, Johnston K (2026) Will artificial intelligence solve the riddle of athlete development? A critical review of how AI is being used for athlete identification, selection, and development. Psychology of Sport & Exercise 82:102978. https://doi.org/10.1016/j.psychsport.2025.102978

Banker RD, Charnes A, Cooper W W (1984) Some models for estimating technical and scale inefficiencies in Data Envelopment Analysis. Management Science 30(9):1078-1092. https://doi.org/10.1287/mnsc.30.9.1078

Bertoletti A, Berbegal-Mirabent J, Agasisti T (2022) Higher education systems and regional economic development in Europe: A combined approach using econometric and machine learning methods. Socio-Economic Planning Sciences 82(Part A):101231. https://doi.org/10.1016/j.seps.2022.101231

Breiman L (2001) Random forests. Machine Learning 45:5-32. https://doi.org/10.1023/A:1010933404324

Cheng B, Liu Y, Jia Y (2024) Evaluation of students' performance during the academic period using the XG-Boost Classifier-Enhanced AEO hybrid model. Expert Systems with Applications 238(Part D):122136. https://doi.org/10.1016/j.eswa.2023.122136

Dyson R G, Allen R, Camanho A S, Podinovski V V, Sarrico C S, Shale E A (2001) Pitfalls and protocols in DEA. European Journal of Operational Research 132:245-259. https://doi.org/10.1016/S0377-2217(00)00149-1

Essien A, Kremantzis M, Joshi D, Zaghloul F (2024) Moneyball: Analyzing the efficiency of English Premier League strikers using data envelopment analysis. In: Business Analytics and Decision Making in Practice. Lecture Notes in Operations Research, Springer Nature. https://doi.org/10.1007/978-3-031-61589-4_11

Flegl M, Jiménez-Bandala C A, Lozano C, Andrade L (2018) Personnel selection in complex organizations: A case of Mexican football team for the 2018 World Cup in Russia. Revista del Centro de Investigación de la Universidad La Salle 13(49):43-66. https://doi.org/10.26457/recein.v13i49.1510

Gaurav V, Chakraborty G (2019) Scouting in soccer with applied machine learning. Proceedings of the SAS Global Forum 2019 Conference, Paper 3828-2019. SAS Institute Inc.

Ju W, Doran D, Hawkins R, Evans M, Laws A, Bradley P S (2023) Contextualised high-intensity running profiles of elite football players with reference to general and specialised tactical roles. Biology of Sport 40(1):291-301. https://doi.org/10.5114/biolsport.2023.116003

McHale IG, Holmes B (2023) Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. European Journal of Operational Research 306(1):389-399. https://doi.org/10.1016/j.ejor.2022.06.033

Oukil A, Govindaluri S M (2017) A systematic approach for ranking football players within an integrated DEA-OWA framework. Managerial and Decision Economics 38(5):663-678. https://doi.org/10.1002/mde.2851

Simar L, Wilson P W (1998) Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. Management Science 44(1):49-61. https://doi.org/10.1287/mnsc.44.1.49

Simar L, Wilson P W (2007) Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics 136(1):31-64. https://doi.org/10.1016/j.jeconom.2005.07.009

Thompson R G, Langemeier L N, Lee C T, Lee E, Thrall RM (1990) The role of multiplier bounds in efficiency analysis with application to Kansas farming. Journal of Econometrics 46(1-2):93-108. https://doi.org/10.1016/0304-4076(90)90048-X

van Arem K, Goes-Smit F, Söhl J (2025) Forecasting the future development in quality and value of professional football players. Applied Sciences 15:8916. https://doi.org/10.3390/app15168916

Villa G, Lozano S (2025) Assessing the efficiency of football clubs using data envelopment analysis. IMA Journal of Management Mathematics 36(2):197-213. https://doi.org/10.1093/imaman/dpae014

How University Students Learn with Generative AI: Usage Patterns, Study Materials, and Risk Awareness

Petr Rozehnal[1], Lucie Klaubenschalk Chytilová[2], Vítězslav Novák[3], Miroslav Langer[4]

Abstract. Artificial intelligence tools are becoming increasingly embedded in university students' study practices, yet their role in the learning process remains insufficiently understood. This study examines how frequently students use AI, what types of study materials they rely on, and whether AI is employed primarily as an autonomous source of information or as a tool for processing user-provided content. Data were collected through a questionnaire survey of 171 students from two Czech universities. The responses were analysed using descriptive statistics, exact binomial tests, paired Wilcoxon signed-rank tests, and logistic regression. The results show that 78.9% of respondents use AI tools at least several times per week, although only 40.9% use them daily. Most students primarily learn from their own materials, prepared using university resources. Students use AI significantly more often to process materials they provide themselves than to generate study content solely from the model's internal knowledge. Frequent or very frequent use was reported by 54.4% and 14.2% of respondents, respectively, and the difference was statistically significant. Daily AI use was also positively associated with paid access and customization of the AI environment. The findings indicate that the surveyed students predominantly use AI as a learning assistant and content-processing tool rather than as an independent source of knowledge. Higher education should therefore focus on developing critical, transparent, and source-aware AI-supported learning practices.

Keywords: artificial intelligence, higher education, student learning, study strategies, generative AI, survey research

JEL Classification: I21, I23, C12

1 Introduction

Artificial intelligence (AI) tools have quickly become a phenomenon of the current era. In all fields of human activity, questions are being asked about the impacts of this technology. Education is no exception. The situation is all the more complicated because even before the advent of AI, discussions were held on teaching methods in the era of ongoing digitalisation.

Traditional education based on student-teacher interaction, or rather communication, sharing and exchange of information (including study materials, student work results, etc.), has been fundamentally disrupted by digitalisation. Now, the impact of AI is a much bigger challenge for education.

The fact that students use AI tools to support their studies is beyond discussion. However, this opens up a wide range of questions. Typically, it concerns the frequency and scope of use, the purposes for which the tools are used, etc. Ward et al. (2024) state that over 40% of students use AI tools frequently, and only 2% never do. According to Kumar (2026), 51% of young people are daily users. Considering the relatively short time that AI tools have been available, these are high numbers. It is assumed that use will continue to expand. This leads to further considerations about the form of use and purpose. Ward et al. (2024) monitored use in the classifications of AI tools: AI tutoring systems, smart note-taking AI tools, planning/scheduling applications and others. Ibieta et

[1] VSB-Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava Poruba, Czechia, petr.rozehnal@vsb.cz.
[2] Czech University of Life Sciences Prague, Faculty of Economics and Management, Kamýcká 129, 165 00 Prague, Czechia, chytilova@pef.czu.cz, VSB-Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava Poruba, Czechia, lucie.klaubenschalk.chytilova@vsb.cz.
[3] VSB-Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava Poruba, Czechia, vitezslav.novak@vsb.cz..
[4] VSB-Technical University of Ostrava, Faculty of Economics, 17. listopadu 2172/15, 708 00 Ostrava Poruba, Czechia, miroslav.langer@vsb.cz.

al. (2025) list the categorisation of use as: Information repository, editing tools and cognitive tools. In our survey, we focused on the creation of digital content (whether in the form of input for study – study materials or as an output of study – projects, etc.). Similarly, Hong (2026) distinguished, in his research, the orientations towards specific tasks in which students used AI (e.g., coding, creating study materials, searching in the context of a quiz/exam, drawing conclusions). In this way, we follow research on tools that help students in the learning process (Wang et al., 2023; Yan and Qianjun, 2025). Or, and this must be admitted, they replace the student's work, which is a negative aspect associated with IT in general. We included this aspect in our survey, in which we also monitor students' efforts in the educational process (or the tendency towards superficial study, in which AI tools are supposed to replace depth of study). Research and analysis from this perspective can further advance more complex considerations, as Chiu (2023) notes in the need to reopen questions about what education will offer in the future. Vieriu and Petrea (2025) state that 83.5% of students believe that AI improves their learning efficiency: it facilitates quick access to educational resources and reduces the time needed to find information. However, this result may also include students who use AI tools to simplify and streamline their studies at the expense of quality, i.e., they prioritise time and efficiency while failing to address the real decline in the quality of knowledge. If AI tools help with searching and reporting information, the context is often omitted; the educational process is straightforward but poorer as a result. This ethical dimension of the use of AI is discussed in (Lund et al. 2025).

Previous studies have demonstrated the rapid adoption of generative AI tools among university students and have examined their use for information retrieval, writing, coding, planning, and academic support. However, AI use is often treated as a single behaviour, with limited attention paid to the degree of student control over the information processed by the AI system. In particular, it remains unclear whether students primarily use AI as an autonomous source of content or as a tool for transforming materials that they have selected and provided themselves.

This distinction is important from an educational perspective. When students provide their own materials, they retain greater control over the sources and scope of the generated content. In contrast, relying exclusively on the model's internal knowledge may increase the risk of inaccurate, incomplete, or unverifiable information. At the same time, both approaches may support learning if students critically evaluate the resulting outputs.

The present study addresses this research gap using questionnaire data collected from 171 students at two Czech universities. The study examines the intensity and configuration of AI use, the sources and formats of students' learning materials, and the ways in which AI is integrated into the preparation of study materials and creative academic outputs. It also considers students' awareness of hallucinations, copyright, plagiarism, and institutional rules. The main contribution lies in connecting patterns of AI use with students' broader learning practices rather than examining the frequency of AI adoption in isolation.

2 Research Questions and Hypotheses

Previous research indicates that generative AI has become a common component of university students' study practices. However, the intensity of its use, the sources of learning materials, and the degree to which students retain control over the information processed by AI may differ substantially. Based on the literature review and the questionnaire design, the following research questions and hypotheses were formulated.

RQ1: How frequently do university students use generative AI tools?

H1: More than half of the surveyed students use AI tools on a daily basis.

RQ2: What types and sources of materials do students primarily use for learning?

H2: More than half of the surveyed students primarily learn from self-prepared materials based on resources provided by their university.

RQ3: Do students use AI primarily as an autonomous source of information or as a tool for processing user-provided materials?

H3: Students use AI more frequently to process materials that they provide themselves than to generate study materials solely from the model's internal knowledge.

RQ4: Does the frequency of AI use differ between the preparation of study materials and the creation of other academic outputs?

H4: Students use AI more frequently for the preparation of study materials than for the creation of projects, seminar papers, and programming code.

In addition, an exploratory analysis was conducted to determine whether paid access and customisation of the AI environment were associated with daily AI use. Students' awareness of hallucinations and risks related to copyright, plagiarism, and institutional rules was examined descriptively.

3 Data and Methods

3.1 Survey Design and Sample

The study was based on an anonymous cross-sectional questionnaire survey conducted among university students in the Czech Republic between 20 March and 20 May 2026. Respondents were students taught by members of the research team at VŠB – Technical University of Ostrava and Silesian University in Opava. The questionnaire was distributed electronically through a web link and was completed outside scheduled classes.

A non-probability convenience sampling strategy was used. Students were recruited across the range of courses taught by the research team during the data-collection period, primarily at the Faculty of Economics and in courses coordinated by the Department of Systems Engineering and the Institute of Mathematics. The sample was therefore not restricted to a single course or study programme. Approximately 500 students were invited to participate. However, because the exact number of students who received or opened the survey link was not recorded, an exact response rate could not be calculated.

A total of 171 questionnaires were submitted and retained in the dataset. A small number of item-level responses were missing and were handled separately for each analysis, as described below.

Participation was voluntary. Before completing the questionnaire, respondents were informed about the purpose of the study, the anonymous processing of their responses, and the intended use of aggregated results for scientific purposes. The survey was conducted with the knowledge and agreement of the management of the participating institutions. No directly identifying personal information was included in the analytical dataset.

The questionnaire was developed by the authors specifically for this exploratory study and was administered in Czech. It contained 23 substantive items, including single-choice, multiple-response, and five-point ordinal questions. The questionnaire was designed to take approximately five minutes to complete. Respondents were not required to answer every question, which resulted in a small number of item-level missing responses. No formal pilot study or prior psychometric validation of the questionnaire was conducted. The complete questionnaire is available from the authors upon reasonable request.

3.2 Variables and Statistical Methods

Frequency-based items concerning AI-supported activities were measured using a five-point ordinal response scale: 1 = never, 2 = rarely, 3 = occasionally, 4 = frequently, and 5 = very frequently. Higher values indicated a greater reported frequency of the respective activity. The risk-awareness items used an analogous five-point scale ranging from "never" to "very frequently aware". For descriptive reporting, the categories "frequently" and "very frequently" were combined to represent frequent use. The original five-category responses were retained for the paired Wilcoxon signed-rank tests. Multiple-response questions concerning the AI tools and learning-material formats used by students were transformed into separate binary variables for each response option.

Descriptive statistics were used to characterise the sample and summarise the frequency, form, and purpose of AI use. Absolute and relative frequencies were reported for categorical variables, while medians and interquartile ranges were used for ordinal variables.

Exact one-sample binomial tests were used to determine whether selected proportions exceeded 50%. These one-sided tests were applied to the proportion of students who used AI daily and to the proportion who primarily learned from self-prepared materials based on university resources.

Paired Wilcoxon signed-rank tests were used to compare the frequency of activities reported by the same respondents. The first test compared the use of AI to generate study materials solely from the model's internal knowledge with its use to process materials provided by the student. The second test compared AI use for creating study materials with its use for projects, seminar papers, and programming code. Effect sizes for the paired comparisons were calculated as $r = \frac{|Z|}{\sqrt{N}}$, where N represented the number of valid paired observations.

For the exploratory binary logistic regression, daily AI use was coded as 1 for respondents reporting use "very frequently, effectively daily" and 0 for all lower-frequency categories. Paid AI access, customisation of the AI environment, higher study level, and male gender were entered as binary predictors. Bachelor's study, female gender, free AI access, and no customisation served as the reference categories. Master's and doctoral students were combined because only one doctoral student was included in the sample. Regression results were reported as

odds ratios with 95% confidence intervals. Overall model significance, Nagelkerke's pseudo-R2, and the Hosmer–Lemeshow goodness-of-fit test were also reported.

The full dataset contained 171 respondents. Missing observations were handled on an analysis-by-analysis basis. Descriptive statistics were calculated using the available valid responses for each item, and valid sample sizes were reported where relevant. The exact binomial tests for H1 and H2 were based on all 171 respondents, whereas each paired Wilcoxon signed-rank test and the binary logistic regression included 169 valid observations. Statistical significance was assessed at the 5% level. Initial data preparation, coding, and descriptive summaries were conducted in Microsoft Excel, while inferential statistical analyses were performed using IBM SPSS Statistics.

4 Results

4.1 Sample Characteristics and General AI Use

The final sample consisted of 171 respondents from two Czech universities. The main characteristics of the sample and selected indicators of AI use are summarized in Table 1.

Variable	Category	N	%
Gender	Men	94	55.0
	Women	76	44.4
University	VŠB–TUO	157	91.8
	Silesian University	14	8.2
Study level	Bachelor's	119	69.6
	Master's	51	29.8
AI use	Daily	70	40.9
	Several times per week	65	38.0
Paid AI access	Yes	96	56.5
AI customization	Yes	53	31.4

Table 1 Sample characteristics and selected indicators of AI use

Note: Only selected categories are displayed. Percentages were calculated from valid responses and may not sum to 100% because of omitted categories and item-level missing responses.

The sample consisted of 171 respondents, of whom 55.0% were men, 44.4% were women, and 0.6% preferred not to indicate their gender. Most respondents studied at VŠB – Technical University of Ostrava (91.8%), while 8.2% attended Silesian University in Opava. Bachelor's students accounted for 69.6% of the sample, master's students for 29.8%, and one respondent was enrolled in a doctoral programme.

AI tools were used daily by 40.9% of respondents, while a further 38.0% used them several times per week. Overall, 78.9% of respondents therefore used AI at least several times per week. However, the proportion of daily users did not exceed 50%; consequently, H1 was not supported.

Paid AI services were used by 56.5% of respondents, whereas 31.4% reported customising their AI environment by creating agents or making other adjustments. Gemini and ChatGPT were the most widely used tools, reported by 75.3% and 71.8% of respondents, respectively. As respondents could select multiple tools, the percentages do not sum to 100%.

4.2 Learning Materials and Study Orientation

Most respondents studied primarily from their own materials prepared based on resources provided by their university (59.1%). A further 36.8% relied primarily on unofficial external resources, such as online platforms and shared student materials, while only 4.1% relied primarily on official external materials. The proportion of students using self-prepared university-based materials was significantly greater than 50% (exact binomial test, $p = 0.011$), supporting H2.

Digitally typed text was the most commonly used learning format, selected by 85.3% of respondents. Handwritten materials, subsequently converted to digital form, were used by 58.8%, while 51.2% used handwritten materials on paper. In terms of study orientation, 55.9% reported that they typically studied primarily to prepare for examinations and course requirements. Only 17.1% reported studying a subject comprehensively, including its wider context and relationships.

4.3 AI-Supported Learning Activities

A substantial difference was observed between the two ways in which AI was used to obtain study materials. Among respondents with valid data for both items ($N = 169$), students used AI significantly more frequently to process materials they provided themselves (*Mdn = 4, IQR 3–4*) than to generate study materials solely from the model's internal knowledge (*Mdn = 2, IQR 2–3*), $Z = -8.606$, $p < 0.001$, $r = 0.662$. The effect size was large, supporting H3.

The results therefore suggest that students generally retain control over the source materials and use AI primarily to transform, summarise, organise, or explain content that they have selected themselves. AI appears to function more frequently as a content-processing assistant than as an autonomous source of knowledge.

4.4 Study Materials, Creative Outputs, and Risk Awareness

The results of the statistical tests used to evaluate the four hypotheses are summarised in Table 2.

Hypothesis	Evaluated relationship	Statistical test	Result	Conclusion
H1	More than 50% use AI daily	Exact binomial test	p = 0.993	Not supported
H2	More than 50% use self-prepared university-based materials	Exact binomial test	p = 0.011	Supported
H3	User-provided materials vs model knowledge	Paired Wilcoxon signed-rank test	p < 0.001, r = 0.662	Supported
H4	Study materials vs other academic outputs	Paired Wilcoxon signed-rank test	p = 0.109, r = 0.123	Not supported

Table 2 Summary of hypothesis testing

AI was used frequently or very frequently to create study materials by 42.9% of respondents. The corresponding proportion for projects, seminar papers, and programming code was 48.8%. Among respondents with valid data for both items (N = 169), the median frequency was 3 for both activities. The interquartile range was 2–4 for the creation of study materials and 3–4 for projects, seminar papers, and programming code. The paired Wilcoxon signed-rank test did not demonstrate a statistically significant difference, $Z = -1.604$, $p = 0.109$, $r = 0.123$. The effect size was small, and H4 was therefore not supported.

Respondents nevertheless reported a relatively high level of awareness of AI-related risks. A total of 87.1% frequently or very frequently considered the risk of inaccurate or hallucinated information. Risks related to copyright, plagiarism, and institutional rules were frequently or very frequently considered by 65.9% of respondents.

4.5 Factors Associated with Daily AI Use

An exploratory binary logistic regression was conducted to examine whether paid AI access and customisation of the AI environment were associated with daily AI use while controlling for study level and gender. The dependent variable distinguished between respondents who used AI daily and those who used it less frequently. The analysis included 169 respondents with complete data.

The overall model was statistically significant, $\chi^2(4) = 27.943$, $p < 0.001$, and yielded a Nagelkerke pseudo-R^2 of 0.205. The Hosmer–Lemeshow test did not indicate a lack of model fit, $\chi^2(7) = 1.822$, $p = 0.969$. The model correctly classified 68.0% of respondents.

Predictor	B	SE	OR	95% CI for OR	p
Paid AI access	0.840	0.363	2.316	1.138 – 4.716	0.021
AI environment customization	1.091	0.376	2.976	1.425 – 6.215	0.004
Master's or doctoral study	0.626	0.369	1.871	0.907 – 3.857	0.090
Male gender	-0.653	0.349	0.520	0.263 – 1.032	0.061

Table 3 Binary logistic regression predicting daily AI use

Note: Dependent variable: daily AI use (1 = daily, 0 = less than daily). Reference categories: free AI access, no customization, bachelor's study, and female gender. N = 169. OR = odds ratio; CI = confidence interval.

Students using paid AI services had significantly higher odds of daily AI use than students relying on free services, *OR = 2.316, 95% CI [1.138, 4.716], p = 0.021*. Customisation of the AI environment was also significantly associated with daily AI use, *OR = 2.976, 95% CI [1.425, 6.215], p = 0.004*. Study level, *OR = 1.871, 95% CI [0.907, 3.857]*, *p = 0.090*, and gender, *OR = 0.520, 95% CI [0.263, 1.032], p = 0.061*, were not statistically significant predictors at the 5% level.

5 Discussion

The findings confirm that generative AI has become a regular component of the surveyed students' study practices. Although fewer than half of the respondents used AI daily, almost four out of five used it at least several times per week. The results, therefore, indicate widespread adoption but do not support the assumption that daily use is already dominant.

The most important finding concerns the distinction between AI as an autonomous information source and AI as a tool for processing student-selected materials. Respondents were substantially more likely to provide their own materials and ask AI to transform them than to rely exclusively on the model's internal knowledge. This pattern suggests that AI-supported learning cannot be understood simply as the replacement of students' work. In many cases, students appear to remain involved in selecting inputs and defining the task's scope.

Nevertheless, the provision of source materials does not guarantee critical or high-quality learning. AI may still omit important context, misinterpret the supplied documents, or produce inaccurate summaries. The high declared awareness of hallucinations is therefore encouraging, but the survey does not demonstrate whether respondents actually verify AI-generated outputs in practice. Awareness and effective verification behaviour should be distinguished in future research.

The absence of a significant difference between the creation of study materials and creative academic outputs is also relevant. AI is not used solely for summarising or preparing notes but is integrated into a broader range of academic tasks, including projects, written assignments, and programming. Universities should therefore avoid limiting AI-related guidance to plagiarism prevention. Greater attention should be paid to source verification, transparent disclosure of AI assistance, responsible prompting, and the interpretation of generated outputs.

The exploratory regression results indicate that daily AI use was positively associated with paid access and customization of the AI environment, even after controlling for study level and gender. This suggests that the depth of AI adoption may depend not only on students' attitudes but also on access to advanced features and their ability to configure AI tools. Such differences may disadvantage students who rely exclusively on freely available tools or possess less advanced AI skills.

6 Conclusion and Limitations

This study examined how university students use generative AI in relation to their broader learning practices. The results show that AI use is widespread among the surveyed students, although daily use does not yet characterise the majority of the sample. Students primarily learn from self-prepared materials based on university resources and use AI considerably more frequently to process materials that they provide themselves than to generate study content solely from the model's internal knowledge.

The findings suggest that generative AI predominantly serves as a learning and content-processing assistant rather than as a fully autonomous source of knowledge. At the same time, its use extends beyond preparing study materials to projects, seminar papers, and programming tasks. The exploratory regression analysis further indicated that daily AI use was positively associated with paid access and customisation of the AI environment. Universities

should therefore support students in the critical and transparent use of AI, with particular emphasis on source verification, hallucination detection, copyright, plagiarism, and compliance with institutional rules.

Several limitations should be considered. The study used a non-probability convenience sample drawn from only two Czech universities and was strongly dominated by respondents from one institution. The results cannot therefore be generalised to the wider population of university students. The study relied on self-reported, cross-sectional data and did not measure actual learning outcomes, academic performance, verification behaviour, or the quality of AI-generated materials. The observed associations between daily AI use, paid access, and customisation should not be interpreted as causal relationships.

The questionnaire was developed specifically for this exploratory study and was not subjected to formal pilot testing or psychometric validation. Although its items were designed to reflect the research questions directly, their interpretation and measurement properties require further examination. In addition, some questionnaire items combined several different activities, particularly projects, seminar papers, and programming code, which limited the interpretation of activity-specific differences. Future research should use larger and more diverse samples, validate and refine the questionnaire, distinguish between individual types of academic tasks, and connect students' reported AI practices with objective measures of learning, verification behaviour, and output quality.

References

Chiu T K F (2023) The impact of Generative AI (GenAI) on practices, policies and research direction in education: a case of ChatGPT and Midjourney. Interactive Learning Environments 32(10): 6187-6203. https://doi.org/10.1080/10494820.2023.2253861

Hong B H (2026) From search to work tasks: exploring students' AI chatbot information practices in higher education. Information Research an International Electronic Journal 31(iConf): 1551–1557. https://doi.org/10.47989/ir31iConf64180

Ibieta A, Aguilera S, Constanza-Medina M, Ignacia-Sepúlveda M, Castellanos-Alvarenga L M (2025) Academic Use of Artificial Intelligence Tools by University Students. In: Tomczyk, Ł. (eds) New Media Pedagogy: Research Trends, Methodological Challenges, and Successful Implementations. NMP 2024. Communications in Computer and Information Science 2537. Springer, Cham. https://doi.org/10.1007/978-3-031-95627-0_23

Kumar N (2026) AI in Education Statistics 2026 (Global Trends & Facts). https://www.demandsage.com/ai-in-education-statistics/

Lund B D, Lee T H, Mannuru N R et al. (2025) AI and Academic Integrity: Exploring Student Perceptions and Implications for Higher Education. J Acad Ethics 23: 1545–1565. https://doi.org/10.1007/s10805-025-09613-3

Vieriu A M, Petrea G (2025) The Impact of Artificial Intelligence (AI) on Students' Academic Development. Education Sciences 15(3): 343. https://doi.org/10.3390/educsci15030343

Wang T, Lund B D, Marengo A, Pagano A, Mannuru N R, Teel Z A, Pange J (2023) Exploring the potential impact of artificial intelligence (AI) on international students in highereducation: generative AI, chatbots, analytics, and international student success. Appl. Sci. 13(16716). https://doi.org/10.3390/app13116716

Ward B, Bhati D, Neha F, Guercio A (2024) Analyzing the Impact of AI Tools on Student Study Habits and Academic Performance. arXiv, Cornell University. https://doi.org/10.48550/arXiv.2412.02166

Forecast-Informed Supply Chain Finance Decisions: Evidence from US-Listed Retail Firms

Xiaotong Sang[1]

Abstract. This paper examines supply chain finance and working capital management among US-listed retail firms. Using quarterly panel data from five sample firms, the study first employs fixed-effects specifications to evaluate the within sample associations between inventory, accounts payable, accounts receivable, cash holdings, firm size, and operating cash flow. Walmart is then used as an illustrative case to compare the reported forecasting performance of Seasonal Naive, ARIMA, Prophet, and XGBoost models. The models are evaluated over the common period from 2024Q1 to 2024Q4 using MAE, RMSE, and MAPE. Within this evaluation period, Prophet produces the lowest reported forecasting errors. The resulting operating cash flow forecast and forecast interval are subsequently incorporated into a goal programming decision analysis covering financing volume, payment term adjustment, and the supplier prepayment ratio under alternative demand, interest rate, financing limit, and cash flow conditions. The empirical results indicate that inventory shows the most consistent negative association with operating cash flow in the sample. The scenario results show that financing and payment configurations vary across operating and financing conditions. The contribution of the paper lies in connecting firm-level working capital evidence, short-term cash flow forecasting, and interpretable supply chain finance decision scenarios.

Keywords: Supply chain finance; Working capital management; Operating cash flow; Cash flow forecasting; Scenario analysis

JEL Classification: C53, C61, G32, M11

1 Introduction

In recent years, Supply Chain Finance (SCF) has developed from narrowly defined accounts receivable financing arrangements into a comprehensive field of study that combines working capital management, cash flow management, and synergistic supply chain governance. Pfohl and Gomm (2009) conceptualize SCF as the optimization of financial flows within supply chains, while Kouvelis (2023) emphasizes the roles of working capital, payment terms, liquidity, and risk management. Chakuu et al. (2019) further classify the mechanisms, actors, and instruments involved in SCF. From a supply chain perspective, working capital management involves the coordination of payment terms and cash to cash cycles across supply chain members rather than the isolated optimization of an individual firm's liquidity position (Hofmann & Kotzab, 2010).

The retail industry presents a typical case for SCF research. This is due to the characteristics of low profit margins, high inventory levels, frequent restocking, and significant cash flow volatility. Any external disruption leading to interrupted restocking or imbalanced payment arrangements can rapidly translate to inventory, cash flow, and sales pressures. Consequently, for retailers, SCF arrangements may help retailers enhance supply chain stability, maintain inventory levels, stabilize cash flow, and minimize value loss. In the context of online retailing, Chen et al. (2019) find that SCF may strengthen competitive advantage by improving platform coordination, bargaining capacity, and collaboration efficiency.

Existing studies provide substantial evidence on individual SCF instruments, trade credit, reverse factoring, working capital coordination, and firm performance. Peng and Zhou (2019) examine the optimal deployment of working capital in a supplier–retailer supply chain facing uncertain demand. Their analysis focuses on coordinating working capital allocation across supply chain members rather than on using a firm-level cash flow forecast as a direct decision input. Gelsomino et al. (2019) develop an optimization strategy for the concurrent adoption of reverse factoring, inventory financing, and dynamic discounting, with working capital requirements and financing costs serving as key assessment parameters. The present study differs from these contributions in three respects.

[1] VSB - Technical University of Ostrava, Finance, 17. listopadu 2172/15, xiaotong.sang@vsb.cz.

First, it uses a firm-level operating cash flow forecast as an explicit input to the decision stage. Second, it jointly considers financing volume, payment term adjustment, and the supplier prepayment ratio. Third, it evaluates the reported decision configuration under alternative demand, interest rate, financing limit, and cash flow conditions. Forecasting studies in the SCF literature have largely focused on the classification and prediction of credit risk, particularly for small and medium-sized enterprises (e.g., Zhu et al., 2019). By contrast, the accounting forecasting literature has examined statistically based models for multi-step-ahead quarterly corporate cash flow prediction (Lorek & Willinger, 2011). The present study connects this cash-flow forecasting perspective with a subsequent SCF decision analysis. Although prior studies separately examine SCF scheme design, working capital optimization, credit risk forecasting, and quarterly cash flow prediction, less attention has been paid to an practical sequence in which a firm-level operating cash flow forecast is used directly to inform financing volume, payment term adjustment, and supplier prepayment decisions.

Against this background, this paper applies a forecast to decision process to a sample of US-listed retail firms. The framework combines fixed-effects analysis, short-term operating cash flow forecasting, and scenario analysis. Firstly, it provides evidence of the relationship between working capital composition and operating cash flow for five large US-listed retailers. Secondly, it compares the forecasting results for Walmart's quarterly operating cash flows generated by a seasonal naive model, an ARIMA model, a Prophet model, and an XGBoost model. Thirdly, under different operating and financing conditions, it links forecast points and intervals to specific financing arrangements, such as financing scale, payment term adjustments, and supplier prepayment rates. The contribution of this paper is to integrate established methods into a forecast to decision process that uses operating cash flow forecasts to support firm-level supply chain finance decisions. The paper aims to examine supply chain finance decisions among US-listed retail firms using a forecast to decision process. Based on this, the paper explores the following research questions:

RQ1: How are working capital variables related to operating cash flow among the sampled US-listed retail firms?

RQ2: How do financing volume, payment term adjustment, and the supplier prepayment ratio vary across alternative operating and financing scenarios?

The structure of this paper is as follows: Section 1 introduces the research background, objectives and research questions; Section 2 describes the research methodology, including fixed-effects models, forecasting models and goal programming model; Section 3 details the data sources, variable construction, and preprocessing procedures; Section 4 reports the empirical results, forecasting results, optimization results, and scenario analysis; Section 5 summarizes the main conclusions, managerial implications, and research limitations.

2 Methodology

This paper employs a research methodology comprising empirical correlation analysis, a comparison of forecasting models, and scenario based decision analysis. Fixed-effects analysis is used to examine the relationship between working capital variables and operating cash flow in the sample. The forecasting stage compares the predictive performance of different models for quarterly Walmart operating cash flows and generates point estimates and forecast intervals. The decision-making stage examines how the allocation of financing and supplier payments changes under different operating and financing scenarios. These three stages are sequentially linked but perform distinct analytical functions.

The empirical stage uses a fixed-effects specification to account for persistent differences across firms and common quarterly conditions. The fixed-effects model is specified as follows:

$$Y_{it} = \alpha + \beta X_{it} + \gamma C_{it} + \mu_i + \lambda_t + \varepsilon_{it} \tag{1}$$

where Y_{it} represents the dependent variable for firm i in quarter t, X_{it} represents the core explanatory variable, C_{it} represents the control variable, μ_i represents the firm specific fixed effect, λ_t represents the time specific fixed effect, ε_{it} is the random error term.

The forecasting models are evaluated over the period from 2024Q1 to 2024Q4. The Seasonal Naive Model, ARIMA model, Prophet model and XGBoost model were compared using the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE). The model with the smallest forecasting error during this period was retained as input for subsequent decision analysis. Seasonal Naive is used as a transparent seasonal benchmark, while ARIMA represents a conventional statistical time series approach (Hyndman & Athanasopoulos, 2021). Prophet models the series through decomposable trend and seasonal

components (Taylor & Letham, 2018), whereas XGBoost provides a tree based nonlinear predictive approach (Chen & Guestrin, 2016). Forecast performance is evaluated using MAE, RMSE, and MAPE, which summarize absolute, squared, and percentage forecast errors, respectively (Hyndman & Koehler, 2006). (The main formulas of these method are shown in the Appendix 1.)

In the final stage, a goal programming model is employed to minimize the standardized adverse deviation between indicators such as financing costs, cash conversion cycle, working capital requirements and liquidity, and the target objectives. The final stage adopts a weighted goal programming formulation in which desired performance levels are represented by target equations and positive or negative deviation variables, while the assigned weights reflect the relative importance of deviations from the specified targets (Tamiz et al., 1998). For each scenario s, the decision variables are financing volume F_s, payment term adjustment ΔT_s, and the supplier prepayment ratio p_s. The model minimizes the normalized undesirable deviations from the financing cost (FC), cash conversion cycle (CCC), working capital requirement (WCR), and liquidity targets (L). The objective function can be written as follows:

$$\min Z_s = \frac{w_1}{FC^*} \times d^+_{FC,s} + \frac{w_2}{CCC^*} \times d^+_{CCC,s} + \frac{w_3}{WCR^*} \times d^+_{WCR,s} + \frac{w_4}{L^*} \times d^-_{L,s} \tag{2}$$

where the deviation terms measure financing cost above its target, the cash conversion cycle above its target, working capital requirements above their target, and liquidity below its target. The deviations are divided by their target values before weighting, so Z_s is a dimensionless normalized objective value. $w_{1\ldots4}$ means weights which represent the importance of each target.

The financial relationships used in the model are defined as follows:

$$FC_s = r_s F_s \tag{3}$$

$$CCC_s = DSO_s + DIO_s - (DPO_0 + \Delta T_s) \tag{4}$$

$$WCR_s = AR_s + INV_s - \frac{COGS_s}{N_s} \times (DPO_0 + \Delta T_s) \tag{5}$$

$$L_s = Cash_0 + OCF_s + F_s + \frac{COGS_s}{N_s} \Delta T_s - r_s F_s - p_s P_s \tag{6}$$

where r_s is the financing rate, DSO_s and DIO_s are the forecast collection and inventory periods, DPO_0 is the initial payment period, N_s is the number of days in the forecast quarter, P_s is projected purchasing expenditure, and L_s is available liquidity. Supplier prepayment is treated as a current cash outflow, whereas payment term extension provides a temporary liquidity buffer.

Under this objective function and financial relationships, the constraints can be written as follows:

$$0 \le F_s \le F_s^{max}, 0 \le \Delta T_s \le \Delta T_s^{max}, 0 \le p_s \le p_s^{max} \tag{7}$$

$$L_s \ge L^{min}, \frac{INV_s + p_s P_s}{(COGS_s / N_s)} \ge CD^{min} \tag{8}$$

where L^{min} represents the minimum liquidity level required to maintain normal operations. INV_s is the projected inventory value, $p_s P_s$ represents the amount of purchasing expenditure prepaid to suppliers. $(COGS_s/N_s)$ represents the average daily cost of goods sold. CD^{min} denotes the minimum inventory coverage days required by the firm. Therefore, the second constraint requires the projected inventory and committed prepaid purchases to cover at least the minimum number of operating days.

3 Data and Preprocessing

This paper utilizes quarterly financial data from five major US publicly listed retail firms, including Walmart, Target, Costco, Best Buy and Home Depot. The selection of the sample was based primarily on three

considerations. Firstly, these firms have relatively comprehensive and consistent public financial disclosures, enabling the construction of variables such as operating cash flow, inventory, accounts receivable, accounts payable, cash holdings, and total assets. Second, the operations of these firms involve inventory replenishment, supplier payments, customer collections, and cash flow management, making them suitable for analyzing supply chain finance and working capital conditions. Third, these firms differ in terms of scale, retail model, and inventory structure, providing a heterogeneous research base for examining the relationship between working capital variables and operating cash flow. Consequently, this paper treats these five firms as a sample of large US publicly listed retail firms of analytical significance.

Corporate financial data were obtained from the SEC system. As the forecast evaluation uses the first to fourth quarters of 2024 (2024Q1–2024Q4) as a uniform testing period, the data sample used for model training and empirical analysis end in 2023Q4. Data processing primarily comprises three steps. Firstly, disclosed financial data is standardized to a quarterly frequency. Flow variables such as operating cash flow and net profit are typically disclosed as cumulative amounts from the beginning of the year to the current period. These are therefore converted into single quarter figures. Stock variables, such as inventories, accounts receivable, accounts payable, cash and cash equivalents, and total assets, are based on end-of-quarter values. Secondly, working capital indicators are constructed using standardized financial variables, including the cash conversion cycle, working capital requirement, inventories, accounts receivable, accounts payable, and cash holdings. Finally, observations are matched by firm and quarter to construct a panel dataset for fixed-effects and heterogeneity analyses. The model incorporates total assets to account for differences in firm size and utilizes firm and time effects to control for relatively stable firm characteristics and the common quarterly environment.

Walmart was selected as the illustrative case for the forecasting and scenario analysis stages. It has the largest operating scale in the sample, continuous quarterly observations, substantial inventory, and supplier-payment activities, and pronounced variation in operating cash flow. These features make Walmart suitable for demonstrating how short-term operating cash flow forecasts and forecast intervals can be connected to case specific decisions concerning financing volume, payment term adjustment, and the supplier prepayment ratio.

4 Empirical Results

This section reports the results of the three stage "forecast to decision" process. Section 4.1 presents the associations between working capital variables and operating cash flow. Section 4.2 compares alternative forecasting models for quarterly operating cash flow of Walmart. Section 4.3 then uses the point forecast and forecast interval to examine financing and supplier payment configurations under alternative scenarios.

4.1 Panel Regression Results

The basic model includes inventory, accounts payable and total assets, and is used to examine the key relationships between operating cash flow, investment in inventory, supplier credit and firm size. The extended model additionally incorporates accounts receivable and cash and cash equivalents. These variables have been added to account for customer payment terms and available cash holdings. Both models use operating cash flow as the dependent variable and include firm-level fixed effects and quarterly fixed effects. By comparing the coefficients of the two models, it is possible to assess whether the estimated relationships for inventory and accounts payable remain consistent after incorporating additional working capital variables, thereby testing their robustness.

First, with operating cash flow as the dependent variable, we examine the relationship between working capital variables and cash flow performance. The basic model is defined as follows:

$$OCF_{it} = \alpha + \beta_1 Inventory_{it} + \beta_2 AP_{it} + \beta_3 Total\ Asset_{it} + u_i + \lambda_t + \varepsilon_{it} \tag{9}$$

and the extended model is defined as follows:

$$OCF_{it} = \alpha + \beta_1 Inventory_{it} + \beta_2 AP_{it} + \beta_3 AR_{it} + \beta_4 Cash_{it} + \beta_5 Total\ Asset_{it} + u_i + \lambda_t + \varepsilon_{it} \tag{10}$$

where, OCF_{it} means operating cash flow of firm i in quarter t, $Inventory_{it}$ means inventory, AP_{it} means accounts payable, AR_{it} means accounts receivable, $Cash_{it}$ means cash and cash equivalents, $Total\ Asset_{it}$ means total assets, α means common intercept, u_i and λ_t represent firm specific fixed effects and time specific fixed effects respectively, and ε_{it} is the random error term.

The results indicate that higher inventory levels are associated with lower operating cash flows. These estimates should be understood as conditional relationships. The coefficient is −0.1461 in the basic model and −0.1419 in the extended model, with corresponding robust standard errors of 0.0464 and 0.0318, respectively. These similar estimates suggest that this negative correlation remains robust after accounting for accounts receivable and cash holdings. By contrast, accounts payable, accounts receivable, cash holdings, and total assets did not show statistically significant correlations with operating cash flow. The results therefore suggest that, within the sample, higher inventory levels are therefore associated with lower operating cash flow within the sample. However, these estimates indicate association.

Heterogeneity analysis indicates that the negative correlation between inventory and cash flow is more pronounced among small enterprises and in the low inventory group, with coefficients of −1.400 and −1.040, respectively, both significant at the 1% level. Supplementary results also suggest that inventory and accounts receivable are associated with lower cash holdings, while accounts payable may provide a partial supplier credit buffer. Overall, by identifying inventory as the working capital variable most closely associated with operating cash flow, this study answers RQ1 and provides an empirical basis for subsequent cash flow forecasting and the analysis of scenario-based supply chain finance decision-making.

4.2 Forecasting Results

Panel data analysis primarily identifies the overall relationship and heterogeneity between working capital variables and operating cash flow, while the forecasting and decision analysis sections require a continuous single firm time series as input. Walmart is used as an illustrative case to demonstrate how operating cash flow forecasts can be translated into decisions concerning financing volume, payment term adjustment, and the supplier prepayment ratio. The forecasting analysis first evaluates the competing models and then generates the cash flow inputs used in the decision analysis. Using the period from 2019Q2 to 2023Q4 as the training set and 2024Q1 to 2024Q4 as the test set, we compare Seasonal Naive, ARIMA, Prophet, and XGBoost forecasts of Walmart's operating cash flow. After comparing the reported forecast errors, Prophet is retained and re-estimated using the complete historical series to generate a point forecast and forecast interval for the next quarter. These outputs are then used in the subsequent scenario-based decision analysis. As Seasonal Naive and ARIMA generate identical forecasts over the reported period, they are presented as a single entry.

Quarter / Metric	Actual	Seasonal Naive / ARIMA	Prophet	XGBoost
2024Q1	16.712	13.143	11.418	9.766
2024Q2	4.249	4.633	5.048	1.866
2024Q3	12.108	13.568	12.042	8.360
2024Q4	6.561	0.813	5.307	3.729
MAE	—	2.790	1.853	3.976
RMSE	—	3.466	2.749	4.358
MAPE	—	32.5%	17.5%	42.9%

Table 1 Comparison of OCF Forecasts with the Model (unit: billion USD)

Table 1 shows that Prophet produced the lowest MAE, RMSE, and MAPE over the four-quarter period and was therefore selected as the forecasting input for the subsequent decision analysis. Following re-estimation using the complete historical series, forecast operating cash flow of Prophet is USD 12.083 billion, with a forecast interval of USD 9.142–15.334 billion. The point forecast and forecast interval are used as inputs for the scenario and sensitivity analysis. This result represents an application specific comparison and does not imply that Prophet is generally superior across firms or forecasting settings.

4.3 Optimization Results

The decision model includes three variables: financing volume, payment term adjustment, and the supplier prepayment ratio. Financing volume represents external short-term funding used to support operating liquidity, while payment term adjustment and supplier prepayment alter the timing of cash transfers between the buyer and its suppliers. These variables correspond to the financing instruments, trade credit arrangements, and working capital coordination mechanisms discussed in the SCF literature (Gelsomino et al., 2016; Xu et al., 2018). The

analysis considers liquidity conditions, financing capacity, payment arrangements, and the forecasted operating cash flow.

Using the weighted goal programming model specified in Equations (2) – (8), the Prophet forecast of USD 12.083 billion is adopted as the central operating cash flow input. The forecast interval ranges from USD 9.142 billion to USD 15.334 billion and is used to examine alternative cash flow conditions. The base and constrained settings use the same operating cash flow forecast and reference target weights, while the constrained setting applies tighter upper bounds to financing volume, payment term adjustment, and the supplier prepayment ratio. The reported model results are subsequently analyzed in relation to demand, financing rates, objective weights, financing limits, and forecast scenarios.

Scenario	Forecast OCF	Payment term adjustment	Prepayment ratio	Financing volume	Objective value
Base	12.083	6.25	0	5.833	0
Constrained	12.083	10.00	0	6.000	1.194

Note: OCF and financing volume are expressed in billions of US dollars; payment terms are expressed in days.

Table 2 Optimization Results

Note: *OCF and financing volume are expressed in billions of US dollars; payment terms are expressed in days.*

Table 2 shows that, under the basic scenario, the model reports a payment term adjustment of 6.25 days, a supplier prepayment ratio of zero, and financing volume of USD 5.833 billion. When the decision bounds are tightened, financing volume increases to USD 6.000 billion and payment term adjustment increases to 10 days, while the prepayment ratio remains zero. The objective value increases from 0 to 1.194, indicating that the constrained setting produces a larger combined deviation from the financing cost, cash conversion cycle, working capital requirement, and liquidity targets.

4.4 Scenario Analysis Results

Table 3 shows that the model-based supply chain finance configuration changes in response to alternative scenario assumptions. When demand decreases, financing volume falls to USD 3.500 billion and the supplier prepayment ratio increases to 0.500. When demand reaches the reference or positive levels, financing volume increases to USD 6.000 billion and payment term adjustment remains at 10 days. This pattern indicates that the reported decision variables respond mainly to the difference between lower demand and reference or higher demand conditions.

The interest rate analysis shows that financing volume remains at USD 6.000 billion when the financing rate changes from 4% to 10%, while the objective value increases from 1.062 to 1.482. This indicates that higher financing rates increase the combined deviation from the model targets without materially changing the reported financing volume. The objective weight and financing limit scenarios further show that management priorities and available credit capacity affect both financing volume and payment term adjustment. The forecast interval results indicate that a higher forecast operating cash flow condition is associated with a configuration involving greater financing volume and a longer payment term adjustment. Overall, the results illustrate how forecast information is translated into financing, payment term, and supplier prepayment decisions under the model assumptions.

The higher OCF scenario should not be interpreted as indicating that internally generated cash mechanically causes greater borrowing. Instead, the forecast operating cash flow represents a broader projected operating condition in which business activity and funding requirements change jointly.

Panel	Scenario	Objective value	Payment term adjustment	Prepayment ratio	Financing volume
Demand shock	-0.25	0.075	0	0.500	3.500 bn
	-0.10	0.075	0	0.500	3.500 bn
	0	1.194	10.0	0	6.000 bn
	0.10	3.188	10.0	0	6.000 bn
	0.25	6.179	10.0	0	6.000 bn
Interest rate shock	4%	1.062	10.6	0.125	6.000 bn
	6%	1.194	10.0	0	6.000 bn
	8%	1.338	10.0	0	6.000 bn
	10%	1.482	10.0	0	6.000 bn
Objective weights	Cost priority	0.787	0.833	0	4.167 bn
	OCF priority	1.859	10.0	0	6.000 bn
Financing limit	4.0 bn	1.512	0.625	0.125	4.000 bn
	7.0 bn	1.041	15.0	0	7.000 bn
Forecast interval	OCF = 9.142 bn	0.075	0	0.500	3.500 bn
	OCF = 15.334 bn	4.258	10.0	0	6.000 bn

Table 3 Scenario Analysis

5 Conclusion

This paper applies a forecast to decision framework comprising empirical association analysis, operating cash flow forecasting, goal programming, and scenario analysis to examine supply chain finance and working capital decisions among US-listed retail firms. The study connects firm-level financial evidence with forward-looking forecasts and model-based financing and payment decisions, thereby addressing the two research questions.

Regarding RQ1, the fixed-effects results indicate that working capital variables are associated with operating cash flow, with inventory showing the most consistent negative relationship across the basic and extended specifications. This finding suggests that higher inventory is accompanied by greater working capital commitment and lower operating cash flow within the sample. The heterogeneity results further show that this relationship is more pronounced among smaller firms and firms with relatively lower inventory levels. Accounts payable appears to provide a potential supplier credit buffer by delaying cash outflows, although its direct association with operating cash flow is less statistically consistent. The supplementary channel results are also consistent with inventory and accounts receivable being related to lower cash holdings, while greater inventory requirements are associated with increased reliance on supplier credit. These findings suggest that retail working capital management should be understood as a coordinated combination of inventory control, supplier payment arrangements, and cash flow management rather than as the isolated adjustment of a single financial variable. In the absence of further diagnostics regarding stationarity and unit roots, these findings should be interpreted as correlations specific to the sample in question.

Regarding RQ2, the forecasting results show that Prophet produces the lowest MAE, RMSE, and MAPE over the four quarters of the 2024 period and is therefore used to generate the operating cash flow input for the decision model. The point forecast of USD 12.083 billion and the corresponding forecast interval are translated into financing volume, payment term adjustment, and supplier prepayment decisions through the goal programming model. The base and constrained results show that tighter decision bounds are accompanied by greater financing requirements, longer payment term adjustments, and a higher normalized deviation from the model targets. The sensitivity analysis further indicates that the decision configuration changes with demand conditions, financing rates, target priorities, financing limits, and forecasted operating cash flow. Lower demand and lower cash flow scenarios are associated with reduced financing volume and greater use of supplier prepayment, whereas benchmark and higher demand conditions are associated with larger financing volumes and longer payment terms.

Financing limits and management priorities also materially affect the balance between external funding and supplier payment flexibility.

Overall, the study contributes a practical forecast to decision approach that links working capital evidence, operating cash flow forecasting, and supply chain finance optimization. Its contribution does not lie in introducing a new econometric, forecasting, or optimization algorithm but in demonstrating how established methods can be combined to transform financial forecasts into interpretable financing and payment decisions. For retail managers, the results indicate that financing scale, payment terms, and supplier prepayment should be considered jointly and adjusted according to projected liquidity, operating conditions, financing capacity, and management priorities.

This study has limitations. The sample comprises only five US-listed retail firms. Therefore, the conclusions are primarily applicable to framework validation and cannot be generalized directly to the entire retail industry. Empirical analyses also have diagnostic limitations. No additional panel data stationarity tests or unit root tests were reported in this revision. Consequently, the fixed-effects estimates are used solely to describe conditional correlations within the sample, and no conclusions are drawn regarding long-term relationships or causal effects. The forecasting and optimization analyses use Walmart as a case study. Future research could include additional firms to enable comparative analysis. Furthermore, certain supply chain finance variables cannot be directly observed from public financial statements and must still be approximated using financial indicators. Future research could incorporate more granular transaction data, such as supplier data and actual financing contract information, to enhance the model's explanatory power and practical value.

Acknowledgements

The SGS Research Project SP2026/003 of VSB Technical University of Ostrava for financial support of this work.

References

Chakuu S, Masi D, Godsell J (2019) Exploring the relationship between mechanisms, actors and instruments in supply chain finance: A systematic literature review. International Journal of Production Economics 216: 35–53. https://doi.org/10.1016/j.ijpe.2019.04.013

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785–794. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chen X, Liu C, Li S (2019) The role of supply chain finance in improving the competitive advantage of online retailing enterprises. Electronic Commerce Research and Applications 33: Article 100821. https://doi.org/10.1016/j.elerap.2018.100821

Gelsomino L M, de Boer R, Steeman M, Perego A (2019) An optimisation strategy for concurrent supply chain finance schemes. Journal of Purchasing and Supply Management 25(2): 185–196. https://doi.org/10.1016/j.pursup.2018.07.004

Gelsomino L M, Mangiaracina R, Perego A, Tumino A (2016) Supply chain finance: A literature review. International Journal of Physical Distribution & Logistics Management 46(4): 348–366. https://doi.org/10.1108/IJPDLM-08-2014-0173

Hofmann E, Kotzab H (2010) A supply chain-oriented approach of working capital management. Journal of Business Logistics 31(2): 305–330. https://doi.org/10.1002/j.2158-1592.2010.tb00154.x

Hyndman R J, Athanasopoulos G (2021) Forecasting: Principles and practice (3rd ed.). OTexts. https://otexts.com/fpp3/

Hyndman R J, Koehler A B (2006) Another look at measures of forecast accuracy. International Journal of Forecasting 22(4): 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Kouvelis P (2023). OM Forum—Supply chain finance redefined: A supply chain-centric viewpoint of working capital, hedging, and risk management. Manufacturing & Service Operations Management 25(6): 2074–2084. https://doi.org/10.1287/msom.2022.0606

Lorek K S, Willinger G L (2011) Multi-step-ahead quarterly cash-flow prediction models. Accounting Horizons 25(1): 71–86. https://doi.org/10.2308/acch.2011.25.1.71

Peng J, Zhou Z (2019) Working capital optimization in a supply chain perspective. European Journal of Operational Research 277(3): 846–856. https://doi.org/10.1016/j.ejor.2019.03.022

Pfohl H-C, Gomm M (2009) Supply chain finance: Optimizing financial flows in supply chains. Logistics Research 1(3–4): 149–161. https://doi.org/10.1007/s12159-009-0020-y

Tamiz M, Jones D, Romero C (1998) Goal programming for decision making: An overview of the current state-of-the-art. European Journal of Operational Research 111(3): 569–581. https://doi.org/10.1016/S0377-2217(97)00317-2

Taylor S J, Letham B (2018) Forecasting at scale. The American Statistician 72(1): 37–45. https://doi.org/10.1080/00031305.2017.1380080

Xu X, Chen X, Jia F, Brown S, Gong Y, Xu Y (2018) Supply chain finance: A systematic literature review and bibliometric analysis. International Journal of Production Economics 204: 160–173. https://doi.org/10.1016/j.ijpe.2018.08.003

Zhu Y, Zhou L, Xie C, Wang G-J, Nguyen T V (2019) Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. International Journal of Production Economics 211: 22–33. https://doi.org/10.1016/j.ijpe.2019.01.032

Appendix 1 Formulations

Model	Formulations	
Seasonal Naive model	$\hat{y}_t = y_{t-4}$	(11)
ARIMA model	$\varphi(B)(1-B)^d y_t = c + \theta(B)\varepsilon_t$	(12)
Prophet	$y_t = g(t) + s(t) + \varepsilon_t$	(13)
XGBoost	$\hat{y}_t = \sum_{m=1}^{M} f_m(X_t)$	(14)

where y_t denotes the observed operating cash flow in quarter t, while $\hat{y}_t$ is its forecast value. For the Seasonal Naive model, y_{t-4} is the operating cash flow observed in the corresponding quarter of the previous year. In the ARIMA model, B is the lag operator, d is the order of differencing, $\varphi(B)$ is the autoregressive polynomial, $\theta(B)$ is the moving average polynomial, c is a constant, and ε_t is the error term. In the Prophet model, $g(t)$ represents the trend component and $s(t)$ represents the seasonal component. In the XGBoost model, M is the number of regression trees, f_m is the tree m, and X_t is the vector of forecasting inputs available for quarter t.

Explainable Artificial Intelligence for Dairy Cow Classification: Neural Networks with Permutation Importance and Local Surrogate Explanations

Jan Saro[1], Vojtěch Šimůnek[2]

Abstract. Explainable artificial intelligence (XAI) is important in precision livestock farming because a prediction must be accompanied by evidence that a farm professional can inspect. This paper evaluates an explainable classification workflow for dairy cow performance and milk quality. Multi-layer perceptron (MLP) classifiers are compared with random forest and gradient boosting models on two balanced, controlled proof-of-concept datasets (n = 1,000 each). The workflow combines held-out permutation feature importance for global model inspection with a LIME-style local weighted-ridge surrogate for individual predictions. On a stratified 80/20 test split, MLP accuracy was 93.5% for dairy cow performance and 84.0% for milk quality; the tree-based baselines were competitive or marginally stronger. Global results identified milk yield, somatic cell count and feed efficiency as the most influential dairy-cow variables, and temperature, colour and pH as the leading milk-quality variables. Local surrogate plots show the case-specific evidence used for a selected prediction. The study contributes a reproducible and explicitly limited XAI workflow rather than a claim of operational deployment. It also clarifies that explanation outputs are model-dependent associations, not causal recommendations.

Keywords: Explainable AI, permutation feature importance, local surrogate explanations, neural networks, dairy farming, precision livestock farming, machine learning

JEL Classification: C45, C63, Q12

1 Introduction

Precision livestock farming increasingly combines animal-level observations, farm records and machine learning to support production, health and welfare decisions (Lokhorst et al, 2019; Mahato and Neethirajan, 2025; Cabrera, 2025). Predictive performance alone, however, is insufficient for adoption when users need to assess whether a model relies on plausible information, understand an individual alert, and identify circumstances in which a prediction should not be trusted. This need motivates explainable artificial intelligence (XAI), which provides post-hoc descriptions of the relationship between model inputs and outputs (Lundberg and Lee, 2017; Ribeiro et al, 2016; Salih et al, 2025).

Dairy applications have demonstrated the value of machine learning for production and health-related prediction, including somatic-cell-count monitoring (Grzesiak et al, 2025; Nagy et al, 2025). Nevertheless, many reports emphasize predictive accuracy while giving limited attention to a consistent explanation workflow that combines global feature screening, local case-level reasoning, and explicit modelling limitations. The research gap addressed here is therefore application-oriented: how to evaluate an opaque classifier on dairy-related tabular data while providing explanations that are transparent about their scope and uncertainty.

This paper makes three contributions. First, it compares an MLP with random forest and gradient boosting classifiers using a leak-free, stratified evaluation protocol. Second, it combines held-out permutation feature importance with local weighted-ridge surrogate explanations. Third, it distinguishes interpretability from explainability and describes limits of both the models and the experimental data. The study is a reproducible proof of concept, not a validation on operational farm data.

[1] Czech University of Life Sciences Prague, Faculty of Economics and Management, Department of Systems Engineering, Kamýcká 129, 165 00 Prague–Suchdol, Czech Republic, saroj@pef.czu.cz

[2] Czech University of Life Sciences Prague, Faculty of Economics and Management, Department of Systems Engineering, Kamýcká 129, 165 00 Prague–Suchdol, Czech Republic, simunekvojtech@pef.czu.cz

2 Related Work and Conceptual Framing

XAI methods are commonly grouped by their scope. Global methods summarize which variables a fitted model relies on across a dataset, whereas local methods approximate or attribute the prediction for one observation. SHAP is an additive attribution framework based on Shapley values (Lundberg and Lee, 2017); LIME fits an interpretable model in the neighbourhood of a selected case (Ribeiro et al, 2016). Both can be useful for tabular models, but their conclusions depend on the fitted model, background data, perturbation design, and correlations between input features (Salih et al, 2025).

The present empirical workflow uses permutation feature importance (PFI), not computed SHAP values. PFI measures the decrease in a selected performance score after the values of one feature are randomly permuted (Fisher et al, 2019). It is model-agnostic and practical for global inspection but can underestimate the importance of correlated variables because information may be shared among them. The local method is LIME-style: a weighted ridge-regression surrogate is fitted around a selected observation. It should be interpreted as a local approximation of the classifier, rather than as a causal model of dairy production or milk quality.

For precision livestock farming, this distinction is substantive. A transparent system must show what the model used, but it must not convert an association in a fitted model into a management prescription without biological validation. The proposed workflow is intended to support review by a domain expert, followed by verification with herd records and clinical or laboratory evidence where appropriate.

3 Materials and Methods

This section describes the datasets, evaluation protocol, predictive models and explanation procedures. All preprocessing operations that estimate data-dependent quantities were fitted within the training data only.

3.1 Datasets and Evaluation Protocol

Two balanced datasets were used (Table 1). The dairy-cow dataset contains synthetic but domain-informed records of Holstein–Friesian cows. Class-related distributions were specified for production and physiological variables and then clipped to plausible bounds. The milk-quality dataset follows the structure of a commonly used teaching dataset, with continuous and binary characteristics generated for three quality grades. The use of 1,000 records per task was deliberate: it enables a controlled and class-balanced proof of concept (approximately 333 observations per class) while keeping the experiments reproducible. It does *not* establish population-level performance and cannot replace evaluation on multi-farm operational data.

Dataset	n	**Features**	**Target and input variables**
Dairy cow performance	1,000	8	Low/medium/high performance (334/333/333). Inputs: milk yield, fat and protein content, lactation days, somatic cell count, body weight, age and feed efficiency.
Milk quality	1,000	7	Low/medium/high quality (334/333/333). Inputs: pH, temperature, taste, odour, fat, turbidity and colour.

Table 1 Datasets used in the proof-of-concept evaluation.

For each task, data were divided into stratified training and held-out test sets (80/20, random state 42). Standardization was fitted on the training set and then applied to the corresponding test set. A shuffled stratified five-fold cross-validation was additionally used to characterize the stability of the MLP result. Accuracy and macro-averaged F1 were selected because all classes were balanced.

3.2 Prediction Models

The primary model was a multi-layer perceptron (MLP) classifier. The dairy-cow MLP used hidden layers of 64, 32 and 16 neurons; the milk-quality MLP used 32, 16 and 8 neurons. Both used ReLU activations, Adam optimization, a maximum of 500 iterations, and early stopping. When early stopping was enabled, 10% of the training subset was reserved internally for validation and training stopped after ten consecutive non-improving iterations. These compact architectures were selected to match the small number of tabular features and to avoid presenting depth as a substantive contribution.

Random forest (100 trees) and gradient boosting (100 estimators, learning rate 0.1, maximum depth 3) served as nonlinear baselines. They were evaluated on exactly the same held-out split as the MLP. This comparison tests

whether the explanatory workflow remains useful even when a neural network is not the most accurate model.

3.3 Explanation Workflow

For global explanation, PFI was calculated on the held-out test set with $K = 30$ random permutations per feature:

$$I_j = \frac{1}{K} \sum_{k=1}^{K} \left(s_0 - s_{j,k} \right), \tag{1}$$

where s_0 is the unpermuted test accuracy and $s_{j,k}$ is accuracy after permuting feature j in repetition k. Error bars report the standard deviation across repetitions.

For a local explanation, $n = 1{,}000$ perturbations were sampled around a test instance x^* on the original feature scale and clipped to the observed training range. Distances were then calculated after standardization. For the predicted class c^*, a weighted ridge surrogate was fitted as

$$\hat{g}_{x^*} = \arg\min_{\beta_0, \beta} \sum_{i=1}^{n} w_i \left[p(c^* \mid z_i) - \left(\beta_0 + \beta^\top \tilde{z}_i \right) \right]^2 + \lambda \|\beta\|_2^2, \tag{2}$$

where $p(c^* \mid z_i)$ is the MLP probability for the predicted class, $\tilde{z}_i$ is the standardized perturbation, and $w_i = \exp[-\|\tilde{z}_i - \tilde{x}^*\|_2^2 / \kappa^2]$ with $\kappa = 0.75\sqrt{p}$ for p features. Displayed contributions are $\beta_j \tilde{x}_j^*$ and therefore describe the surrogate's local score relative to its standardized origin.

4 Results and Interpretation

The results are reported first for predictive performance and then for global and local explanations. This separation is important: an explanation is useful only in relation to a model whose predictive behaviour has been evaluated.

4.1 Predictive Performance

Table 2 shows that all models performed strongly on the controlled dairy-cow task. Gradient boosting gave the highest dairy-cow accuracy (96.5%), whereas random forest gave the strongest milk-quality accuracy (85.5%). The MLP remained competitive but was not the leading model in this corrected, training-only preprocessing protocol. Its five-fold cross-validation accuracy was 0.922 ± 0.016 for dairy-cow classification and 0.837 ± 0.029 for milk-quality classification (mean ± standard deviation). These results reinforce that model selection should be empirical rather than based on model family alone.

Model	**Dairy cow**		**Milk quality**	
	Accuracy	Macro F1	Accuracy	Macro F1
MLP neural network	0.935	0.934	0.840	0.842
Random forest	0.960	0.960	**0.855**	**0.857**
Gradient boosting	**0.965**	**0.965**	0.850	0.852

Table 2 Held-out classification performance (stratified 80/20 split). Bold denotes the best score within a dataset.

4.2 Global and Local Explanations

Figure 1 identifies milk yield (0.182), somatic cell count (0.167), and feed efficiency (0.068) as the leading PFI variables for the dairy-cow MLP. For milk quality, the leading variables were temperature (0.160), colour (0.093), and pH (0.076). These rankings are consistent with the way the controlled data-generating process separates classes, but they must not be interpreted as universal biological effect sizes. In particular, PFI distributes importance imperfectly when predictors are correlated (Fisher et al, 2019; Salih et al, 2025).

The corrected dairy-cow confusion matrix contains 187 correct classifications out of 200. Most errors occur for observations near the low/medium boundary (nine medium observations predicted as low). For milk quality, 168 of 200 observations are correctly classified; the most common errors are between low and medium grades. This pattern supports a cautious interpretation: adjacent, discretized categories are inherently more difficult to separate than extreme cases.

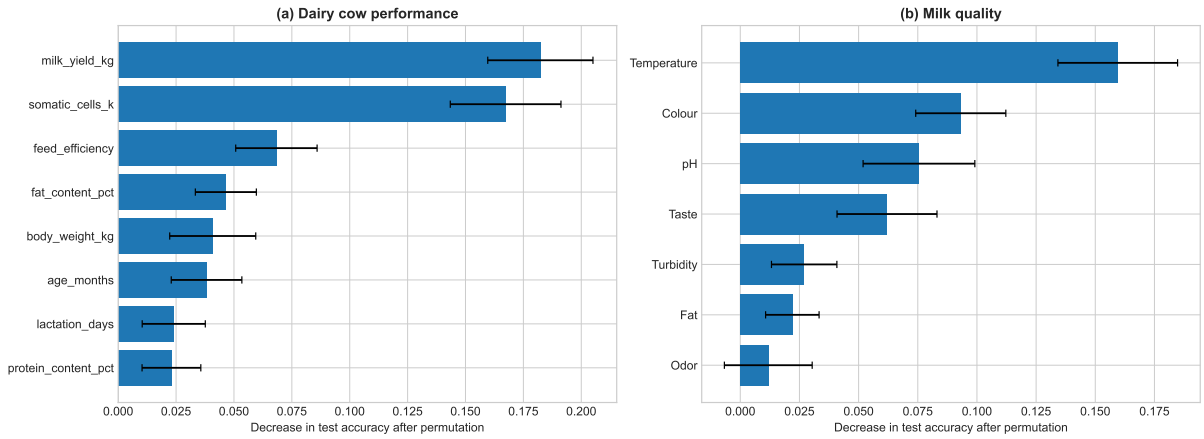


Figure 1 Held-out permutation feature importance for the MLP classifiers. Error bars show the standard deviation across 30 permutations.

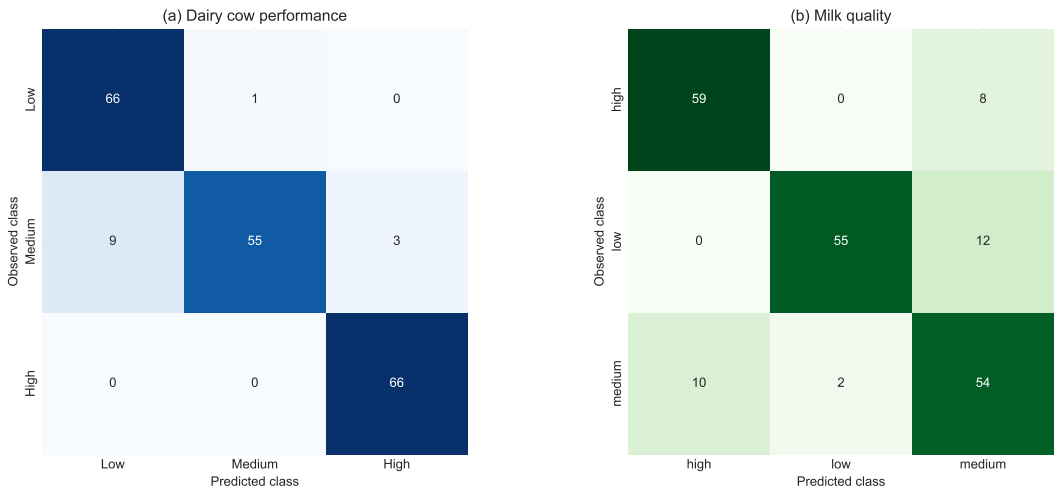


Figure 2 Confusion matrices for the MLP classifiers using the held-out test sets.

Figure 3 gives one correctly classified low-performance dairy-cow case and one correctly classified high-quality milk case. Positive bars increase the local surrogate score for the predicted class; negative bars decrease it. The plots do not state that changing one variable will cause a different class, because the bars reflect a fitted local approximation and not an intervention. Their value for a decision-support workflow is diagnostic: they identify inputs that a farm expert may want to verify before acting on a model output.

5 Limitations and Future Work

The main limitation is data realism. Both datasets are synthetic and balanced; they omit missingness, measurement error, herd effects, management differences, seasonality, temporal dependence, class imbalance and label uncertainty. The sample size of 1,000 records per task is appropriate only for the present controlled demonstration. Larger operational datasets should include multiple farms, repeated observations from the same animals, external temporal validation and pre-specified target definitions.

There are also model and explanation limitations. MLP results can change with architecture, initialization, early-stopping behaviour and hyperparameter selection; the three tested models do not exhaust plausible alternatives. PFI can be unstable or misleading with strongly correlated features, and the local surrogate depends on the perturbation distribution, kernel width and ridge penalty. Neither method supplies causal evidence, uncertainty calibrated for individual decisions, or a guarantee that users will interpret explanations correctly. Future work should compare explanations across model families, quantify explanation stability, include calibrated probabilities and confidence intervals, and co-design explanation displays with veterinarians and farm managers.

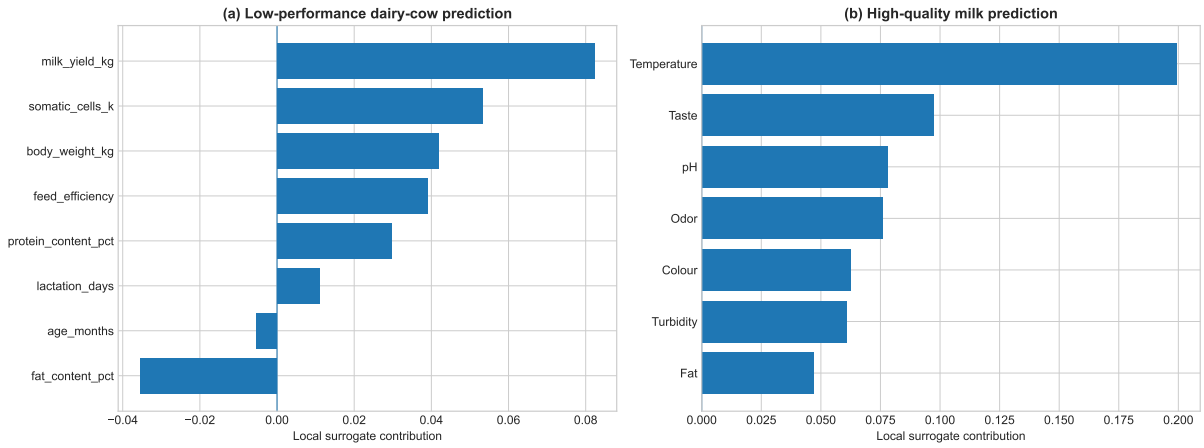


Figure 3 LIME-style local weighted-ridge surrogate contributions for selected correctly classified cases. Contributions are case-specific and non-causal.

6 Conclusion

This study presents a transparent XAI workflow for dairy-related classification that combines out-of-sample evaluation, global permutation importance and local surrogate explanations. Rather than treating interpretability as a visual add-on, the workflow makes its assumptions and limitations explicit. The results show that tree-based models can match or exceed the MLP on these controlled tasks, while the explanation procedure remains applicable to any fitted classifier. Before deployment, the approach requires validation on larger, multi-farm real-world data and expert assessment of whether the explanations are useful, stable and safe for decision support.

Acknowledgements

This research was supported by the Czech University of Life Sciences Prague, Faculty of Economics and Management.

References

Cabrera V E (2025) Artificial intelligence applied to dairy science: Insights from the Dairy Brain Initiative. Animal Frontiers 14(6):60–63. https://doi.org/10.1093/af/vfae040

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 20(177):1–81

Grzesiak W, Zaborski D, Pluciński M, et al (2025) The use of selected machine learning methods in dairy cattle farming: A review. Animals 15(14):2033

Lokhorst C, de Mol R M, Kamphuis C (2019) Invited review: Big Data in precision dairy farming. Journal of Dairy Science 102(10):8774–8787

Lundberg S M, Lee S I (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp 4765–4774

Mahato S, Neethirajan S (2025) Integrating artificial intelligence in dairy farm management - biometric facial recognition for cows. Information Processing in Agriculture 12(3):312–325. https://doi.org/10.1016/j.inpa.2024.10.001

Nagy S Á, Csabai I, Varga T, et al (2025) Neural network-aided milk somatic cell count increase prediction. Veterinary Sciences 12(5):420. https://doi.org/10.3390/vetsci12050420

Ribeiro M T, Singh S, Guestrin C (2016) "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1135–1144, https://doi.org/10.1145/2939672.2939778

Salih A M, Raisi-Estabragh Z, Galazzo I B, et al (2025) A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems 7(1):2400304. https://doi.org/10.1002/aisy.202400304

Radial Basis Functions: Analytical Characterization of Polynomial Augmentation

Vaclav Skala[1]

Abstract. Radial Basis Function (RBF) interpolants approximate an unknown function $f(\mathbf{x})$ via a linear combination of conditionally positive definite kernels, $\varphi(\|\mathbf{x} - \mathbf{c}_i\|)$, augmented by a low-order polynomial term, $\sum_{j=1}^{q} a_j p_j(\mathbf{x})$, to ensure the unique solvability of the interpolation matrix and to control the asymptotic behavior of the estimator in the flat-tail regime. Although existing treatments treat polynomial augmentation as a purely numerical regularization device, we derive closed-form *analytical expressions* for the coefficient vector $\mathbf{a} \in \mathbb{R}^q$, with explicit emphasis on the zero-degree case a_0.

We demonstrate that a_0 is not an artifact of the linear system but rather a global bias-correction term that absorbs the mean of the target function orthogonal to the span of the kernel.

Crucially, our formulation provides an explicit decomposition of the RBF estimator into a parametric polynomial component (global) and a non-parametric kernel residual (local), thereby clarifying the trade-off between polynomial order k and the shape parameter α in determining the estimator's spectral properties and approximation order.

This analytical treatment reframes polynomial augmentation from a mere solvability condition into a structural component of the estimator.

Keywords: Radial Basis Functions, RBF interpolation, Non-parametric interpolation, Polynomial Augmentation, Reproducing Kernel Hilbert Space (RKHS), Multidimensional scattered data

JEL Classification: C63

1 Introduction

We derive a novel closed-form analytical expression for the polynomial coefficient vector within the RBF framework, placing particular emphasis on the zero-degree (constant) term. We demonstrate that this coefficient functions similarly to a global intercept term in a linear regression. However, in the RBF architecture, it acts as a bias-correction mechanism that accounts for mean shifts in the underlying data-generating process. Importantly, there is a direct connection between this zero-degree polynomial coefficient and the sigmoid activation function commonly used in RBF-based neural networks.

By providing explicit formulas for these coefficients—rather than relying on numerical conditioning—our analytical approach enhances the understanding of the bias-variance trade-off inherent in high-dimensional function interpolation. This finding has direct implications for empirical researchers using RBFs in reduced-form estimation, especially in situations that require robust spatial interpolation (such as geographic spillovers), flexible controls in difference-in-differences design, or approximating policy response surfaces in computational structural models. The anticipated applications extend across applied econometrics, spatial statistics, machine learning for causal inference, and numerical solutions for dynamic stochastic general equilibrium (DSGE) models.

2 RBF interpolation

Radial basis function (RBF) methods provide a flexible, mesh-free framework for interpolation and approximation, with broad applicability in numerical and empirical settings Biancolini, 2017, Biancolini, 2017,Zhang et al, 2000. An extensive overview can also be found in Skala, 2026.

1 [0000-0001-8886-4281]

c/o University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Computer Science and Engineering, Univerzitni 8, CZ 301 00 Pilsen

c/o University of Defence, Faculty of Military Technologies, Department of Military Robotics, Kounicova 65, CZ 662 10 Brno, Czech Republic, WEB: www.VaclavSkala.eu

Their defining feature is that the interactions depend solely on pairwise distances between observations in the data domain $\Omega = \{\mathbf{x}_i\}_{i=1}^n$ Fasshauer, 2007, Hardy, 1971, Hardy, 1990, Schaback, 1998, Fasshauer, 2007, Wendland; Wendland, 1995; 2004, Franke, 1979, Iske, 2004, Floater and Iske, 1996, Wu, 1995, Belytschko et al, 1996, Majdisova and Skala, 2017. This distance-based structure makes RBFs particularly suitable for irregularly spaced data.

The interpolant is given by:

$$h(\mathbf{x}) = \sum_{j=1}^{n} \lambda_j \varphi(\|\mathbf{x} - \mathbf{x}_j\|) = \sum_{j=1}^{n} \lambda_j \varphi(r_j) \quad , \quad r_j = \|\mathbf{x} - \mathbf{x}_j\| \quad , \quad \mathbf{x}_j = [{}^1x_j, \ldots, {}^dx_j]^T \quad , \quad \mathbf{x}_j, \mathbf{x} \in \mathbb{R}^d \tag{1}$$

where λ_j are weights, $\varphi(\cdot)$ is a radial kernel and r_j is the distance of the points $\mathbf{x}$ and $\mathbf{x}_j$. Because the argument depends on Euclidean distance, the representation is intrinsically non-separable across dimensions.

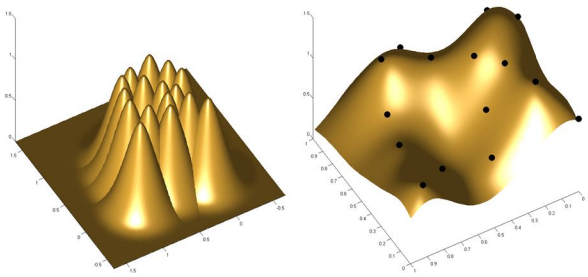


Figure 1 Kernel functions and final RBF interpolation(Smolik and Skala, 2020)

Imposing the interpolation conditions $h(\mathbf{x}_i) = h_i$ yields:

$$h_i = h(\mathbf{x}_i) = \sum_{j=1}^{n} \lambda_j \varphi(\|\mathbf{x}_i - \mathbf{x}_j\|) = \sum_{j=1}^{n} \lambda_j \varphi(r_{ij}), \quad r_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|, \quad i, j = 1, \ldots, n, \quad \mathbf{x}_i \neq \mathbf{x}_j \ \ \forall i / \neq j \tag{2}$$

which can be written compactly as $\mathbf{A}\boldsymbol{\lambda} = \mathbf{h}$, where $\det \mathbf{A} \neq 0$.

RBF kernel functions: there are many RBF kernel functions $\varphi(r)$ which can be split into two main groups:

- "**global**" functions with unlimited influence over the data domain,
  e.g. $\varphi(r) = \frac{1}{2}r^2 \log r^2$, $\varphi(r) = e^{-\alpha r^2}$, $\varphi(r) = \sqrt{1 + \alpha r^2}$, $\varphi(r) = 1/\sqrt{1 + \alpha r^2}$,
  The matrix $\mathbf{A}$ is usually "full" and might be ill-conditioned,
- "**local**" functions with a limited influence only for $r \in [0, 1]$, mostly in the form
  $\varphi(r) = (1 - r)^q P_k(r)$, e.g. $(1 - r)^5(8r^2 + 5r + 1)_+$ "$_+$" means limitation to the interval $[0, 1]$. A shape parameter $\alpha > 0$ is a multiplicative factor to r used to extend/shrink the function influence to $[0, \alpha]$.
  The matrix $\mathbf{A}$ is usually "sparse" (depends on α).
- "**combined**" - Buhmann, 1993 $\varphi(r) = (\frac{1}{3} + r^2 - \frac{4}{3}r^3 + 2r^2 \log r)_+$, Menandro, 2019, etc.

The evaluation of $h(\mathbf{x})$ in any $\mathbf{x}$ follows directly from Eq. 1. The formulation reduces to solving a linear system, allowing the use of standard numerical methods.

Polynomial augmentation is often introduced to improve stability and conditionality Fasshauer, 2007,Flyer et al, 2016, Skala, 2017b, Skala, 2018b, although it can worsen conditioning in the case of large domains Skala, 2017b. Then, with a polynomial $P_k(\mathbf{x})$ the RBF interpolations if given as:

$$h(\mathbf{x}_i) = \sum_{j=1}^{n} \lambda_j \varphi(\|\mathbf{x}_i - \mathbf{x}_j\|) + P_k(\mathbf{x}_i) \quad , \quad i = 1, \ldots, n \tag{3}$$

It should be noted that the use of polynomials can lead to unexpected consequences in the case of a large data domain or a polynomial of higher degree Skala, 2017a, Skala, 2018a.

If the polynomial $P_1(x, y)$ (a bilinear form) is used:

$$P_1(x, y) = a_0 + a_1 x + a_2 y + a_3 xy \tag{4}$$

then the orthogonality conditions are added:

$$\sum_{j=1}^{n} \lambda_j = 0 \quad \sum_{j=1}^{n} \lambda_j x_j = 0 \quad \sum_{j=1}^{n} \lambda_j y_j = 0 \quad \sum_{j=1}^{n} \lambda_j x_j y_j = 0 \tag{5}$$

and the linear system Eq.2 can be expressed in the block system as:

$$\begin{bmatrix} \mathbf{A} & \mathbf{P} \\ \mathbf{P^T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \lambda \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{h} \\ \mathbf{0} \end{bmatrix} \quad , \quad \left[\begin{array}{cccc|cccc} \varphi_{11} & \varphi_{12} & \dots & \varphi_{1n} & 1 & x_1 & y_1 & x_1 y_1 \\ \varphi_{21} & \varphi_{22} & \dots & \varphi_{2n} & 1 & x_2 & y_2 & x_n y_n \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \varphi_{n1} & \varphi_{n2} & \dots & \varphi_{nn} & 1 & x_1 y_1 & y_n & x_n y_n \\ \hline 1 & 1 & \dots & 1 & 0 & 0 & 0 & 0 \\ x_1 & x_2 & \dots & x_n & 0 & 0 & 0 & 0 \\ y_1 & y_2 & \dots & y_n & 0 & 0 & 0 & 0 \\ x_1 y_1 & x_2 y_2 & \dots & x_n y_n & 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} c_1 \\ c_2 \\ \vdots \\ c_n \\ \hline a_0 \\ a_1 \\ a_2 \\ a_3 \end{array}\right] = \left[\begin{array}{c} h_1 \\ h_2 \\ \vdots \\ h_n \\ \hline 0 \\ 0 \\ 0 \\ 0 \end{array}\right] \tag{6}$$

While polynomial terms can enhance numerical behavior, they introduce scale dependence and may reduce robustness over large domains.

Distances are defined for space or space-time as:

$$r_{ij} = \sqrt{\sum_{k=1}^{d} ({}^k x_i - {}^k x_j)^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)} \quad \text{, or} \quad r_{ij} = \sqrt{\sum_{k=1}^{d-1} ({}^k x_i - {}^k x_j)^2 + \beta^2 (t_i - t_j)^2} \tag{7}$$

where ${}^k x$ means the k-st coordinate of the vector $\mathbf{x}$; $t_i \equiv {}^d x_i$.

Squared Normalized RBF (SN-RBF) is defined as:[1]

$$h(\mathbf{x}) = \frac{\sum_{j=1}^{n} \lambda_j \varphi(\|\mathbf{x} - \mathbf{x}_j\|)}{\sqrt{\sum_{j=1}^{n} \varphi^2(\|\mathbf{x} - \mathbf{x}_j\|)}} = \frac{\sum_{j=1}^{n} \lambda_j \varphi(r_j)}{\sqrt{\sum_{j=1}^{n} \varphi^2(r_j)}} \tag{8}$$

which improves conditioning by normalizing the scale of the basis functions.

RBF approximation - the linear system Eq.2 is over-determined. It leads to Eq.9 if a polynomial $P_k(\mathbf{x})$ is used:

$$\begin{bmatrix} \mathbf{A} & \mathbf{P} \end{bmatrix} \begin{bmatrix} \lambda \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{h} \end{bmatrix} \tag{9}$$

where q is the number of the $P_k(\mathbf{x}_i)$ polynomial elements.
Formal application of the Least Squared Error (LSE)[2] leading to Eq.10 **cannot be used** as the matrix elements have different physical units and Lagrange multipliers should be used for a solution Majdisova and Skala, 2016, Skala and Kansa, 2021, Skala, 2016.

$$\begin{bmatrix} \mathbf{A}^T\mathbf{A} & \mathbf{A}^T\mathbf{P} \\ \mathbf{P}^T\mathbf{A} & \mathbf{P}^T\mathbf{P} \end{bmatrix} \begin{bmatrix} \lambda \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^T\mathbf{h} \\ \mathbf{P}^T\mathbf{h} \end{bmatrix} \tag{10}$$

3 RBF polynomial coefficients

In RBF interpolation, the polynomial component $P_k(\mathbf{x})$ captures the global (low-frequency) structure of the data, while the radial basis terms provide localized adjustments. This separation yields a flexible representation that combines a coarse deterministic trend with non-parametric corrections.
Importantly, the coefficients of $P_k(\mathbf{x})$ do not correspond to conventional statistical summaries, e.g., means or moments. Instead, they arise endogenously from the constrained optimization problem underlying RBF interpolation. In particular, when $\varphi(*)$ is conditionally positive definite, these coefficients act as Lagrange multipliers, ensuring the solvability and uniqueness of the system.

[1] The Squared Normalized RBF (SN-RBF) was introduced in Skala and Mourycova, 2023
[2] Using the pseudo-inverse for a solution of an over-determined system $\mathbf{Ax} = \mathbf{b}$ as $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$.

- *Algebraically*, the polynomial coefficients enforce the orthogonality between the RBF component and the space of polynomials of degree $\leq k$, guaranteeing identification.
- *Geometrically*, they determine the global structure of the approximation: a_0 represents a constant offset, whereas higher-order terms capture linear trends and curvature.

In what follows, we characterize the dependence of the polynomial coefficients **a** on the spatial distribution of observations **x** and the associated values **h**.

RBF polynomial - general case

Consider the block system:

$$\begin{bmatrix} \mathbf{A} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{h} \\ \mathbf{0} \end{bmatrix} \quad , \quad \mathbf{M}\,\boldsymbol{\xi} = \boldsymbol{\eta} \tag{11}$$

where **A** is symmetric positive definite (for standard RBFs) and non-singular ($\det(\mathbf{A}) \neq 0$ and $\det(\mathbf{M}) \neq 0$)[3], **P** has full column rank, i.e. $\mathbf{P} \in \mathbb{R}^{n \times q}$, q is the number of polynomial coefficients and $\text{rank}(\mathbf{P}) = q$.
This system implies:

$$\mathbf{A}\boldsymbol{\lambda} + \mathbf{P}\mathbf{a} = \mathbf{h} \quad , \quad \mathbf{P}^T\boldsymbol{\lambda} = \mathbf{0} \tag{12}$$

Solving Eq. 12 for $\boldsymbol{\lambda}$:

$$\mathbf{A}\boldsymbol{\lambda} = \mathbf{h} - \mathbf{P}\mathbf{a} \quad , \quad \boldsymbol{\lambda} = \mathbf{A}^{-1}(\mathbf{h} - \mathbf{P}\mathbf{a}) \tag{13}$$

Substituting $\boldsymbol{\lambda}$ from Eq.13 to $\mathbf{P}^T\boldsymbol{\lambda} = \mathbf{0}$ yields:

$$\mathbf{P}^T\mathbf{A}^{-1}\mathbf{h} - \mathbf{P}^T\mathbf{A}^{-1}\mathbf{P}\mathbf{a} = \mathbf{0} \tag{14}$$

Thus, the polynomial coefficients **a** are as follows:

$$\boxed{\mathbf{a} = \left(\mathbf{P}^T\mathbf{A}^{-1}\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{A}^{-1}\mathbf{h}} \tag{15}$$

The matrix $\mathbf{P}^T\mathbf{A}^{-1}\mathbf{P}$ corresponds to the Schur complement of **A** in **M**.
Eq.15 shows that the polynomial coefficients **a** represent a linear projection of **h** in the polynomial space, with weights determined by $\mathbf{A}^{-1}$, which is determined by the mutual distances of the points $\mathbf{x}_i$ and by the RBF kernel function $\phi(*)$ used.

SN-RBF polynomial - general case

For the squared-normalized RBF (SN-RBF), see Eq.8, the linear system becomes:

$$\boxed{\mathbf{a} = \left(\mathbf{P}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{h}} \tag{16}$$

where $\mathbf{Q} = diag[q_1, \ldots, q_n]$ rescales the observations.

This yields:

$$\mathbf{A}\boldsymbol{\lambda} + \mathbf{Q}\mathbf{P}\mathbf{a} = \mathbf{Q}\mathbf{h} \quad , \quad \mathbf{P}^T\boldsymbol{\lambda} = \mathbf{0} \tag{17}$$

Proceeding analogously:

$$\boldsymbol{\lambda} = \mathbf{A}^{-1}\mathbf{Q}(\mathbf{h} - \mathbf{P}\mathbf{a}) \tag{18}$$

Substitution gives:

$$\mathbf{P}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{h} = \mathbf{P}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{P}\mathbf{a} \tag{19}$$

Hence:

$$\boxed{\mathbf{a} = \left(\mathbf{P}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{h}} \tag{20}$$

Compared to the "standard" RBF case, **Q** introduces a data-dependent weighting, modifying the projection structure.

[3] For some kernel functions $\varphi(*)$, **A** is strictly positive definite.

Zero-oder polynomial

Higher-degree polynomials may introduce an undesirable dependence on coordinate transformations. A zero-degree polynomial $P_0(\mathbf{x}) = a_0$ provides a parsimonious alternative, capturing only a global offset.
A constant polynomial $P_0(\mathbf{x}) = a_0$ often provides a robust baseline:

$$\begin{bmatrix} \mathbf{A} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ a_0 \end{bmatrix} = \begin{bmatrix} \mathbf{h} \\ 0 \end{bmatrix} \tag{21}$$

This specification preserves invariance and provides a parsimonious adjustment to the RBF representation.

RBF interpolation

For $P_0(\mathbf{x}) = a_0$, we obtain:

$$a_0 = \left(\mathbf{1}^T\mathbf{A}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}^T\mathbf{A}^{-1}\mathbf{h} = \frac{\mathbf{1}^T\mathbf{A}^{-1}\mathbf{h}}{\mathbf{1}^T\mathbf{A}^{-1}\mathbf{1}} \quad , \quad \boxed{a_0 = \frac{\sum_{i,j=1}^{n}(\mathbf{A}^{-1})_{ij}h_j}{\sum_{i,j=1}^{n}(\mathbf{A}^{-1})_{ij}} = \frac{\sum_{j=1}^{n} h_j \sum_{i=1}^{n}\alpha_{ij}}{\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_{ij}}} \tag{22}$$

Thus, a_0 is a weighted average of $\mathbf{h}$, with weights determined by $\mathbf{A}^{-1}$.

SN-RBF interpolation

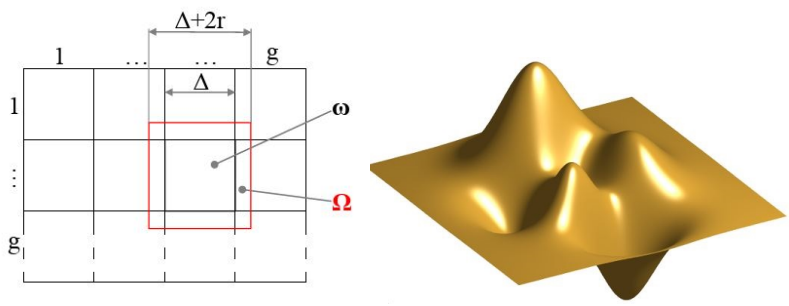
In the SN-RBF case:

$$a_0 = \left(\mathbf{1}^T\mathbf{A}^{-1}\mathbf{Q}\,\mathbf{1}\right)^{-1}\mathbf{1}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{h} = \frac{\mathbf{1}^T\mathbf{A}^{-1}\mathbf{Q}\mathbf{h}}{\mathbf{1}^T\mathbf{A}^{-1}\mathbf{Q}\,\mathbf{1}} \quad , \quad \boxed{a_0 = \frac{\sum_{j=1}^{n} h_j q_j \sum_{i=1}^{n}\alpha_{ij}}{\sum_{j=1}^{n}(q_j \sum_{i=1}^{n}\alpha_{ij})}} \tag{23}$$

This corresponds to a weighted average with observation-specific scaling.

Space subdivision

In the case of large data processing, subdivision of the data domain can be used, see Fig.2(a) Smolik and Skala, 2017. If the compactly Supported RBF (CS-RBF), then the data domain can be split to slightly overlapping cells and interpolation can be computed as ”cell-by-cell” interpolation with additional blending on borders.



(a) Data domain subdivision. (b) Blended interpolation

Figure 2 CS-RBF interpolation - domain subdivision and blending

It leads to significant computational speedup, lower memory demands, and better numerical stability. Fig.2 presents an example of CS-RBF with the data domain subdivision. In the case of Fig.3(a) (in the appendix), the RBF coefficients computation speedup was $\nu \approx 2.4\ 10^5$, and the RBF evaluation speedup was $\nu \approx\ 172$.

4 Conclusion

This paper derives ***closed-form*** analytical formulae for the polynomial $P_k(\mathbf{x})$ in the RBF and SN-RBF interpolation, with particular emphasis on the zero-degree case. The results clarify that the polynomial coefficients, specifically a_0, are not statistical functionals of the data, e.g. mean or median, but equilibrium objects determined jointly by the kernel matrix $\mathbf{A}$, the sampling geometry $\{\mathbf{x}_i\}$, and the target values $\mathbf{h}$.

For the polynomial degree $k = 0$, the interpolant preserves rotational and translational invariance, and a_0 admits a transparent representation as a weighted aggregation induced by $\mathbf{A}^{-1}$ (and $\mathbf{Q}$ in the SN-RBF case). This characterization supports the interpretation of a_0 as a global bias term arising from orthogonality constraints, rather than a descriptive statistic.

The *analytical formulae* have three primary implications.

1. Interpolation with conditionally positive definite kernels makes explicit the role of a_0 in guaranteeing solvability and uniqueness via projection into the polynomial null space.
2. In meshless PDE solvers (Kansa-type methods), the explicit treatment of the constant polynomial improves conditioning, supports boundary enforcement, and informs preconditioning and adaptive strategies
3. For RBF neural networks, they enable direct (non-iterative) determination of the bias term, improve initialization, and facilitate hybrid analytical–learning schemes that separate low- and high-frequency components.

More broadly, the derivations expose a structural link between kernel matrix inversion and projection onto polynomial subspaces, connecting RBF interpolation to RKHS theory and orthogonal decomposition. The resulting closed-form expressions reduce computational overhead by avoiding iterative estimation, which is advantageous in large-scale, online, and streaming settings. Extensions to approximation (non-interpolatory) regimes follow naturally from the same framework.

RBF interpolation and approximation are naturally suitable for scalar and vector data scattered in space-time without the need for tessellation in the data domain. An extensive overview can be found in Skala 2026.

Acknowledgements

The author thanks colleagues and students for their critical comments and constructive suggestions, anonymous reviewers for valuable comments and advice, and also several authors of recently published relevant articles for sharing their views and providing valuable advice.

Funding: This research received no external funding. **Data availability**: No specific data have been used.
AI use disclosure: The core idea, derivations, etc. are entirely the author's work. Large language models (LLMs)(including GPT, Grok, Gemini) were used mainly for language editing and coding assistance.
Conflict of interest: The author declares no conflict of interest.

References

Belytschko T, Krongauz Y, Organ D, et al (1996) Meshless methods: an overview and recent developments. Computer methods in applied mechanics and engineering 139(1-4):3–47. https://doi.org/10.1016/S0045-7825(96)01078-X

Biancolini ME (2017) Fast Radial Basis Functions for Engineering Applications. Springer, https://doi.org/10.1007/978-3-319-75011-8

Buhmann M (1993) On quasi-interpolation with radial basis functions. Journal of Approximation Theory 72(1):103 – 130. https://doi.org/10.1006/jath.1993.1009

Fasshauer G (2007) Meshfree Approximation Methods with Matlab, 1st edn. World Scientific, https://doi.org/10.1142/6437

Floater MS, Iske A (1996) Multistep scattered data interpolation using compactly supported radial basis functions. Journal of Computational and Applied Mathematics 73(1-2):65–78. https://doi.org/10.1016/0377-0427(96)00035-0

Flyer N, Fornberg B, Bayona V, et al (2016) On the role of polynomials in RBF-FD approximations: I. interpolation and accuracy. Journal of Computational Physics 321:21–38. https://doi.org/10.1016/j.jcp.2016.05.026

Franke R (1979) A critical comparison of some methods for interpolation of scattered data. Tech. rep., Naval Postgraduate School Monterey CA

Hardy RL (1971) Multiquadric equations of topography and other irregular surfaces. Journal of geophysical research 76:1905–1915. https://doi.org/10.1029/JB076i008p01905

Hardy RL (1990) Theory and applications of the multiquadric-biharmonic method 20 years of discovery 1968-1988. Computers & Mathematics with Applications 19:163–208. https://doi.org/10.1016/0898-1221(90)90272-L

Iske A (2004) Multiresolution methods in scattered data modelling, vol 37. Springer, https://doi.org/10.1007/978-3-642-18754-4

Majdisova Z, Skala V (2016) A radial basis function approximation for large datasets. In: SIGRAD 2016, pp 9–14

Majdisova Z, Skala V (2017) Big geo data surface approximation using radial basis functions: A comparative study. Computers and Geosciences 109:51–58. https://doi.org/10.1016/j.cageo.2017.08.007

Menandro F (2019) Two new classes of compactly supported radial basisfunctions for approximation of discrete and continuous data. Engineering Reports 2019;1:e12028:1–30. https://doi.org/doi.org/10.1002/eng2.12028

Schaback R (1998) Optimal geometric hermite interpolation of curves, mathematical methods for curves and surfaces. ii. Innov Appl Math pp 417–428. http://num.math.uni-goettingen.de/schaback/research/papers/OGHIoC.pdf

Skala V (2016) Total least square error computation in E2: A new simple, fast and robust algorithm. ACM International Conference Proceeding Series 28-June-01-July-2016:1 – 4. https://doi.org/10.1145/2949035.2949036

Skala V (2017a) Projective geometry, duality and plücker coordinates for geometric computations with determinants on GPUs. ICNAAM 2017 1863. https://doi.org/10.1063/1.4992684

Skala V (2017b) RBF interpolation with CSRBF of large data sets. In: ICCS 2017, Proceedia Computer Science, vol 108. Elsevier, pp 2433–2437, https://doi.org/10.1016/j.procs.2017.05.081

Skala V (2018a) Geometric algebra, extended cross-product and laplace transform for multidimensional dynamical systems. Advances in Intelligent Systems and Computing 661:62 – 75. https://doi.org/10.1007/978-3-319-67618-0_7

Skala V (2018b) RBF interpolation and approximation of large span data sets. In: MCSI 2017 – Corfu. IEEE, pp 212–218, https://doi.org/10.1109/MCSI.2017.44

Skala V (2026) Meshless methods for computer graphics, DSc. Thesis, Czech Academy of Sciences, Institute of Information Theory and Automation (ÚTIA AV ČR)

Skala V, Kansa E (2021) Why is the least square error method dangerous? Computacion y Sistemas 25(1):149–151. https://doi.org/10.13053/CYS-25-1-3473

Skala V, Mourycova E (2023) Meshfree interpolation of multidimensional time-varying scattered data. Computers 12(12). https://doi.org/10.3390/computers12120243

Smolik M, Skala V (2017) Large scattered data interpolation with radial basis functions and space subdivision. Integrated Computer-Aided Engineering 25(1):49 – 62. https://doi.org/10.3233/ICA-170556

Smolik M, Skala V (2019) Efficient simple large scattered 3d vector fields radial basis functions approximation using space subdivision. Lecture Notes in Computer Science 11619 LNCS:337 – 350. https://doi.org/10.1007/978-3-030-24289-3_25

Smolik M, Skala V (2020) Efficient speed-up of radial basis functions approximation and interpolation formula evaluation. Lecture Notes in Computer Science 12249 LNCS:165 – 176. https://doi.org/10.1007/978-3-030-58799-4_12

Wendland H (1995) Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Advances in computational Mathematics 4(1):389–396. https://doi.org/10.1007/BF02123482

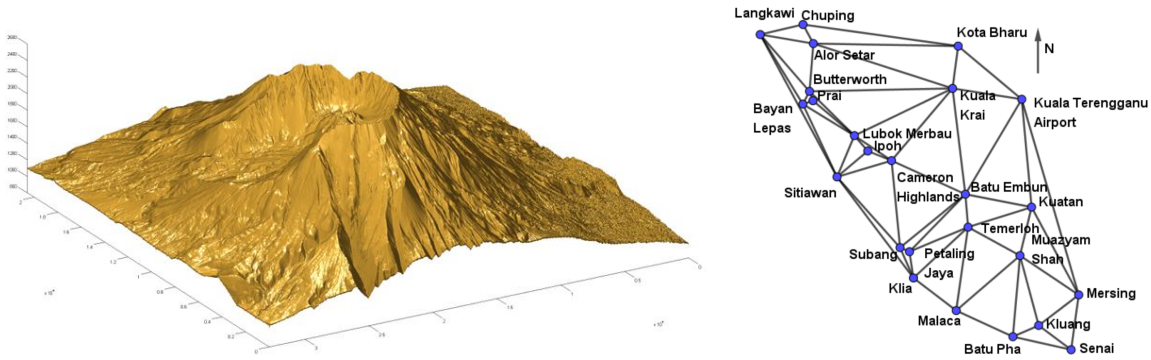
Wendland H (2004) Scattered data approximation, vol 17. Cambridge university press, https://doi.org/10.1017/CBO9780511617539

Wu Z (1995) Compactly supported positive definite radial functions. Advances in computational mathematics 4(1):283–292. https://doi.org/doi.org/10.1007/BF03177517

Zhang X, Song KZ, Lu MW, et al (2000) Meshless methods based on collocation with radial basis functions. Computational mechanics 26:333–343. https://doi.org/10.1007/s004660000181

Appendix Examples of RBF applications

The following examples demonstrate the applicability and flexibility of the RBF interpolation and approximation. Fig.3(a) presents an approximation of a very large GIS dataset of 6.7 mil. points Smolik and Skala 2017 and Fig.3(b) presents the Malaysian rainfall area[4].

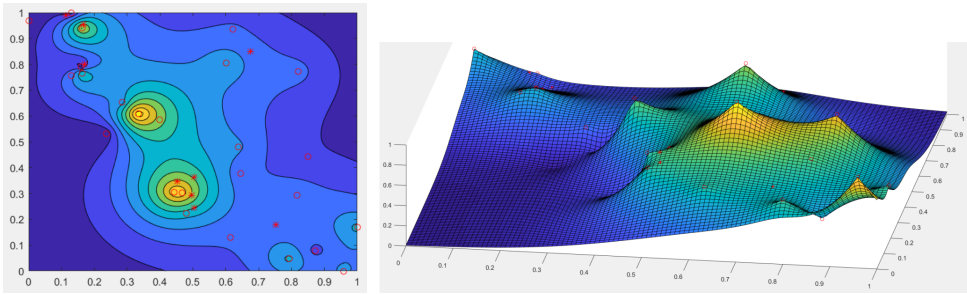


(a) Mount St. Helens; 756 150 points; CS-RBF interpolation with subdivision; original data set 6 743 176 points

(b) Malaysia rainfall area

Figure 3 RBF GIS data interpolation (Smolik and Skala 2017) and Malaysia rainfall area

Fig.4(a) presents a contour plot of the interpolated data, and Fig.4(b) presents the interpolated values.

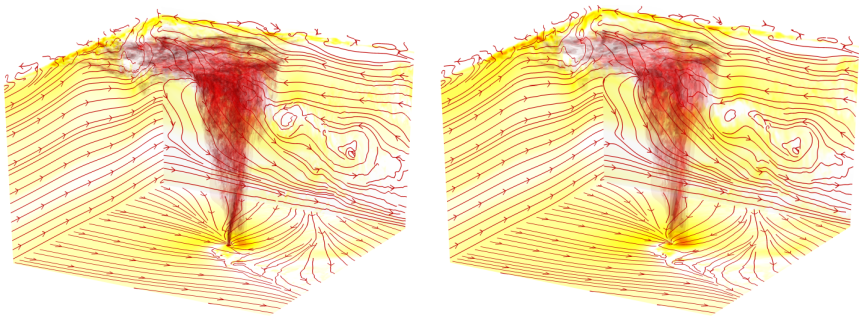


(a) Rainfall data in Peninsular Malaysia.

(b) The interpolation of rainfall dataset.

Figure 4 Malaysia rainfall data in May 2002 (Skala and Mourycova 2023)

Fig.5(a) presents RBF interpolation of a very large vector dataset; fig.5(b) demonstrates efficient compression using RBF approximation.



(a) Tornado 3D original data; 5.5 10^8 points.

(b) Tornado 3D approximated data; approx. error 0.1%, compression rate 1 : 10^3.

Figure 5 RBF Tornado data approximation (Smolik and Skala 2019)

[4] Data taken from: Saaban, A.; Majid, A.A.; Mt Piah, A.R. Visualization of rainfall data distribution Using quintic triangular Bézier patches. Bull. Malays. Math. Sci. Soc. 2009, 32, 137–150

Fractional Economic Dynamics: Deriving Production Functions via Caputo-Type Memory Invariants

Roman G. Smirnov [1]

Abstract. Standard dynamical economic models rely on integer-order differential equations, assuming output depends solely on instantaneous states while ignoring long-term memory effects from policy and technology. This paper extends the exponential growth framework by introducing the Caputo fractional derivative ($0 < \alpha \leq 1$) into the dynamical systems governing factor inputs. This non-local approach models economic trajectories based on the system's entire history, showing that the Mittag-Leffler function emerges as the natural growth solution to generalize the classical exponential model. Finally, we use this framework to derive a new class of time-independent invariants (in a certain sense) as generalized production functions, which converge exactly to the classical Cobb-Douglas function as the fractional order approaches unity.

Keywords: Caputo fractional derivative, production functions in economics, mathematical modeling, economic growth models

JEL Classification: E17, F43, F64, F68, O21, O41, O44, O47

1 Introduction

In their seminal 1928 paper, Charles Cobb and Paul Douglas (Cobb and Douglas, 1928) established a mathematical foundation for modern macroeconomic growth theory. By fitting their namesake production function to historical U.S. manufacturing data spanning from 1899 to 1922, they validated the model's structural integrity through pioneering statistical and econometric analysis. Using ordinary least squares on indices of production, labor, and capital, they demonstrated a robust fit with the functional form: $Y = AL^{0.75}K^{0.25}$, where the scaling parameter was determined to be $A = 1.01$. This specific parametric configuration notably exhibits constant returns to scale, as the exponents (output elasticities) sum exactly to unity ($0.75 + 0.25 = 1$), implying that a proportional increase in all inputs yields an identical proportional increase in total output. Mathematically, this property ensures that the function is homogeneous of degree one, satisfying Euler's theorem for distribution and aligning perfectly with classical competitive equilibrium conditions.

Historically, this work represented a monumental, data-driven departure from the heuristic-based, purely deductive economic modeling that dominated the late 19th century. Instead of deriving market relationships from abstract philosophical tenets, Cobb and Douglas steered the discipline toward rigorous, inductive empirical verification. This methodology contrasted sharply with the fundamental-principles approach typical of theoretical physics, where models are frequently constructed a priori from universal laws rather than being induced directly from observation. Consequently, their framework not only formalized the relationship between factor inputs and aggregate output but also catalyzed the development of modern econometrics by bridging the gap between formal economic theory and statistical measurement. For a century, production functions have underpinned economic growth theory, yet their traditional derivation often lacks uniqueness and robustness. Standard methods fail to determine if alternative functions could fit data equally well or to define the complete set of compatible models. To address this, we have developed a framework — extending Sato's approach (Sato, 1981) across several studies (Smirnov and Wang, 2020, 2021; Smirnov et al, 2022; Smirnov and Wang, 2024; Smirnov, 2025) — that treats production functions as time-independent invariants of dynamical systems. Rather than relying on simple curve-fitting, our approach derives functions directly from systems governed by empirical data and mathematical axioms.

Regarding the exponential model studied, for example, in standard frameworks (Sato, 1981; Smirnov et al, 2022; Smirnov, 2025), that naturally gives rise to the Cobb-Douglas production function, various generalizations exist, most notably the transition to a logistic framework where the exponential model serves as a limiting case (Smirnov and Wang, 2024). However, another promising direction is particularly relevant when the exponential model is applied to economic systems. In this paper, we explore a novel generalization by

[1] Dalhousie University, Department of Mathematics and Statistics, Halifax, 6297 Castine Way PO, BOX 15000 Canada B3H 4R2, Roman.Smirnov@dal.ca

replacing the standard derivative with a Caputo fractional derivative of order $0 < \alpha < 1$. The resulting solution — the Mittag-Leffler function— facilitates a more flexible growth trajectory capable of capturing the "stretched" patterns typical of long-term economic data. While this approach aligns with recent advancements in the "fractionalization" of economic models like the Solow-Swan framework (Cheow et al, 2024; Tarasova and Tarasov, 2017; Traore and Sene, 2020), our work introduces a distinct paradigm shift. Rather than applying fractional calculus to existing macroeconomic equations, we use fractional dynamics to *derive a completely new class of production functions from first principles*. To the best of our knowledge, this marks the first time a production function has been explicitly constructed as an invariant in some sense derived via fractional factor input trajectories. We stress that unlike the classical Solow–Swan framework where the production function is imposed *a priori* as an exogenous structural primitive, our approach derives the generalized Cobb–Douglas function endogenously as a time-independent geometric invariant inherent to the underlying fractional trajectories.

This paper is organized as follows: Section 2 presents the generalization of exponential growth trajectories by introducing the Caputo fractional derivative to capture historical memory effects. Section 3 derives the explicit growth solutions governed by Mittag-Leffler dynamics via the Laplace transform method and provides the economic and mathematical rationale for restricting the fractional order to $0 < \alpha \leq 1$. Section 4 establishes the structural framework for treating production functions as time-independent invariants, highlighting the necessity of algebraic parameter elimination over traditional Lie derivative techniques due to the absence of a semigroup flow. The material presented in this section also synthesizes these trajectories into a novel generalized production function using a weighting parameter, and demonstrates its exact convergence to the classical Cobb-Douglas model as the fractional orders approach unity. Finally, Section 6 concludes the paper and outlines future directions for empirical data calibration.

2 Generalizing the Exponential Model

Before introducing the fractional framework, it is necessary to recall the classical foundation of the underlying growth trajectories. In standard macroeconomic growth theory, under conditions of constant continuous growth rates and perfect competition, the factor inputs of labor and capital, along with aggregate output, are assumed to expand at steady proportional rates over time. Mathematically, this memoryless, instantaneous compounding structure is expressed as a system of first-order linear differential equations, $L'(t) = b_1 L(t)$, $K'(t) = b_2 K(t)$, and $Y'(t) = b_3 Y(t)$, which uniquely yields the classical exponential growth benchmark — see, for instance, (Smirnov and Wang, 2024).

To capture the historical path-dependence and institutional frictions that these integer-order models omit, we extend this standard system by building upon the framework established in (Sato, 1981; Smirnov et al, 2022; Smirnov and Wang, 2024; Smirnov, 2025). We generalize the dynamical systems that yield the classical Cobb-Douglas production function by substituting the local time derivatives with their Caputo fractional counterparts:

$$ {}^{C}D_t^{\alpha_1} L(t) = b_1 L(t), \quad {}^{C}D_t^{\alpha_2} K(t) = b_2 K(t), \quad {}^{C}D_t^{\alpha_3} Y(t) = b_3 Y(t), \tag{1} $$

where $0 < \alpha_i \leq 1$ for $i = 1, 2, 3$. Here, ${}^{C}D_t^{\alpha}$ denotes the Caputo fractional derivative of order α, defined for $n - 1 < \alpha < n$ ($n \in \mathbb{N}$) as:

$$ {}^{C}D_t^{\alpha} f(t) = \frac{1}{\Gamma(n-\alpha)} \int_0^t \frac{f^{(n)}(\tau)}{(t-\tau)^{\alpha+1-n}} d\tau. \tag{2} $$

In this definition, $\Gamma(\cdot)$ is the Gamma function and $f^{(n)}(\tau)$ represents the standard n-th order integer derivative. For the specific interval considered in this paper ($0 < \alpha \leq 1$), the operator simplifies to the case where $n = 1$.

Because the Caputo operator (2) is defined via a convolution integral over the function's past states, it effectively models systems characterized by non-locality or "memory." This mathematical structure is particularly apt for economic frameworks, where current growth rates are often a byproduct of historical policy decisions, infrastructure investments, and cumulative technological adoption (Cheow et al, 2024; Tarasova and Tarasov, 2017; Traore and Sene, 2020).

3 Mittag-Leffler Dynamics and Growth Solutions

To solve the fractional system (1), we apply the Laplace transform method, which naturally incorporates standard initial conditions. Applying this operator directly to each equation yields:

$$ \mathcal{L}\{{}^{C}D_t^{\alpha} f(t); s\} = s^{\alpha} F(s) - \sum_{k=0}^{n-1} s^{\alpha-k-1} f^{(k)}(0). \tag{3} $$

We note that for the specific range of the fractional order considered in this paper ($0 < \alpha \leq 1$), the integer n is defined as the ceiling of α, namely $n = \lceil \alpha \rceil = 1$. Consequently, the operator (2) simplifies to: ${}^{C}D_t^{\alpha} f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{f'(\tau)}{(t-\tau)^{\alpha}} d\tau$. While (3) reduces to: $\mathcal{L}\{{}^{C}D_t^{\alpha} f(t); s\} = s^{\alpha} F(s) - s^{\alpha-1} f(0)$.

Applying this transform to the representative equation ${}^{C}D_t^{\alpha} x(t) = bx(t)$ from our system, we obtain the algebraic relation in the s-domain: $s^{\alpha} X(s) - s^{\alpha-1} x(0) = bX(s)$. Rearranging for $X(s)$ yields:

$$X(s) = x(0) \frac{s^{\alpha-1}}{s^{\alpha} - b}. \tag{4}$$

The inverse Laplace transform of (4) is recognized as the one-parameter Mittag-Leffler function, $E_{\alpha}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k+1)}$. Specifically, the closed-form solutions for our generalized model (1) are:

$$L(t) = L_0 E_{\alpha_1}(b_1 t^{\alpha_1}), \quad K(t) = K_0 E_{\alpha_2}(b_2 t^{\alpha_2}), \quad Y(t) = Y_0 E_{\alpha_3}(b_3 t^{\alpha_3}), \tag{5}$$

where L_0, K_0,Y_0 denote the initial values at $t = 0$. These Mittag-Leffler trajectories generalize the classical exponential growth e^{bt}, introducing a heavier tail and a "stretched" growth profile as α deviates from unity. This transition from local to non-local dynamics provides the necessary flexibility to account for the hereditary effects inherent in long-term economic development (Haubold et al, 2011). We note that the formulas (5) do not define a one-parameter semigroup action (like the classical exponential growth e^{bt}), because the Mittag-Leffler function $E_{\alpha}(bt^{\alpha})$ used in (5) does not satisfy the property $\phi_s(\phi_t\ (t)) = \phi_{s+t}\ (t)$ for $\alpha < 1$, that is $E_{\alpha}(b(t+s)^{\alpha}) \neq E_{\alpha}(bt^{\alpha}) E_{\alpha}(bs^{\alpha})$(Haubold et al, 2011).

Furthermore, within an economic context, restricting the fractional order to $0 < \alpha \leq 1$ is advantageous for several mathematical, empirical, and conceptual reasons. First, the Caputo derivative (2) for $0 < \alpha \leq 1$ (where $n = 1$) characterizes a system in which the current state is a weighted average of its entire history; this is known as the *memory effect*. Setting $\alpha > 1$ (where $n = 2$) would necessitate defining the initial "velocity" of growth ($L'(0)$, $K'(0)$, and $Y'(0)$). While it is economically intuitive that current capital depends on historical investment (memory), it is difficult to justify why the current growth rate should depend on a pre-existing velocity constant established at the start of the observation period.

Second, classical exponential growth models (e^{bt}) assume an instantaneous, memoryless compounding structure that can overestimate the velocity of early-stage structural changes. In contrast, the Mittag-Leffler solution $E_{\alpha}(bt^{\alpha})$ for $0 < \alpha < 1$ offers a highly flexible, non-local growth profile. For short- and medium-term horizons, the fractional order α induces a sub-exponential retardation effect, capturing a "stretched" trajectory where institutional frictions, infrastructure bottlenecks, or slow technology diffusion introduce an initial structural drag.

Crucially, for a positive growth parameter ($b > 0$), the long-term asymptotic behavior of this trajectory along the positive real axis does not exhibit power-law decay. Instead, as $t \to \infty$, it adheres to the well-established asymptotic expansion: $E_{\alpha}(bt^{\alpha}) \sim \frac{1}{\alpha} \exp\left((bt^{\alpha})^{1/\alpha}\right) = \frac{1}{\alpha} \exp\left(b^{1/\alpha} t\right)$. This reveals that while the economy's path is initially constrained by historical memory and friction, it eventually converges to a pure exponential trajectory operating at an accelerated modified rate of $b^{1/\alpha}$. Setting the upper bound strictly at $\alpha = 1$ ensures that this model remains a nested generalization of standard economic theory, collapsing perfectly back into the classical, localized exponential growth benchmark when memory effects are absent.

Additionally, economic variables such as labor (L) and capital (K) must strictly remain positive. Solutions to fractional differential equations with $0 < \alpha \leq 1$ are "completely monotonic," ensuring they behave predictably and remain positive for all t. Conversely, for $\alpha > 1$, the Mittag-Leffler function may exhibit oscillatory behavior. While such oscillations are relevant to business cycle theory, they are generally avoided in fundamental growth models where factor inputs should not exhibit "vibrations" or negative values. Finally, in physics, $\alpha < 1$ represents *sub-diffusion*. Economically, this serves as an apt metaphor for *institutional friction* — such as bureaucratic red tape, slow technological adoption, and market inefficiencies — that prevents an economy from achieving its theoretical exponential potential. More specifically, in physical systems, fractional sub-diffusion describes particle transport that is slower than classical Brownian motion due to complex, crowded environments or trapping sites. In an economic context, this non-local behavior translates to structural path-dependency: a high concentration of institutional "traps" — such as regulatory red tape, lengthy patent approval pipelines, and mismatched labor skills — holds back growth. This ensures that current output is tightly bound to historical bottlenecks, transforming the memoryless exponential benchmark into a stretched, sub-exponential trajectory.

To validate the fractional framework, we fit the generalized trajectories given by (5) directly to the original, exponentiated manufacturing data studied by Cobb & Douglas (Cobb and Douglas, 1928), and spanning from 1899 to 1922. The non-linear optimization of these Mittag-Leffler trajectories yields a striking dichotomy between

the structural dynamics of physical factors and aggregate industrial performance. Most notably, the optimization boundary solution for capital converges exactly to unity ($\alpha_2 = 1.0000$), demonstrating that capital accumulation behaves as a localized, memoryless compounding mechanism characterized by steady, classical exponential growth ($b_2 = 0.0652$). Conversely, the fractional indexes for labor and output drop significantly below the integer baseline ($\alpha_1 = 0.5773$ and $\alpha_3 = 0.6340$, respectively). Rather than reflecting smooth, continuous expansion, these heavily compressed sub-exponential parameters reveal strong hereditary memory effects. This mathematical sub-diffusion captures how labor deployment and aggregate output were profoundly disrupted and shaped by the cyclical economic ups and downs — such as post-WWI industrial rebalancing and localized market panics — that characterized the 1899–1922 period. By shifting the true baseline initial conditions downward ($L_0 = 96.23$ and $Y_0 = 95.33$), the fractional framework effectively integrates this institutional drag and macro-volatility, proving that historical memory plays a fundamental, stabilizing role in long-term production dynamics.

4 Derivation of Fractional Production Function Invariants

In simple terms, an invariant is a permanent economic relationship that remains completely unchanged, even as individual variables change rapidly over time. Think of it as the underlying "economic DNA" or the structural rules of a system. While labor, capital, and total output fluctuate day by day due to market cycles, policy shocks, or technological growth, an invariant is the fixed mathematical surface to which these variables are permanently bound.

To ground this concept for a general quantitative economics audience, consider a standard market system where factor inputs like capital (K) and labor (L) are constantly fluctuating due to business cycles, policy shifts, or localized shocks. While these variables are highly dynamic and time-dependent when viewed individually, economic theory dictates that they do not drift arbitrarily; they are constrained by an underlying technological or structural reality. In a data-driven dynamical systems framework, the production function $F(L, K, Y) = 0$ is precisely this structural reality—it is the time-independent geometric manifold, or "invariant surface," to which the system's trajectories are globally confined.

Deriving the production function as an invariant means we are extracting the permanent, time-free relationship governing production directly from the differential paths of its inputs. In classical frameworks where growth follows a standard local derivative, this invariant surface can be mapped using an infinitesimal generator and its associated Lie derivative. However, because our fractional framework introduces non-local memory via the Caputo operator, the system lacks a one-parameter semigroup flow. Consequently, the invariant cannot be derived using local Lie algebraic techniques. Instead, we must rely strictly on the systematic algebraic elimination of the temporal parameter t. This process strips away the time-dependent "clock" of the economy to uncover the absolute, memory-independent structural DNA linking labor, capital, and aggregate output.

In the framework of data-driven dynamical systems (Sato, 1981; Smirnov, 2025), a production function $F(L, K, Y) = 0$ is viewed as a time-independent invariant of the underlying growth trajectories. Having established the Mittag-Leffler solutions in (5), we now seek to eliminate the temporal parameter t to derive the generalized functional relationship between labor, capital, and output.

Starting from the solution $x(t) = x_0 E_\alpha(bt^\alpha)$, where x represents any of the variables L, K, or Y, we solve for t by utilizing the inverse Mittag-Leffler function, $E_\alpha^{-1}(\cdot)$: $bt^\alpha = E_\alpha^{-1}\left(\frac{x(t)}{x_0}\right) \implies t = \left[\frac{1}{b} E_\alpha^{-1}\left(\frac{x(t)}{x_0}\right)\right]^{1/\alpha}$. Now, solving for the *corresponding* time histories in (5), we arrive at

$$t_L = \left[\frac{1}{b_1} E_{\alpha_1}^{-1}\left(\frac{L}{L_0}\right)\right]^{1/\alpha_1}, t_K = \left[\frac{1}{b_2} E_{\alpha_2}^{-1}\left(\frac{K}{K_0}\right)\right]^{1/\alpha_2}, t_Y = \left[\frac{1}{b_3} E_{\alpha_3}^{-1}\left(\frac{Y}{Y_0}\right)\right]^{1/\alpha_3}. \tag{6}$$

By equating the expression for t along the path where $t = t_K = t_L = t_Y$ for the three trajectories in (5), we obtain the fundamental system of invariants for the fractional model:

$$\left[\frac{1}{b_1} E_{\alpha_1}^{-1}\left(\frac{L}{L_0}\right)\right]^{1/\alpha_1} = \left[\frac{1}{b_2} E_{\alpha_2}^{-1}\left(\frac{K}{K_0}\right)\right]^{1/\alpha_2} = \left[\frac{1}{b_3} E_{\alpha_3}^{-1}\left(\frac{Y}{Y_0}\right)\right]^{1/\alpha_3}. \tag{7}$$

In the absence of a one-parameter group action for the regime $0 < \alpha < 1$, the term "invariant" is utilized here in the *structural and geometric sense*. Specifically, these invariants denote a time-independent functional relationship that remains fixed as the system evolves. While the lack of the semigroup property precludes the use of an infinitesimal generator $\mathbf{V}$ and the associated Lie derivative $\mathcal{L}_{\mathbf{V}}$, the algebraic elimination of the temporal parameter t identifies the unique manifold in the (L, K, Y) state space to which the fractional trajectories are globally confined.

Thus, the resulting production function represents a fixed structural constraint that is independent of the system's memory-dependent "clock". This situation contrasts sharply with the derivation of fundamental invariants within the framework of a regular Lie group action. In the classical setting, one can either employ the method of moving frames — which reduces to purely algebraic elimination of the group parameters — or integrate the infinitesimal generators of the Lie group to determine the identical fundamental invariants. For instance, this dual approach is illustrated by the group-invariant classification of Killing tensors of valence two under the isometry group of the Euclidean plane, achieved via algebraic manipulations within the framework of the moving frames method in (Deely et al, 2004) and through infinitesimal generators in (McLenaghan et al, 2002). In the current fractional regime such a duality is broken: the invariants can only be derived algebraically, introducing inherent limitations to the geometric analysis of the underlying dynamical system.

As we have already mentioned above, the surface defined in the (L, K, Y) space by (7) can be interpreted as the "economic DNA" of the system. While the growth trajectories $(L(t), K(t), Y(t))$ evolve dynamically, this surface represents a fixed, structural relationship that remains invariant. Crucially, these formulas are derived algebraically by eliminating the temporal parameter t. In this fractional context, however, the invariant surface cannot be rederived via the standard Lie derivative approach — where $\mathcal{L}_V(f) = V \cdot \nabla f = 0$ — because the solutions (5) do not constitute a flow in the classical group-theoretical sense. This distinguishes the current approach from previous generalizations of the Cobb-Douglas production function, such as those in (Smirnov and Wang, 2020) and (Smirnov and Wang, 2024); those models relied on local dynamical properties and allowed for the invariant to be established through both the Lie derivative and the algebraic elimination of time.

5 A Generalization of the Cobb-Douglas Production Function

We begin by operationalizing the invariant relations established in Equation (7). While a single chronological time variable t underlies the entire dynamical system, the separate production factors accumulate memory and experience at asymmetric scales governed by their distinct fractional orders ($\alpha_1 \neq \alpha_2$). To construct a production function that handles out-of-equilibrium input configurations where labor and capital may deviate from their strict steady-state paths, we invert the system to define factor-specific historical metrics. Let $t_L^{\alpha_3}$ denote the virtual operational history implied by the observed labor stock, and let $t_K^{\alpha_3}$ denote the corresponding history derived from (6) implied by the capital stock: $t_L^{\alpha_3} = \left[\frac{1}{b_1} E_{\alpha_1}^{-1}\left(\frac{L}{L_0}\right)\right]^{\frac{\alpha_3}{\alpha_1}}$ and $t_K^{\alpha_3} = \left[\frac{1}{b_2} E_{\alpha_2}^{-1}\left(\frac{K}{K_0}\right)\right]^{\frac{\alpha_3}{\alpha_2}}$. In an idealized economic equilibrium, these trajectories coincide such that $t_L^{\alpha_3} = t_K^{\alpha_3} = t^{\alpha_3}$. However, to accommodate independent variations in factor deployment, we follow a structural ansatz by taking a weighted linear combination of these separate historical metrics within the characteristic space of the output function $Y = Y_0 E_{\alpha_3}(b_3 t^{\alpha_3})$. This synthesis yields the generalized production function Y:

$$Y = Y_0 E_{\alpha_3}\left\{b_3\left(\theta\left[\frac{1}{b_1} E_{\alpha_1}^{-1}\left(\frac{L}{L_0}\right)\right]^{\frac{\alpha_3}{\alpha_1}} + (1-\theta)\left[\frac{1}{b_2} E_{\alpha_2}^{-1}\left(\frac{K}{K_0}\right)\right]^{\frac{\alpha_3}{\alpha_2}}\right)\right\}, \tag{8}$$

where the weight parameter $\theta \in (0, 1)$. In the limit as $\alpha_i \to 1$ for all i, this fractional relation converges to the classical Cobb-Douglas family:

$$Y = A L^{\frac{b_3}{b_1}\theta} K^{\frac{b_3}{b_2}(1-\theta)}, \tag{9}$$

in view of $\lim_{\alpha\to 1} E_\alpha^{-1}(z) = \ln(z)$ and $\lim_{\alpha\to 1} E_\alpha(z) = e^z$, where $A = \frac{Y_0}{L_0^{\theta\frac{b_3}{b_1}} K_0^{(1-\theta)\frac{b_3}{b_2}}}$. Notably, by selecting the specific weight $\theta = \frac{b_1(b_2-b_3)}{b_3(b_2-b_1)}$, provided $b_2 > b_3 > b_1$, the exponents satisfy the condition $\theta\frac{b_3}{b_1} + (1-\theta)\frac{b_3}{b_2} = 1$. This particular configuration recovers the standard Cobb-Douglas production function characterized by constant returns to scale. We note that the family of functions (8) is an invariant in the sense that in the limit it yields a true time-independent family of invariants (9) preserved by the exponential model studied in Sato, 1981; Smirnov and Wang, 2021; Smirnov et al, 2022; Smirnov and Wang, 2024.

6 Conclusions

In this paper, we have derived a new generalization of the Cobb-Douglas production function by extending the underlying dynamical systems framework into the fractional regime. By replacing traditional integer-order growth rates with the Caputo fractional derivative of order $0 < \alpha_i \leq 1$ for $i = 1, 2, 3$, we have successfully integrated non-local "memory effects" into the modeling of labor (L), capital (K), and output (Y).

Our derivation demonstrates that the Mittag-Leffler function emerges as the natural growth solution, providing the necessary mathematical flexibility to account for historical policy impacts, institutional frictions, and the sub-exponential "stretched" growth patterns often observed in maturing economies. We have established that this fractional model serves as a nested generalization. Specifically, as the fractional orders approach unity, the derived time-independent invariants (7) and the resulting composite surface (8) converge precisely to the classical Cobb-Douglas functional form. This ensures that the framework remains consistent with established economic theory while offering a more robust lens for analyzing long-term structural relationships.

Acknowledgments

The author is grateful to the organizers of the MME2026 conference, hosted by the Faculty of Economics at VSB - Technical University of Ostrava, for providing a platform to present and discuss these findings; he also wishes to thank the anonymous reviewers whose constructive feedback has helped to improve the presentation.

References

Cheow Y H, Ng K H, Chang P, et al (2024) The application of fractional calculus in economic growth modelling: An approach based on regression analysis. Heliyon 10(15):e35379

Cobb C, Douglas P (1928) A theory of production. American Economic Review 18 (Supplement):139–165

Deely R J, Horwood J T, McLenaghan R G, et al (2004) Theory of Algebraic Invariants of Vector Spaces of Killing Tensors: Methods for Computing the Fundamental Invariants. In: Proceedings of Institute of Mathematics of NAS of Ukraine. Mathematics and its Applications. Institute of Mathematics of NAS of Ukraine, Kyiv, pp 1079–1086

Haubold H J, Mathai A M, Saxena R K (2011) Mittag-Leffler functions and their applications. Journal of Applied Mathematics 2011:298628, 51 pages

McLenaghan R G, Smirnov R G, The D (2002) Group invariant classification of separable Hamiltonian systems in the Euclidean plane and the O(4)-symmetric Yang-Mills theories of Yatsun. Journal of Mathematical Physics 43(3):1422–1440

Sato R (1981) Theory of Technical Change and Economic Invariance. Application of Lie Groups. Academic Press, New York

Smirnov R, Wang K (2020) In search of a new economic model determined by logistic growth. European Journal of Applied Mathematics 31:339–368

Smirnov R, Wang K (2021) The Cobb-Douglas Production Function Revisited. In: The 5th International Conference on Applied Mathematics, Modeling and Computational Science 2019. Conference Proceedings. Springer International Publishing, Cham, pp 725–734

Smirnov R, Wang K (2024) The Cobb-Douglas production function and the old Bowley's law. Symmetry, Integrability, and Geometry: Methods and Applications (SIGMA) 20:045, 20 pages

Smirnov R, Wang K, Wang Z (2022) The Cobb-Douglas Production Function for an Exponential Model. In: Advances in Econometrics, Operational Research, Data Science and Actuarial Studies. Contributions to Economics. Editor Terzioˇglu M.K. Springer International Publishing, Cham, pp 1–12

Smirnov R G (2025) Deriving Production Functions in Economics Through Data-Driven Dynamical Systems. In: Conference Proceedings of the 43rd International Conference on Mathematical Methods in Economics. Editor Hrabec, D. Czech Society of Operational Research, Tomas Bata University in Žlin, Faculty of Management and Economics, pp 291–296

Tarasova V V, Tarasov V E (2017) Economic interpretation of fractional derivatives. Progress in Fractional Differ-entiation and Applications 3(1):1–6

Traore A, Sene N (2020) Model of economic growth in the context of fractional derivative. Alexandria Engineering Journal 59(6):4843 – 4950

Weekday and Daytime Influence on Work Contours in Projects

Tomas Subrt[1], Jan Bartoska[2], Josef Kunhart[3], Irena Strausova[4], Simona Huguet[5]

Abstract. A substantial portion of the authors' prior research has been devoted to analysing the distribution of work effort over the duration of individual activities or groups of activities within projects, leading to the formulation of work effort models. In addition to parameters derived from the technology of specific activities (e.g., Front-Loaded, Back-Loaded, Turtle, Bell etc.) and scheduling constraints derived from the objective (un)availability of resources resulting levelling process, work effort is also significantly influenced by subjective factors.
Among the principal subjective influences are, in line with approaches such as the Critical Chain method generating Student Syndrome and Parkinson's Law, as well as effects stemming from weekly work cycles namely, variations in work performance and willingness to exert effort across different times of day and days of the week.
This research analyses the evolution of work effort as a function of daytime phase (morning vs. afternoon) and weekday (Monday through Friday), in combination with both subjective and objective factors. In addition to the task completion horizon, the temporal dimension of the working week is explicitly considered.
Given that work motivation at the level of the working week relates not to a specific task but rather to engagement with the current project portfolio, the influence of individual task deadlines is assumed to be marginal in this context.
An exploratory study was conducted involving secondary and university students, as well as project professionals, employees, and managers. The resulting model constitutes a comprehensive mathematical representation of the expected work effort of a project worker across a portfolio of tasks, under the constraint of a limited weekly working time capacity.
We assume that the results of the model will be provided to project managers authority to adjust key milestones and deadlines for project tasks and to increase the motivation of project resources.

Keywords: resource work effort, work contour, daytime phases contour model, deadline optimization

JEL Classification: C63, D91

1 Theoretical foundations and connection to previous research

Previous theoretical research in the field of project management proves that the behaviour of the human factor, shaped primarily by "Parkinson's Law" and "Student Syndrome", represents a fundamental source of deviations from original plans (Smith, 2010; Leach, 1999). The tendency to postpone work until deadlines approach is consistent with broader research on procrastination and self-regulation. According to Steel (2007), procrastination represents a pervasive behavioral phenomenon associated with temporal discounting and reduced task motivation. Similar behavioral patterns have also been observed in project environments where deadlines strongly influence effort allocation. The authors base their research on the premise of Parkinson's Law, which describes the phenomenon where work expands to fill all the time available for its completion (Parkinson, 1991). The authors

[1] University of South Bohemia, Faculty of Economics, Department of Applied Mathematics and Informatics, Studentska 787/13 Ceske Budejovice, tsubrt@ef.jcu.cz
[2] Department of Systems Engineering FEM CZU, Kamýcká 129, Prague, bartoska@pef.czu.cz .
[3] Department of Systems Engineering FEM CZU, Kamýcká 129, Prague, kunhart@pef.czu.cz.
[4] University of South Bohemia, Faculty of Economics, Department of Applied Mathematics and Informatics, Studentska 787/13 Ceske Budejovice, istrausova@ef.jcu.cz
[5] Department of Systems Engineering FEM CZU, Kamýcká 129, Prague, huguet@tf.czu.cz

closely follow up on this assumption with a detailed investigation of the Student Syndrome (often associated with procrastination), which characterizes the tendency of workers to postpone tasks until the last possible moment (Bartoška and Šubrt, 2011). Results from empirical studies based on real-world data have shown that worker behaviour can be divided into three phases, with a critical drop in performance in the second third of the task's duration (Bartoška and Šubrt, 2011). To accurately quantify this phenomenon, the authors proposed a mathematical model combining trigonometric and exponential functions (Bartoška and Šubrt, 2011), which they subsequently refined into a 4th-degree polynomial function (Kučera et al., 2014).

The authors directly follow up on their previous research and further integrate these behavioural models with Work Contour models for resource allocation purposes (Kučera et al., 2014). While theoretical planning standardly operates with idealized curves (e.g., Front loaded, Back loaded, Bell, Turtle), the authors demonstrated with their models that the actual course of work effort must be mathematically expressed as a weighted sum of the ideal curve and the degree of the worker's susceptibility to yielding to the Student Syndrome (Kučera et al., 2014; Šubrt et al., 2021). The results clearly show that even when the workload is planned for the beginning of the task, the inclusion of human factor behaviour leads to a radical deformation of the work effort course (Kučera et al., 2014). The authors most recently follow up on the research of these exact contour models in the agile environment of the Scrum methodology, where they propose incorporating the expected work effort course directly into the Planning Poker technique, which enables teams to better manage the variability of the total workload during a sprint and collaborate more effectively (Bartoška et al., 2024).

The practical consequences of these theoretical findings manifest primarily in advanced control mechanisms, such as Earned Value Management (EVM) (Bartoška et al., 2015; Bartoška et al., 2016). In this area as well, the authors follow up on the research of quantifying the Student Syndrome and work contours with the aim of non-linearly modifying the parameter of Budgeted Cost for Work Scheduled (BCWS) (Bartoška et al., 2015). The authors explicitly document that the classic assumption of linear BCWS growth is completely flawed in a real-world environment and leads to falsely optimistic estimates (Bartoška et al., 2016). The implementation of the authors' non-linear models reveals hidden project crises and threats of delay much earlier than traditional methods (Bartoška et al., 2016). Based on a precise understanding of these performance fluctuations, the authors subsequently propose optimized multi-criteria strategies for human resource management and motivation, which consider worker criticality and prevent the accumulation of risks in projects (Bartoška et al., 2018).

2 Work Contours Analytical Functions

The primary objective of this paper is to extend our previous research on the scheduling of project resource workloads. Our prior work focused on various factors influencing the intensity of work performed by project resources on individual tasks or groups of tasks. These influences can be examined in isolation, grouped according to their nature, or understood in a comprehensive framework. All factors affecting work scheduling can be broadly categorized into objective and subjective ones. Objective factors are determined by technology, resource unavailability (not limited to human resources), and temporal constraints. Subjective factors arise from patterns of human behaviour and include the syndromes described by Eliyahu M. Goldratt (Parkinson's Law and the Student Syndrome), as well as the effect of the phase of the working week. It is just this latter factor, and its integration into the broader context of other influences, that constitutes the focus of this paper.

2.1 Backgrounds for Construction of Analytical Functions

In accordance with established approaches, let us assume that the work effort curve is represented by a function $y = f(x)$, where $x$denotes time during which a resource performs work on a task, and y represents the level of effort. Time may be measured either in standard time units or as a percentage of task duration, i.e., the duration of resource allocation. The vertical axis may represent work effort in percentage terms. Both axes may also be normalized to the interval (0, 1).

A. Discrete Approach

In real-world settings, work is assigned to resources in discrete time. That is, over a non-zero interval (t_{i-1}, t_i), work is performed at a constant intensity. The total volume of work performed is thus given by the sum of the areas of rectangles defined by $t_{i-1}, t_i, f(t_i), f(t_{i-1})$.

B. Continuous Approach

Within the continuous framework, the instantaneous work intensity of a resource is expressed as the derivative of the function $f(x)$ at point t_i. The total volume of work performed is then given by the definite integral of $f(x)$ between the start and the end of the resource's allocation to the activity.

C. Principle of Constant Workload

Let us assume that the total volume of work is determined ex ante in the form of task workload and remains invariant under any scheduling of work. Workload is therefore a constant, while task duration (i.e., resource allocation duration) and work intensity are variables. It is evident that the duration of a task (or resource allocation) is minimized under uniform utilization at 100% capacity (i.e., a flat work contour). Furthermore, we assume that work intensity cannot exceed 100% at any time.

2.2 Work Effort and Willingness to Work

Existing research and modelling of work scheduling functions have primarily focused on the intensity of work effort over the entire duration of a task. Work effort is influenced not only by objective constraints but also by intrinsic motivation and individual perceptions of meaningful work. Amabile and Kramer (2011) demonstrate that daily fluctuations in motivation and progress significantly affect employee performance and engagement. Our approach examines work distribution within the working week, regardless of whether the task is in its initiation, execution, or termination phase. The approach is based on the premise that the behaviour of work resources is subjectively influenced by the phase of the working week.

For the purposes of this research, the week was divided into ten phases, i.e. each working day split into morning and afternoon segments (e.g., MoM – Monday morning, MoA – Monday afternoon, etc.). The study investigated individuals' willingness and motivation to work across these phases.

Three groups of respondents were surveyed:

1. secondary school students whose primary occupation is studying,
2. university students, the majority of whom are employed alongside their studies, and
3. employees (excluding managerial positions).

Only two questions were posed: which phase of the working week is the most preferred and associated with the highest work performance, and conversely, which phase is the least preferred and associated with the lowest productivity. It was necessary to clearly distinguish between preference and willingness to work, as illustrated by Friday afternoon. A considerable proportion of respondents identified Friday afternoon as their preferred time (logically), yet simultaneously acknowledged that their work performance during this period is likely to be very low

The survey was conducted as an exploratory pilot study aimed at identifying general patterns of perceived willingness to work across different phases of the working week. Respondents were recruited using convenience sampling. The sample consisted of approximately 60 secondary school students and approximately 70 university students, employees and managers. Participation was voluntary and anonymous. The questionnaire contained two closed questions asking respondents to identify (i) the phase of the working week in which they perceive their highest willingness and productivity and (ii) the phase associated with their lowest willingness and productivity.

The objective of the survey was not to provide statistically representative estimates of workforce behavior, but rather to obtain initial empirical data for the construction of analytical work-contour functions. The resulting model should therefore be interpreted as exploratory and illustrative.

2.3 Survey among Secondary School Students

Students from an economics-oriented secondary school participated in the survey, yielding the following results (see Figure 1). The critical phase, i.e. the period with the lowest potential work performance, concentrates around the second day of the working week. A small peak in Monday afternoon, peak of potential work activity occurs on Wednesday, followed by a moderate decline on Thursday afternoon and a secondary peak on Friday morning, slowly decreasing toward the weekend.

2.4 Survey among University Students and Employees

University students from an economics-oriented faculty and employees (managers) from several branches yielding the following results in graphical form (see Figure 1). Given that most university students are simultaneously engaged in employment alongside their studies, the authors decided to merge these categories into a single analytical output. The results do not differ significantly from those of respondents whose primary occupation is studying. Nevertheless, the critical phase (the period with the lowest potential work performance) again concentrates around the second day of the working week. The peak of potential work activity occurs on Wednesday, followed by a moderate decline on Thursday afternoon and a secondary peak on Friday morning, with a subsequent rapid decrease toward the weekend.

2.5 All Together Approximation

In order to represent the curve of expected work effort, it is possible either to use average values (most preferred and less preferred working time) or to compute differences between individual values and subsequently normalize them. Given the qualitative nature of the question (best–worst) combined with quantitative evaluation (number of respondents), the difference-based approach with normalization to the interval (0, 1), or equivalently (0,100%), appears more appropriate.

To aggregate multiple influences (e.g., the Student Syndrome and Parkinson's Law), it is necessary to approximate the results using a continuous function. Suitable candidates include higher-degree polynomials or spline functions.

A cubic or, more generally, higher-degree polynomial represents a global approximation model in which a single analytical function describes the entire dataset—this being both its principal strength and its main limitation. Its advantages include simplicity of representation, ease of implementation, compactness in the form of a single expression, and relatively good performance for smooth data without pronounced local variations. A cubic polynomial, in particular, often captures the fundamental trend, including a single inflection point. Although a third-degree polynomial provides a simple and computationally convenient approximation for rough orientation, it fails to capture the mid-week maximum (i.e., the peak in willingness to work on Wednesday).

As the degree of the polynomial increases, the approximation capability improves and the model adapts more closely to the data, but at the cost of reduced stability and interpretability. Higher-degree polynomials tend to produce artificial oscillations, especially near the boundaries of the interval, are sensitive to small perturbations in the data, and typically exhibit poor extrapolation properties. An exact ninth-degree polynomial would be highly unstable, and its oscillatory behaviour would not yield meaningful or interpretable values

In contrast to pure polynomials, a cubic spline represents a local model composed of low-degree polynomials defined on subintervals and connected by smoothness conditions. As a result, it interpolates the data exactly, preserves the natural shape without introducing spurious extrema, is numerically stable, and behaves well at the boundaries. Its disadvantage is that it is not described by a single formula but rather by a system of functions, making implementation more demanding without specialized tools.

Cubic splines were selected because they preserve local extrema and avoid artificial oscillations typical of higher-degree polynomial approximations. Since the empirical data exhibit multiple local maxima and minima, spline interpolation provides a more realistic representation of behavioral patterns than a single global polynomial function.

Due to the similar functional patterns observed across both respondent groups, spline-based approximation appears to be the most appropriate method for modelling the resulting work effort profile (See Figure 1). Spline functions for 9 intervals for All Together Model – see Formula (1).

$$
\begin{aligned}
&\boldsymbol{Interval}\ \langle 0,1\rangle: S_0(x) = -13.8182x^3 + 34.9547x^2 - 8.1364x + 12\\
&\boldsymbol{Interval}\ \langle 1,2\rangle: S_1(x) = -13.8182(x-1)^3 - 6.5(x-1)^2 + 20.3182(\mathrm{x}-1) + 25\\
&\boldsymbol{Interval}\ \langle 2,3\rangle: S_2(x) = 57.0911(x-2)^3 - 47.9547(x-2)^2 - 34.1364(\mathrm{x}-2) + 25\\
&\boldsymbol{Interval}\ \langle 3,4\rangle: S_3(x) = -79.5461(x-3)^3 + 123.3186(x-3)^2 + 41.2275(\mathrm{x}-3) + 0\\
&\boldsymbol{Interval}\ \langle 4,5\rangle: S_4(x) = 49.0935(x-1)^3 - 115.3198(x-1)^2 + 49.2263(\mathrm{x}-1) + 85\\
&\boldsymbol{Interval}\ \langle 5,6\rangle: S_5(x) = -17.8278(x-1)^3 + 31.9607(x-1)^2 - 34.1328(\mathrm{x}-1) + 68\\
&\boldsymbol{Interval}\ \langle 6,7\rangle: S_6(x) = 32.2179(x-1)^3 - 21.5228(x-1)^2 - 23.6950(\mathrm{x}-1) + 48\\
&\boldsymbol{Interval}\ \langle 7,8\rangle: S_7(x) = -40.0436(x-1)^3 + 75.1307(x-1)^2 + 29.9129(\mathrm{x}-1) + 35\\
&\boldsymbol{Interval}\ \langle 8,9\rangle: S_8(x) = -40.0436(x-1)^3 - 45(x-1)^2 + 60.0436(\mathrm{x}-1) + 100
\end{aligned}
\tag{1}
$$

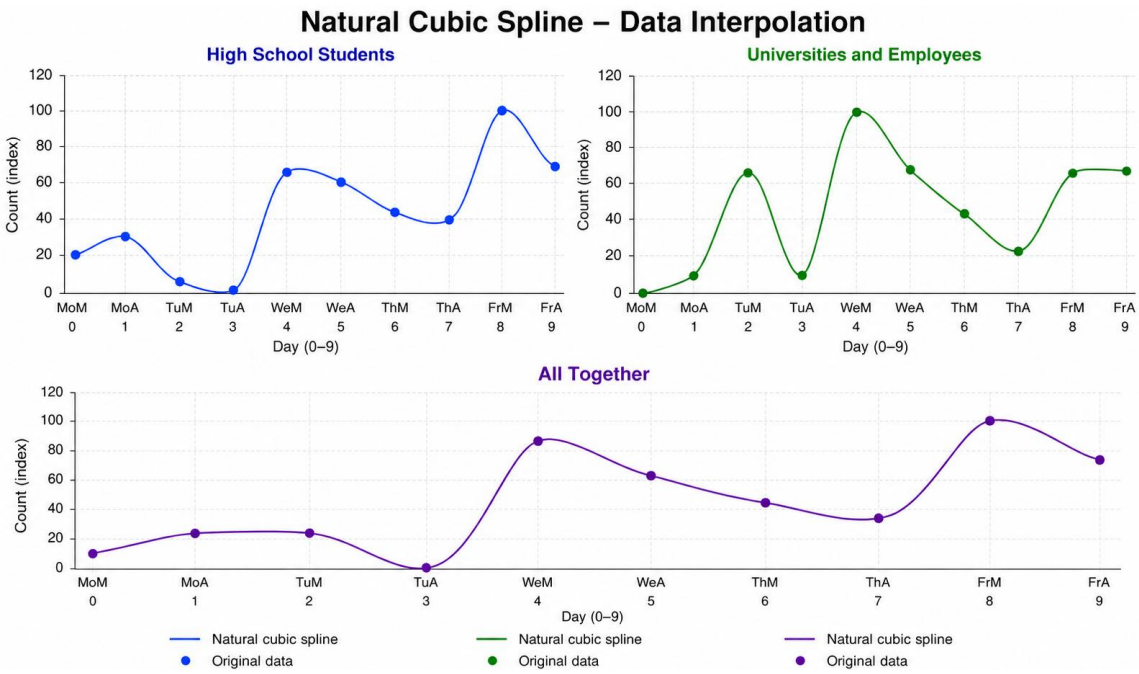


Figure 1 Cubic Spline for Work Contour Data Sets

2.6 The Impact of Student Syndrome on Work Patterns during the Week

Let us assume a task whose duration corresponds to a full working week. Based on the characteristics of the Student Syndrome, whether understood as procrastination or as the inability to complete tasks on time due to overload, it follows that, after an initial increase in work effort at the beginning of the week (Monday afternoon and Tuesday morning), a decline occurs, followed by a substantial escalation toward Friday afternoon.

A comparison with Figure 1 reveals a noticeable tendency for performance to decrease specifically on Tuesday afternoon, while, conversely, the willingness (in this case necessity) to work is relatively high on Friday. From the perspective of the Student Syndrome, these peaks are broadly consistent with expected behaviour; however, the potentially productive period of Wednesday remains underutilized.

In (Kučera et al., 2014) the authors proposed an approximation of the Student Syndrome using the formula (2)

$$\boldsymbol{y} = \left| \sin(8\pi(x-0.5)^3) \left(x + \frac{z-2}{2}\right) e^x \right| \tag{2}$$

The parameter z represents the propensity to succumb to the Student Syndrome: the higher the value of z, the greater the tendency of the workforce toward procrastination.

After rescaling the domain of the function from the interval (0, 1) to (0, 100), in accordance with the previous research, and setting the parameter $z = 1$ (representing a standard propensity to succumb to the Student Syndrome), the resulting profile of work effort is as follows (Figure 2).

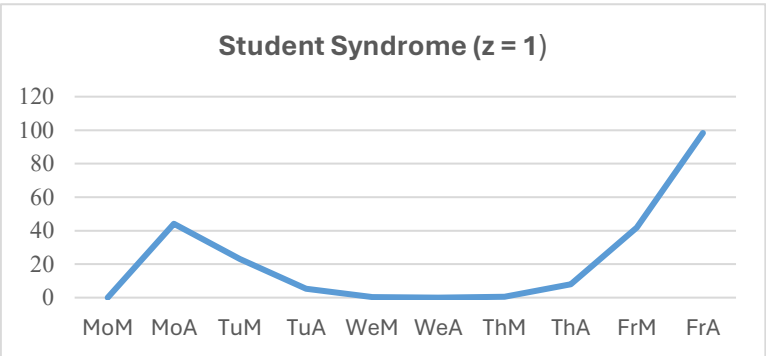


Figure 2 Student Syndrome adjusted to the Monday Morning -Friday Afternoon Scale

2.7 The Impact of Parkinson's Law on Work Patterns during the Week

Parkinson's Law states that work expands to fill the time available for its completion. The general increase in work effort under this principle follows a concave pattern, i.e., a decelerating growth trajectory, which can be suitably approximated, for example, by a polynomial function:

$$y \approx 2x^2 - 0.15x^3 \tag{3}$$

A secondary implication of this law is that if a task is completed ahead of schedule, the remaining capacity is typically not reallocated to subsequent tasks. Within the framework of our research, it is evident that the elevated work effort, i.e., the increased willingness to work observed during Wednesday (the peak of effort), may both reflect and reinforce this phenomenon. The anticipated reward is a free Friday afternoon and the postponement of further work until Monday morning. This dynamic may, however, be in tension with the procrastination tendencies associated with the Student Syndrome.

2.8 Summary of Factors Influencing the Potential Level of Work Effort During the Week

Work effort within the working week can be understood as a combination of "voluntary" willingness to work (i.e., the subjective influence of individual phases of the week and the motivation to complete tasks earlier) and the objective necessity to perform work under the influence of the Student Syndrome.

One possible approach to modelling the evolution of work effort (while simultaneously accounting for subjective influences and the objective requirement to meet deadlines) is to construct an aggregate work-effort function defined as a convex linear combination (with weights of importance as combination coefficients) of the willingness-to-work function and the Student Syndrome function (i.e., a combination of necessity and the subjectivity of procrastination). Both component functions can be expressed using splines or polynomials or others (goniometric).

However, the determination of the weights of these functions (v_1 and v_2) is inherently subjective, and even small variations in these weights may lead to significant distortions in the resulting model.

A more realistic and, for a project manager, more meaningful application lies in the use of the envelope curve of these influences, representing the willingness (respectively the necessity) to exert work effort across the individual phases of the working week. For the project manager, this corresponds to the maximum available work capacity.

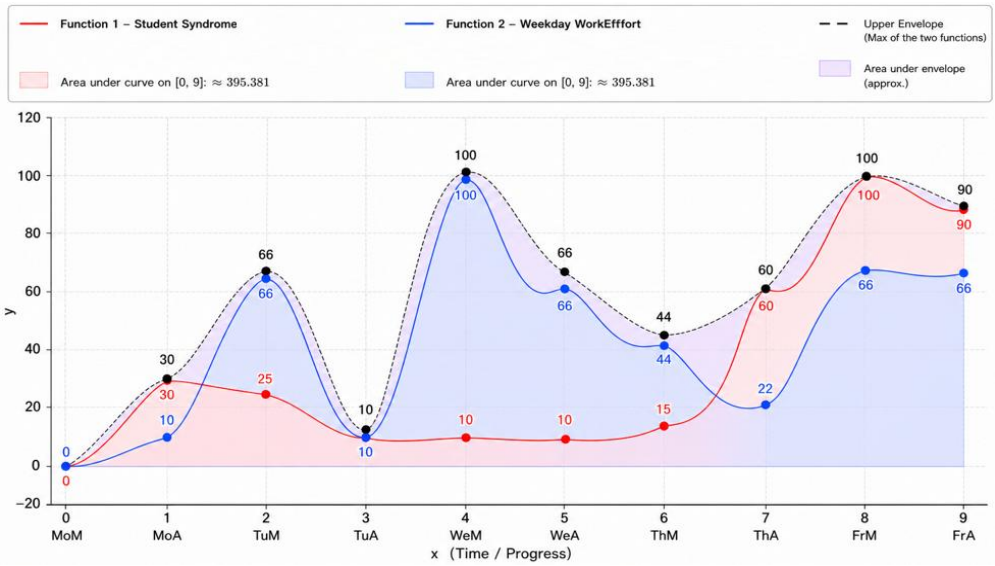


Figure 3 Envelope of Weekday work Effort and Student Syndrome

The area under the envelope curve represents the total work effort (i.e., the potential work output provided by the resource) available to the project manager over the course of the working week.

Let us examine the envelope function in greater detail. Here, too, it is possible to perform an approximation using spline functions. However, based on a detailed analysis, a trigonometric approximation was selected (Formula 4)

instead, as trigonometric functions are inherently global in nature (a single sinusoid influences the entire course of the function). The given data are thus well represented by a trigonometric functional approximation with a maximum error of ±0.009.

$$y = 49{,}07 - 13{,}2\sin(1{,}13x + 0.56) + 32{,}33\sin(0{,}23x - 0{,}68) - 27{,}19\sin(1{,}95x + 1{,}99) + \\ + 21{,}93\sin(3{,}55x + 0{,}14) \quad (4)$$

Subsequently, to determine the available work capacity, let us compute the area under this envelope curve

$$\int y\,dx = 49.07x + 11{,}48\cos(1.13x + 0.56) - 140{,}57\cos(0.23x - 0.68) + 13{,}94\cos(1.95x + \\ + 1.99) - 6{,}18\cos(3.55x + 0.14) + C \quad (5)$$

For the total amount of available work, we then evaluate the definite integral over the interval (0, 9), i.e., from Monday morning to Friday afternoon - equivalently $F(9) - F(0)$:

$$\int_0^9 y(x)\,dx \approx 530{,}4 \quad (6)$$

This represents the available number of time units of work, considering all factors influencing potential work effort per week.

3 Results and Discussion

Work effort over the course of the working week is the aggregate outcome of numerous objective and subjective factors. These influences act simultaneously and may, but need not, exhibit synergistic effects. This paper has examined willingness to work and the intensity of work effort during the working week, irrespective of the specific start or completion deadlines of individual (project) tasks.

However, practical experience indicates that many deadlines are aligned with Friday, and occasionally with Monday (following the weekend). The findings of the survey suggest a managerial implication: the highest willingness to work (and thus peak performance) is concentrated on Wednesday and Friday morning (even when accounting for the influence of the Student Syndrome). Conversely, critical periods characterized by minimal motivation tend to occur on Tuesday and Friday afternoon.

Several limitations should be acknowledged. The survey sample was relatively small and based on convenience sampling. The questionnaire relied on self-reported perceptions rather than observed productivity data. Consequently, the resulting functions should be interpreted as indicative approximations of behavioral tendencies rather than precise predictors of actual work performance. Future research should validate the proposed model using larger samples and real project performance data.

From a managerial perspective, the results suggest that project milestones and internal deadlines should preferably be aligned with periods of naturally higher work willingness, particularly Wednesday and Friday morning. Conversely, assigning critical deliverables to Tuesday afternoon may increase the risk of lower productivity and delayed completion.

4 Conclusion

This paper is a part of broader research aimed at mapping subjective and objective influences on the scheduling of project resource work contours. In contrast to most of the studies discussed above, it does not focus on a specific work schedule within an individual project task, but rather on willingness to work (and thus the availability of work effort) over the horizon of the working week. In combination with the Student Syndrome and Parkinson's Law, this availability is modelled using spline functions and trigonometric approximations.

The proposed model is a pilot in nature and should not be interpreted as a fully validated model of productivity, but it highlights an interesting interaction between natural weekly work rhythms and behavioral phenomena such as the Student Syndrome. While the Student Syndrome tends to shift effort toward the deadline, the weekly contour indicates that the most productive period may occur earlier. This mismatch may lead to inefficient utilization of available human capacity.

The proposed contour may be implemented in project scheduling software similarly to existing resource calendars or work-contour definitions. Instead of assuming constant daily productivity, resource availability coefficients may be adjusted according to the estimated weekday effort profile.

References

Amabile T M, Kramer S J (2011) The Progress Principle: Using Small Wins to Ignite Joy, Engagement, and Creativity at Work. Harvard Business Review Press

Bartoška J, Kunhart J, Pilný J (2024) Extension of Planning Poker by Work Contour Models in Project Management. In: Proceedings of the 42nd International Conference on Mathematical Methods in Economics (MME 2024). Jan Evangelista Purkyně University Ústí nad Labem

Bartoška J, Kučera P, Šubrt T (2016) Modification of the EVM by Work Effort and Student Syndrome Phenomenon. In: Proceedings of the 34th International Conference on Mathematical Methods in Economics (MME 2016). Technical University Liberec

Bartoška J, Kučera P, Šubrt T (2015) Modification of the BCWS by Work Effort. In: Proceedings of the 33rd International Conference on Mathematical Methods in Economics (MME 2015), pp 19–24. University of West Bohemia, Plzeň

Bartoška J, Šubrt T, Kučera P (2018) Resource Allocation and Motivation Strategies in Project Management. In: Proceedings of the 36th International Conference on Mathematical Methods in Economics (MME 2018). University of Economics, Jindřichův Hradec

Bartoška J, Šubrt T (2011) The Effect of Human Agent in Project Management. Central European Journal of Operations Research 19 (3): 369–382. ISSN 1435-246X

Kučera P, Bartoška J, Šubrt T (2014) Mathematical Models of the Work Contour in Project Management. In: Proceedings of the 32nd International Conference on Mathematical Methods in Economics, pp 518–523. Palacký University, Olomouc

Leach L P (1999) Critical Chain Project Management Improves Project Performance. Project Management Journal 30 (2): 39–51

Parkinson C N (1991) Parkinson's Law and Other Selected Writings on Management. Singapore: Federal Publications

Smith D (2010) The Effects of Student Syndrome, Stress, and Slack on Information Systems Development Projects. Issues in Informing Science and Information Technology 7:489–494

Steel P (2007) The Nature of Procrastination: A Meta-Analytic and Theoretical Review of Quintessential Self-Regulatory Failure. Psychological Bulletin 133(1)

Šubrt T, Bartoška J, Chamrada D (2021) Work Contour Models in Projects. In: Proceedings of the 39th International Conference on Mathematical Methods in Economics, pp 475–480. Czech University of Life Sciences Prague

Fuzzy Multi-Attributes Decision Model for Athletes Position Suitability

Vojtěch Šimůnek [1], Helena Brožová [2], Jan Saro [3]

Abstract. The paper presents a fuzzy multi-criteria decision model to help floorball or ice hockey coaches assign players to positions in the team - defender, forward, and goalkeeper. Inputs include relative results of four cognitive tests and two game situation tests, fuzzified with three-level linguistic variables. Each athlete is evaluated using a weighted sum of these fuzzified tests results, with weights describing the importance of the tests set by the coach according to position specific skills. The Athlete Placement Model provides suitability scores for athletes for each position, supporting decisions on team composition and training.
The numerical results reported in the paper were obtained using a prototype Decision Support System (DSS) in which the Athlete Placement Model is embedded together with the athlete testing tools required to acquire the input variables, complementary educational modules addressed to the athletes themselves and additional tools for coach decision-making. This DSS is intended to support coaches throughout the full cycle of evaluation, role assignment and longitudinal monitoring.

Keywords: Data Analytics for Sports, Multi-Criteria Decision Making, Fuzzy approach

JEL Classification: C44, Z20

1 Introduction

Team sports are characterised by complex, multidimensional decision problems in which a coach must integrate heterogeneous sources of information (physiological measurements, psychometric indicators, tactical context and longitudinal observations) into operational decisions concerning team composition, training design and in-game adjustments. The increasing availability of digital assessment instruments and performance analysis tools has substantially enriched the information environment in which coaches operate, while also amplifying the cognitive load associated with assimilating this information. Consequently, structured decision support frameworks have become increasingly relevant as a means of translating diverse data sources into transparent, reproducible recommendations that can be interpreted by domain experts.

Multi-criteria decision-making (MCDM) methods are particularly well suited to such problems because they explicitly accommodate the trade-offs between competing criteria that arise in athletic evaluation (Talašová 2003). Classical aggregation procedures, however, presuppose precise numerical inputs and crisp evaluation thresholds, assumptions that are rarely satisfied in practice. Athletic performance measurements are inherently noisy: results depend not only on stable individual abilities but also on transient factors such as physical condition, motivation and the recent training load. Linguistic categorisations habitually used by coaches (for example, classifying a reaction time as 'fast', 'average' or 'slow') further illustrate the imprecision that pervades the evaluative process.

Fuzzy set theory, originally proposed by Zadeh (1965) and subsequently developed into a comprehensive framework for reasoning under uncertainty (Ross 2010; Wang and Klir 1992; Zimmermann 2001), offers a natural mathematical apparatus for representing such graded, linguistically described information. Recent contributions confirm that fuzzy decision support approaches can improve the effectiveness of sport-related interventions across diverse application domains. An (2024) demonstrates the use of fuzzy systems for enhancing the design of fitness-training programmes, Novatchkov and Baca (2013) review fuzzy logic applications in strength training and broader sport science, and Liu and Wang (2026) develop a T-spherical fuzzy FUCA framework for real-time basketball analytics, Talašová (2003) also pointed usefulness of fuzzy approaches in sport. Taken together, these

Department of Systems Engineering, Faculty of Economics and Management, Czech University of Life Sciences Prague, Kamýcká 129, 165 00 Prague-Suchdol, Czech Republic
[1] simunekvojtech@pef.czu.cz
[2] brozova@pef.czu.cz
[3] saroj@pef.czu.cz

works support the view that fuzzy modelling provides an effective bridge between expert knowledge (typically articulated in linguistic terms) and the quantitative requirements of formal DSS.

Despite this growing body of work, comparatively few contributions address the specific problem of assigning players to game-specific roles using an integrated cognitive and tactical evaluation. The present study contributes to this strand by proposing a fuzzy multi-criteria decision model the Athlete Placement Model that evaluates the suitability of an individual athlete for the roles of forward, defender and goalkeeper in floorball or ice hockey. The model combines four cognitive test results with two game-situation scores, fuzzified them using trapezoidal membership functions, normalises them onto a common linguistic scale, and aggregates the normalised values via role-specific weighted sums. The final output is a suitability score that is subsequently mapped onto a five-level linguistic scale by means of a distance-based defuzzification rule.

The paper is organised as follows. Section 2 describes the input variables and the role-specific weight elicitation that together define the athlete placement process. Section 3 develops the formal fuzzy model in three steps: fuzzification, aggregation and defuzzification. Section 4 presents an illustrative application of the model and discusses the resulting role-specific suitability profiles. Section 5 concludes by summarising the main contributions, the limitations of the proposed approach and avenues for further research.

2 The Athlete Placement Process

A wide range of tests is used to assess the abilities of athletes in different playing positions (Mănescu et al., 2026, Xu and Hao, 2025). The selected tests for this model were chosen based on the expert experience of coaches working with youth players. The DSS will also enable analyses based on a larger number of tests. The athlete placement process is generally structured around two complementary groups of indicators used also in this process. Cognitive indicators capture stable, role independent components of athletic performance (perceptual speed, decision accuracy, response inhibition and working-memory capacity) that have repeatedly been associated with success in fast-paced team sports. Game situation indicators, by contrast, are role specific behavioural scores derived from controlled video or on-ice scenarios that probe the athlete's ability to resolve offensive and defensive tactical problems. The simultaneous use of cognitive and tactical indicators is intended to balance generic ability with role relevant skill, in keeping with widely held principles of talent identification in team sports.

The coach's decision about a player's position is derived from the results of six tests concerning the qualities and skills of the athletes. Four tests represent general cognitive abilities and two represent performance in specific game situations (Table 1).

Variable	Name	Type	Note
RT	Reaction time	Cognitive	Lower value is better
ACC	Accuracy	Cognitive	Higher accuracy in % is better
INH	Inhibition	Cognitive	Good / Poor
WM	Working memory	Cognitive	Working memory in % (higher is better)
RGSO	Game situation score - offense	Game situation	Score in attacking situations (higher is better)
RGSD	Game situation score - defense	Game situation	Score in defense situations (higher is better)

Table 1 List of tests

For the forward, the highest weights are assigned to RT and to the test of RSGO. A good forward must react quickly to changes in the situation and to the movements of teammates, defenders or the goalkeeper, while at the same time deciding rapidly on a shot, pass or change of direction. ACC is of medium importance, because the forward must be precise both when finishing and when combining with teammates, but accuracy alone, without quick decision-making, may not be sufficient for this role. INH and WM have lower weights, because the offensive role is associated more with dynamics, fast reactions and the resolution of attacking situations. RGSD has only supplementary importance for the forward.

For the defender, the highest weight is assigned to the test of RGSD. The defender must above all correctly evaluate defensive situations, cover space, read the opponent's movement and choose appropriate defensive solutions. INH and WM also play important roles here. The defender must be precise in the timing of interventions, in passing and in positional decision-making, while at the same time being able to suppress hasty or inappropriate reactions. This is important, for example, in one-on-one defending, where premature movement can create space for the opponent. WM carries a higher weight for the defender than for the forward, because the defender must monitor several pieces of information simultaneously - opponents, teammates, space and the development of the game

situation. RT has a lower weight, because for the defender it is not only the speed of the reaction that is decisive, but above all the correctness, control and timing of decisions. RGSO has only supplementary importance, because the offensive resolution of situations is not the main criterion of the defensive role.

For the goalkeeper, the highest weights are assigned to RT and to RGSD. This corresponds to the specific demands of the goalkeeper's role, where the ability to react very rapidly to a shot, deflection, pass in front of the goal or a sudden change in the direction of play is crucial. The goalkeeper often works in time-constrained situations in which the stimulus must be evaluated immediately, and an adequate reaction must be selected. At the same time, the ability to correctly read defensive game situations, anticipate the development of the play, take up suitable positions and react to the behaviour of opponents and teammates is essential. INH also plays an important role, because the goalkeeper must be able to suppress reactions to premature or deceptive actions, for example when an opponent fakes a shot, a pass or a change in the direction of movement. ACC is of medium importance, since it complements the speed and correctness of decisions during interventions, build-up play and the positional resolution of situations. WM has only supportive importance for the goalkeeper, since immediate perception, reaction speed, inhibition and defensive decision-making are more important for this role. The RGSO variable is not included for the goalkeeper, because the test of offensive game situations is not relevant for the evaluation of the goalkeeper's role.

Based on this expert analysis, the weights shown in Table 2 were determined. The DSS will also provide coaches with more precise methods for quantifying weights, such as Saaty's pairwise comparison method.

Variable	Weight for attacker's role	Weight for defender's role	Weight for goalkeeper's role
RT	0.30	0.10	0.20
ACC	0.15	0.20	0.15
INH	0.10	0.20	0.20
WM	0.10	0.15	0.05
RGSO	0.30	0.05	0.00
RGSD	0.05	0.30	0.40

Table 2 Default test weights by game role

3 The Athlete Placement Model

Fuzzy modelling and fuzzy logic provide a well-established framework for decision support in settings in which decisions must be derived from data that are distorted by measurement uncertainty or that are most naturally articulated in linguistic terms (Ross 2010; Wang and Klir 1992; Zadeh 1965). In contrast with classical models that rely on crisp inputs, fuzzy tools accommodate concepts such as Poor, Medium and Good through membership functions defined over an appropriate universe of discourse. Because the results of each athletic test reflect not only the player's underlying ability but also their current physical and psychological condition, individual measurements should be interpreted as values located within a band of plausible realisations rather than as point estimates; trapezoidal fuzzy numbers are particularly suitable for representing such bounded ambiguity. The Athlete Placement Model proposed below is therefore organised as a three-step process (fuzzification, weighted aggregation and defuzzification) that converts crisp test results into an interpretable assessment of role suitability.

Step 1 – Fuzzification

The first step of the model consists in the fuzzification of the individual test results. Each test outcome is initially represented by a trapezoidal fuzzy number, the parameters of which are calibrated by the coaching staff so as to reflect the expected distribution of performance within the monitored team. For five of the six tests (RT, ACC, WM, RGSO and RGSD) a three-term linguistic variable (Poor, Medium and Good) is employed; Table 3 reports illustrative parameter values, which must be re-estimated whenever the model is transferred to a different team or competitive level. The INH test, which has a naturally bipolar interpretation, is handled separately via a single membership function for the term Good; the complementary term Poor is then defined as its negation, in accordance with standard fuzzy logic conventions.

	Fuzzy numbers $(t_{1j}, t_{2j}, t_{3j}, t_{4j})$ corresponding to the linguistic terms		
Test	Term Poor	Term Medium	Term Good
RT	$Poor_{RT} = (480,530,590,590)$	$Medium_{RT} = (380,420,480,530)$	$Good_{RT} = (320,320,380,420)$
ACC	$Poor_{ACC} = (0,0,50,60)$	$Medium_{ACC} = (50,60,75,85)$	$Good_{ACC} = (75,85,100,100)$
INH			$Good_{INH} = (30,45,55,70)$
WM	$Poor_{WM} = (0,0,50,60)$	$Medium_{WM} = (50,60,70,80)$	$Good_{WM} = (70,80,100,100)$
RGSO	$Poor_{RGSO} = (0,0,40,50)$	$Medium_{RGSO} = (40,50,70,80)$	$Good_{RGSO} = (70,80,100,100)$
RGSD	$Poor_{RGSD} = (0,0,40,50)$	$Medium_{RGSD} = (40,50,70,80)$	$Good_{RGSD} = (70,80,100,100)$

Table 3 Linguistic terms for the test results in a specific team

Because the linguistic terms defined in Table 3 are anchored to different physical scales, the resulting fuzzy numbers cannot be directly compared or aggregated across tests. A normalisation step is therefore required, and it is most conveniently implemented by mapping all test outcomes onto a common dimensionless linguistic scale. The scale used in the present model, reported in Table 4, partitions the unit interval into three normalised fuzzy terms Poor, Medium and Good that subsequently serve as building blocks for the weighted aggregation procedure.

Fuzzy numbers $(t_{1j}, t_{2j}, t_{3j}, t_{4j})$ corresponding to the linguistic terms		
Poor test result	Medium test result	Good test result
$T_{Poor} = (0,0,0.10,0.25)$	$T_{Medium} = (0.20,0.35,0.50,0.70)$	$T_{Good} = (0.60,0.75,1.00,1.00)$

Table 4 Linguistic terms for normalisation

This secondary fuzzification proceeds as follows. For each test result, the values of the membership functions of the individual linguistic terms are first determined. These membership function values are then used as weights by which the fuzzy numbers corresponding to the fuzzy linguistic scale for normalisation are multiplied. The fuzzy values of the individual indicators are calculated as the weighted sum of all values of the linguistic variable used for normalisation, where the weights are the degrees of membership of the given indicator to the initial values of the linguistic variable; for example, for reaction time the test result is normalised using the formula:

$$RT_i = (rt_{i1}, rt_{i2}, rt_{i3}, rt_{i4}) = \mu_{RT}^{Poor}(x_{RT}) \cdot T_{Poor} + \mu_{RT}^{Medium}(x_{RT}) \cdot T_{Medium} + \mu_{RT}^{Good}(x_{RT}) \cdot T_{Good} \quad (1)$$

where $RT_i = (rt_{i1}, rt_{i2}, rt_{i3}, rt_{i4})$ is the fuzzification of the reaction time x_{RT} measured for the i-th athlete; T_{Poor}, T_{Medium} and T_{Good} are the fuzzy representations of the normalised linguistic terms for the RT test; and $\mu_{RT}^{Poor}(x_{RT})$, $\mu_{RT}^{Medium}(x_{RT})$ and $\mu_{RT}^{Good}(x_{RT})$ are the values of the membership functions of the linguistic terms of the RT test for the monitored team.

The fuzzy values of the other tests are calculated similarly. For the INH test, only two complementary terms Good and Poor are used. The value of the membership function for the term Poor is $\mu_{INH}^{Poor}(x) = 1 - \mu_{INH}^{Good}(x)$. For the INH test result, due to the binomial structure, the assignment Poor → T_{Low} and Good → T_{High} is used.

$$INH_i = (1 - \mu_{INH}^{Good}(x_{INH})) \cdot T_{Poor} + \mu_{INH}^{Good}(x_{INH}) \cdot T_{Good} \quad (2)$$

Step 2 – Aggregation

In the second step the overall potential of an athlete for a given playing role is computed as a weighted aggregation of the normalised partial fuzzy values produced in the previous step. The weights can be either taken from the default profile reported in Table 2 or tailored to the specific philosophy of the coaching staff. Because the underlying operations are linear in the components of the trapezoidal representation, the result is itself a trapezoidal fuzzy number and the model retains a closed form interpretation. The aggregated potential for the three roles is obtained from the following equations:

$$PPF_i = vf_{RT} \cdot RT_i + vf_{ACC} \cdot ACC_i + vf_{WM} \cdot WM_i + vf_{RGSO} \cdot RGSO_i + vf_{RGSD} \cdot \mathrm{RGSD}_i + vf_{INH} \cdot INH_i \quad (3)$$
$$PPD_i = vd_{RT} \cdot RT_i + vd_{ACC} \cdot ACC_i + vd_{WM} \cdot WM_i + vd_{RGSO} \cdot RGSO_i + vd_{RGSD}.\mathrm{RGSD}_i + vd_{INH}.INH_i \quad (4)$$
$$PPG_i = vg_{RT} \cdot RT_i + vg_{ACC} \cdot ACC_i + vg_{WM} \cdot WM_i + vg_{RGSD} \cdot \mathrm{RGSD}_i + vg_{INH} \cdot INH_i \quad (5)$$

where PPf_i denotes the potential of athlet i for the forward role, PPD_i denotes the potential of player i for the defender role, and PPG_i denotes the potential of player i as a goalkeeper, $vf_{RT}, \dots, vd_{RT}, \dots, vg_{RT}, \dots$ are the weights from the Table 2

Step 3 – Defuzzification

The final step of the Athlete Placement Model consists in defuzzifying the aggregated potentials to obtain an interpretable linguistic assessment of role suitability. The potentials PPF, PPD and PPG are classified using the five-level linguistic scale of suitability, the membership functions of which are reported in Table 5. Since the output universe of the model is the unit interval ⟨0,1⟩, the classification scale is constructed over the same domain, and each suitability level is represented by a trapezoidal fuzzy number.

Linguistic term	Fuzzy number
Unsuitable	$V_1 = (0.00,0.00,0.10,0.20)$
Weakly suitable	$V_2 = (0.10,0.20,0.30,0.40)$
Moderately suitable	$V_3 = (0.30,0.40,0.55,0.65)$
Suitable	$V_4 = (0.55,0.65,0.80,0.90)$
Highly suitable	$V_5 = (0.80,0.90,1.00,1.00)$

Table 5 Linguistic terms for the final recommendation of role suitability for the player

The linguistic term that best characterises the aggregated potential is identified by computing the distance between the resulting trapezoidal fuzzy number and each fuzzy number of the suitability linguistic scale, and by selecting the term for which this distance is minimal. The distance between two trapezoidal fuzzy numbers is defined as the sum of the absolute differences between their corresponding components, normalised by the sum of the core and support widths of the reference fuzzy number:

$$d(PPF_i, V_j) = \frac{\sum_{i=1}^{4} |PPF_{ik} - v_{jk}|}{(v_{j4} - v_{j1}) + (v_{j3} - v_{j2})} \tag{6}$$

where PPF_{ik} denotes the kth component of the athlete i resulting fuzzy number PPF_i and v_{jk} the corresponding component of the k-th reference fuzzy number V_j of the language scale.

The resulting linguistic term is determined as the one for which the calculated distance is minimal:

$$\hat{V}_i = \arg\min_j d(PPF_i, V_j) \tag{7}$$

This step converts the resulting fuzzy potential into a comprehensible linguistic interpretation $\hat{V}_i$, which allows for a simpler assessment of a player's suitability for a given game position with the measure $d(PPF_i, \hat{V}_i)$.

4 Results

The behaviour of the proposed model is illustrated below on a representative athlete drawn from a monitored squad. All numerical inputs, fuzzy parameterisations and weight profiles correspond to those defined in the previous section.

Figure 1 illustrates the fuzzification stage of the Athlete Placement Model for the selected athlete. The crisp input values of the RT, ACC, INH, WM and RGSO and RGSD are first located on the team-specific trapezoidal fuzzy scales calibrated in Tables 3. The resulting degrees of membership in the linguistic terms Poor, Medium, and Good (or Good and Poor in the case of INH) are reported beneath each test panel. These membership degrees provide the input to the subsequent normalisation and aggregation steps and ensure that information expressed on heterogeneous physical scales can be compared and combined within a single evaluative framework.

Figure 2 summarises the output of the aggregation step. For each of the three target roles, the figure displays the trapezoidal fuzzy potential together with a representative midpoint that simplifies visual comparison. The mid-point values indicate that the goalkeeper role yields the highest potential, followed by the defender role, while the forward role is associated with the lowest aggregated potential. The width of the fuzzy intervals additionally conveys information about the residual uncertainty inherent in each evaluation: narrower intervals correspond to more decisive recommendations, whereas wider intervals signal a need for additional evidence before a final placement is made.

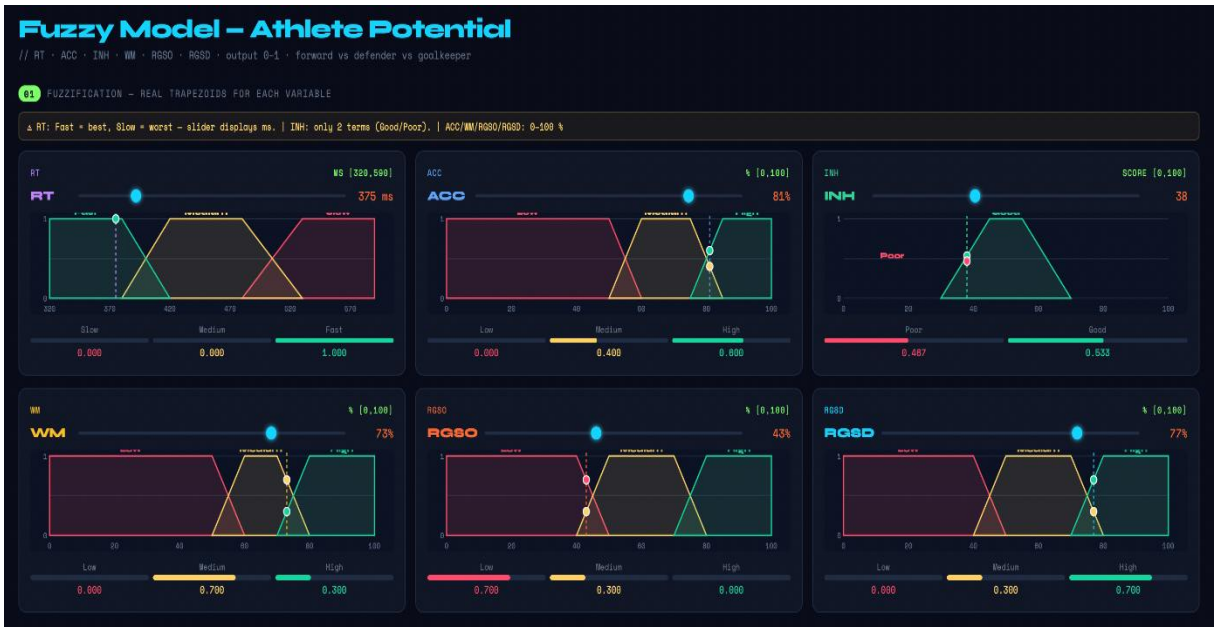


Figure 1 Test results mapped onto the initial fuzzy linguistic scale

Figure 2 Overall potential score

The defuzzification step is summarised in Figure 3, which reports the linguistic suitability level assigned to each role.

Within the five-level suitability scale, the evaluated athlete is classified as moderately suitable for the forward role, whereas the defender and goalkeeper roles both reach the Suitable category. The highest absolute potential is nonetheless obtained for the goalkeeper role, which would therefore be recommended as the primary placement, with the defender role as a viable secondary option.

Figure 3 Final results of role potential score

The numerical results reported above were obtained using a prototype DSS in which the Athlete Placement Model is embedded together with the athlete testing tools required to acquire the input variables and complementary

educational modules addressed to the athletes themselves. The system is intended to support coaches throughout the full cycle of evaluation, role assignment and longitudinal monitoring.

5 Conclusion

The introduced Athlete Placement Model, a fuzzy multi-criteria decision model that evaluates the suitability of an athlete for the roles of forward, defender and goalkeeper in floorball or ice hockey articulates four cognitive indicators and two role-specific game-situation scores within a three-step process (fuzzification, weighted aggregation and distance-based defuzzification). The use of trapezoidal fuzzy numbers ensures that the variability of athletic performance is propagated consistently from the raw measurements to the final linguistic recommendation, while the role-specific weight profiles allow the coaching staff to encode domain-specific priorities in a transparent and reproducible manner. The example confirmed that the resulting suitability is interpretable and actionable, supporting both primary placement decisions and the identification of plausible secondary roles.

Several limitations should be acknowledged. First, the parameters of the fuzzy linguistic scales and the role-specific weights have been elicited from a single coaching context and must be recalibrated whenever the model is transferred to a different team or to athletes of a different competitive level. Second, the empirical evidence presented in this article is limited to a single illustrative athlete; systematic validation against expert role assignments on a larger sample is needed to assess the predictive performance of the model. Third, the underlying tests are considered conditionally independent, so future research will investigate this dependence also.

References

An P (2024) Fuzzy decision support systems to improve the effectiveness of training programs in the field of sports fitness. International Journal of Computational Intelligence Systems 17:1–14. https://doi.org/10.1007/s44196-024-00555-z

Liu D, Wang W (2026) Advancing sports performance evaluation through T-spherical fuzzy FUCA MCDM approach to real-time basketball analytics. Scientific Reports 16. https://doi.org/10.1038/s41598-025-32934-7

Mănescu D C., Filip C, Nae C I, Ciomag R V (2026). Optimizing Athletic Performance: A Systems Framework for Adaptive Training, Load Management, and Decision-Making. Journal of Functional Morphology and Kinesiology 11(3), 245. https://doi.org/10.3390/jfmk11030245

Novatchkov H, Baca A (2013) Fuzzy logic in sports: a review and an illustrative case study in the field of strength training. International Journal of Computer Applications 71:8–14. https://doi.org/10.5120/12360-8675

Ross T J (2010) Fuzzy logic with engineering applications. John Wiley & Sons, Chichester

Talašová J (2003) Fuzzy metody vícekriteriálního hodnocení a rozhodování. Olomouc: Vydavatelství Univerzity Palackého.

Wang Z, Klir G J (1992) Fuzzy measure theory. Plenum Press, New York

Xu J, Hao F (2025). The role of personality traits in athlete selection: A systematic review. Acta Psychologica, 260, 105478. https://doi.org/10.1016/j.actpsy.2025.105478

Zadeh L A (1965) Fuzzy sets. Information and Control 8:338–353

Zimmermann H J (2001) Fuzzy set theory and its applications, 4th ed. Kluwer Academic Publishers, Boston

Legitimacy–Cost Gap in Reshoring: Evidence from Central Europe

Šindlerová Markéta[1], Urminský Jaroslav[2], Černoch Jakub[3], Melecký Ondřej[4]

Abstract. This paper examines the discrepancy between normative support for reshoring and consumers' willingness to bear its economic consequences, referred to as the legitimacy–cost gap. Using survey data from 2,000 respondents in Czechia and Slovakia, a gap index is constructed to capture this discrepancy. Its existence and distribution are assessed using complementary mean, median, and proportion-based tests, followed by a categorical analysis of integer-threshold attitude profiles.
The results indicate that normative support significantly exceeds economic willingness and that paradox-type attitudes dominate the sample. Regression analysis further indicates that economic patriotism is positively associated with both normative support and economic willingness, while the relationship between normative support and economic willingness varies across age groups. Older consumers exhibit a stronger association between normative support and economic willingness than younger respondents. Overall, the findings suggest that reshoring enjoys broad symbolic support, whereas consumers' declared willingness to bear its economic costs remains conditional and varies across consumer groups.

Keywords: Reshoring, Legitimacy–cost gap, Economic patriotism, Consumer attitudes

JEL Classification: F23, L23, M31, D12

1 Introduction

1.1 Reshoring and the Demand-Side Perspective

Reshoring represents firms' strategic decisions to relocate previously offshored activities to another country, including the home country. Its forming part of a broader spatial reconfiguration of global value chains (GVCs). While earlier offshoring waves were primarily driven by labour cost differentials and efficiency-seeking strategies, recent developments—including supply chain disruptions, geopolitical tensions etc.—have shifted attention toward resilience, reliability, and strategic autonomy (Fratocchi et al., 2016; Fratocchi & Di Stefano, 2019). Rather than a reversal of globalization, reshoring reflects an adaptive and contingent adjustment of globally fragmented production systems, shaped by multiple factors and uncertainty (Henkel et al., 2022).

Reshoring literature remains predominantly supply-side oriented, focusing on firm-level determinants such as costs, capabilities, and risk management. Less attention has been devoted to the demand side, although reshoring often involves higher production costs that may be passed on to consumers through higher prices. Consequently, the economic feasibility of reshoring depends not only on firm considerations but also on consumer acceptance.

Consumers can thus be understood as a conditional driver of reshoring. Prior research shows that consumers generally respond positively to reshoring initiatives, associating them with desirable outcomes such as job creation, domestic economic development, and improved ethical or environmental standards (Grappi et al., 2015; 2018; Gillani et al., 2022). Such normative endorsement does not necessarily correspond to consumers' declared willingness to bear the associated economic costs. Consumers may support domestic production in principle, while remaining unwilling to bear the associated costs in practice. This tension motivates the concept of a legitimacy–cost gap, defined as the discrepancy between normative support for reshoring and consumers' willingness to accept

[1]VSB-TUO, Department of Systems Engineering and Informatics, 17. Listopadu 2172/15, 708 00 Ostrava-Poruba, Czechia marketa.sindlerova@vsb, ORCID: https://orcid.org/0000-0001-6741-2657
[2]VSB-TUO, Department of Marketing and Business, 17. Listopadu 2172/15, 708 00 Ostrava-Poruba, Czechia jaroslav.urminsky@vsb.cz,, ORCID: https://orcid.org/0000-0002-8254-4305
[3]VSB-TUO, Student of Department of Systems Engineering and Informatics, 17. Listopadu 2172/15, 708 00 Ostrava-Poruba, Czechia
[4] VSB-TUO, Student of Department of Marketing and Business, 17. Listopadu 2172/15, 708 00 Ostrava-Poruba, Czechia

its economic consequences, such as higher prices. The concept reflects a discrepancy between declared normative support and declared economic willingness. This gap represents a critical demand-side constraint, even if reshoring is perceived as desirable and legitimate, insufficient willingness to internalize its costs may limit its practical feasibility.

1.2 Attitudinal Framework and Research Hypotheses

To operationalize the demand-side perspective, three attitudinal dimensions are distinguished, capturing complementary aspects of consumer reasoning related to reshoring.

Although conceptually related, economic patriotism primarily reflects consumers' orientation toward supporting the domestic economy through purchasing decisions, whereas normative support evaluates reshoring as a desirable economic strategy. Economic patriotism reflects consumers' preference for domestic production and their motivation to support the national economy through purchasing decisions. It is conceptually related to the literature on consumer ethnocentrism, which emphasizes identity-based motivations to prefer domestic over foreign products (Siamagka & Balabanis, 2015). From this perspective, economic patriotism represents a value-based orientation rooted in national identity and perceptions of economic solidarity.

Normative support represents the perceived desirability and societal legitimacy of returning production activities to the home country. This dimension reflects value-based evaluations related to broader socio-economic outcomes, such as employment creation, economic stability, and national resilience (Grappi et al., 2015; 2018). While normative support reflects legitimacy perceptions, the economic viability of reshoring ultimately depends on whether consumers are willing to internalize its costs. Economic willingness to bear reshoring costs therefore captures the behavioural-economic dimension of reshoring support. It reflects consumers' readiness to accept higher prices or other trade-offs associated with domestic production (Grappi et al., 2018; Su & Huang, 2023). Building on this framework, the study addresses two interrelated research questions with five hypotheses.

RQ_1: Does a systematic legitimacy–cost gap exist between consumers' normative support for reshoring and their willingness to bear its economic consequences, and how is this gap structured across distinct attitude profiles?

- **H_1:** The mean level of normative support for reshoring exceeds the mean level of consumers' economic willingness to bear its costs.
- **H_2:** The distribution of the legitimacy–cost gap across integer-threshold categories is systematically non-uniform, **with** paradox categories occurring more frequently than the category in which economic willingness exceeds normative support.

RQ_2: What factors determine consumers' economic willingness to bear the costs associated with reshoring?

- **H_3: Economic** patriotism is positively associated with normative support for reshoring.
- **H_4: Economic** patriotism is positively associated with economic willingness, controlling for normative support.
- **H_5: Age** positively moderates the relationship between normative support and economic willingness, indicating a stronger positive association at higher age.

2 Research Design

2.1 Data

The empirical analysis is based on primary quantitative data collected in November 2025 through an online survey using the Computer-Assisted Web Interviewing (CAWI) method. Data collection was administered by the professional research agency MNFORCE, s.r.o., via its proprietary online panel in an omnibus design. A multi-stage quality control procedure was implemented, including verification of unique access links, detection of duplicate registrations or minimum completion time controls. Only complete and consistent questionnaires were retained.

The final sample comprises N = 2000 respondents (n_{cz} = 1000; n_{sk} = 1000). Quota sampling based on gender, age, education level, region, and municipality size was applied to approximate key population characteristics.

Attitudinal variables were measured on a seven-point Likert scale (1 = strongly disagree; 7 = strongly agree), allowing for sufficient discriminative power and approximate interval-level interpretation suitable for econometric analysis. Although Likert scales are ordinal by definition, they are commonly treated as approximately interval-level measures in applied survey research when composite indices are constructed. To reduce dependence on this assumption, the analysis of the legitimacy–cost gap is complemented by non-parametric robustness tests, including

the Wilcoxon signed-rank and binomial tests. The whole questionnaire included 20 items structured into five theoretical constructs (each 4 items) and five identification questions. Three reverse-coded items were incorporated to reduce acquiescence bias.

The five conceptual dimensions originally included economic patriotism, normative support for reshoring, supply chain trust and economic security, ethical–environmental orientation, and economic willingness to bear reshoring costs. However, the empirical analysis primarily focuses on the relationships among economic patriotism (EP), normative support (NS), and consumers' willingness to bear its economic consequences (EW). The final specification therefore prioritizes these theoretically central constructs in order to preserve model parsimony and to directly examine the legitimacy–cost discrepancy addressed in this study.

2.2 Measurement Model

The first analytical step assesses the measurement properties of the three theoretical constructs. The analysis examines whether the constructs represent empirically distinct dimensions of consumer attitudes to reshoring. The internal consistency of the scales was assessed using Cronbach's alpha (Cronbach, 1951), defined as:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} \sigma_j^2}{\sigma_T^2}\right),$$

where k denotes the number of items in the scale, σ_j^2 the variance of individual items, and σ_T^2 the variance of the total scale score. Reliability values above 0.70 were considered acceptable. Composite construct scores were calculated as the mean of the retained items and used in the subsequent econometric analysis. To provide additional evidence of construct distinctiveness, Pearson correlations among the principal constructs are reported in Appendix A.

2.3 Structural Model

Following the assessment of the measurement model, the structural relationships between the key constructs were examined. Table 1 summarizes the research framework and empirical strategy.

RQ_x	H_x	Formulation	Empirical method
RQ_1	H_1	$\mu_{NS} > \mu_{EW}$	Paired samples t-test
	H_2	$Gap_i = NS_i - EW_i$,	Gap index construction
		Typology based on integer thresholds	χ^2 test, post-hoc residuals
		Robustness: $\mu_{Gap} > 0; Me_{Gap} \neq 0; P(Gap_i > 0) > 0.5$	One-sample t-test, Wilcoxon, Binomial test
RQ_2	H_3	$NS_i = \beta_0 + \beta_1 EP_i + \varepsilon_i$, $\beta_1 > 0$	Linear regression (robust SE)
	H_4	$EW_i = \beta_0 + \beta_1 NS_i + \beta_2 EP_i + \varepsilon_i$, $\beta_2 > 0$	Linear regression (robust SE)
	H_5	$EW_i = \beta_0 + \beta_1 NS_{c,i} + \beta_2 Age_{c,i} + \beta_3 (NS_{c,i} \times Age_{c,i}) + \varepsilon_i$, $\beta_3 > 0$	Moderated linear regression
		$\frac{\partial EW_i}{\partial NS_{c,i}} = \beta_1 + \beta_3 Age_{c,i}$	Marginal effect

Notes: NS_i – normative support; EW_i – economic willingness; EP_i – economic patriotism; $NS_{c,i}$ – mean-centered normative support; $Age_{c,i}$– mean-centered age used in the moderation analysis; Gap_i – legitimacy–cost gap.

Table 1 Research Framework and Empirical Strategy

Construction and Evaluation of the Legitimacy–Cost Gap

Normative support and economic willingness of respondent are measured on an identical seven-point Likert scale (NS_i, $EW_i \in [1,7]$). The legitimacy–cost gap is defined as:

$$Gap_i = NS_i - EW_i$$

with theoretical range $Gap_i \in [-6,6]$. Positive values indicate higher normative support relative to economic willingness, zero denotes consistency, and negative values indicate that declared economic willingness exceeds normative support. Given the common metric, the index can be interpreted as a signed distance between two evaluative criteria. Formally, for $\mathrm{x}_i = (NS_i, EW_i)$,

$$Gap_i = \mathrm{w}'\mathrm{x}_i, \mathrm{w} = (1, -1)$$

representing a linear projection into a one-dimensional discrepancy space.

For interpretative clarity, the index is transformed into an ordinal typology based on integer distances of the Likert metric:

$$\begin{cases} Gap_i < 0 & \text{Higher economic willingness} \\ 0 \leq Gap_i < 1 & \text{Mild paradox} \\ 1 \leq Gap_i < 2 & \text{Strong paradox} \\ Gap_i \geq 2 & \text{Symbolic support} \end{cases}.$$

Thresholds (0, 1, 2) are defined ex ante according to the unit structure of the scale, ensuring metric interpretability and cross-sample comparability. As a one-unit difference corresponds to a shift of one Likert category, thresholds reflect integer distances rather than sample-dependent cut-offs. The typology is intentionally asymmetric, consistent with the analytical focus on the positive domain reflecting the analytical focus on situations in which normative support exceeds declared economic willingness. The proposed typology is intended as an interpretative tool for describing patterns in declared attitudes rather than as a classification of actual consumer behaviour.

The existence and structure of the gap (RQ_1) are evaluated using complementary inferential procedures (see Table 1), including mean and median-based tests, a proportion test of sign dominance, and chi-square analysis of the typology. This layered approach ensures distributional robustness.

Regression Models

To examine the determinants of reshoring-related attitudes (RQ2), three linear regression models with heteroskedasticity-robust (Huber–White) standard errors were estimated (see Table 1). Model 1 evaluates the association between economic patriotism and normative support for reshoring. Model 2 examines economic willingness by including both normative support and economic patriotism, thereby assessing the direct contribution of economic patriotism beyond its association with normative endorsement. Model 3 introduces age as a moderator of the relationship between normative support and economic willingness. The interacting variables are mean-centred to facilitate interpretation and reduce non-essential multicollinearity. Together, the models examine the associations among economic patriotism, normative support for reshoring, and consumers' declared economic willingness. All statistical analyses were conducted using IBM SPSS Statistics 31.0.

3 Empirical Analysis

3.1 Reliability

The whole questionnaire included three reverse-coded items intended to reduce acquiescence bias. Only one of these items was part of the constructs analysed in the present study. Reliability testing indicated that this item slightly lowered the internal consistency of the respective scale. It was therefore removed prior to constructing the final indices, resulting in improved reliability. Cronbach's alpha was used to assess the internal consistency of the measurement constructs. As shown in Table 2, all retained constructs exhibit satisfactory reliability ($\alpha > 0.7$), indicating acceptable internal consistency.

Construct	Number of Items	Cronbach's α
Economic patriotism	3	0.833
Normative support for reshoring	4	0.913
Economic willingness	4	0.828

Table 2 Reliability of Measurement Constructs

Composite construct scores were subsequently computed as the mean of the retained items and used in the econometric analysis. Pearson correlations among the principal constructs were all positive and statistically

significant ($p < 0.001$), while remaining below the commonly accepted threshold of 0.80, thereby providing additional evidence of the empirical distinctiveness of the analysed constructs (see Appendix A).

3.2 Evidence of the Legitimacy-Cost Gap

After confirming the reliability of the constructs, the analysis proceeded to examine (H_1) the relationship between normative support for reshoring and the willingness to bear its economic consequences (see Table 3).

Normative support is significantly higher than economic willingness (mean difference = 0.895, $t = 32.033$, $p < 0.001$), with a substantial effect size (Cohen's $d = 0.716$). These results support H1 and indicate a systematic discrepancy between the two attitudinal dimensions. The result is robust across alternative distributional assumptions. A one-sample test confirms that the mean of the gap index is strictly positive ($\mu_{Gap} > 0$, $p < 0.001$), the Wilcoxon signed-rank test rejects median equality ($p < 0.001$), and the proportion of respondents with $Gap_i > 0$ significantly exceeds 0.5 (binomial test, $p < 0.001$). Together, these tests indicate that the observed discrepancy is robust across different analytical approaches.

Variable	Mean	Std. Dev.
Normative support for reshoring	5.375	1.415
Economic willingness	4.480	1.366
Test	**Value**	
Mean difference	0.895	
t-statistic	32.033	
p-value	<.001	

Table 3 Paired Comparison of Normative Support and Economic Willingness

The discrepancy between normative support for reshoring and the willingness to bear its economic consequences was operationalized as the legitimacy–cost gap index, with respondents classified into four categories reflecting different levels of consistency between normative support and declared economic willingness.

In the sample (N = 2000), the **mild paradox** category is most prevalent (39.4%), followed by **strong paradox** (26.3%), **symbolic support** (20.3%), and **higher economic willingness** (14.1%). The dominance of paradox categories indicates that reshoring support is frequently conditional: although normative endorsement of domestic production is widespread, it is not always accompanied by a corresponding willingness to accept higher prices or related trade-offs. These findings indicate that declared support for reshoring is frequently conditional and varies in its consistency with consumers' declared economic willingness. To examine potential cross-country differences in gap typology, a χ^2 test was conducted. The results indicate a statistically significant association between country and attitude type (χ^2 (3) = 9.397, $p = 0.024$). Descriptive results suggest that Czech respondents are relatively more represented in the categories of symbolic support and strong paradox, while Slovak respondents exhibit higher shares in the mild paradox and higher economic willingness category. However, post-hoc analysis based on standardized residuals does not indicate any statistically significant cell-level deviations (all $|z| < 1.96$). Therefore, the observed differences cannot be interpreted as systematic. This interpretation is further supported by a small effect size (Cramer's $V = 0.069$).

3.3 Determinants of Economic Willingness and Moderating Role of Age

The regression results are presented in Table 4. The first model examines the relationship between economic patriotism and normative support for reshoring. The results indicate a strong and statistically significant positive association ($B = 0.626$, $p < 0.001$), providing support for H_3 and suggesting that respondents expressing stronger patriotic economic attitudes also report higher normative support for domestic production and reshoring initiatives.

Variables	Model 1	Model 2	Model 3
Dependent variable:	Normative support	Economic willingness	Economic willingness
Economic patriotism	0.626***	0.471***	0.469***
Normative support		0.256***	
NS_c			0.276***
Age_c			-0.033**
NS_c × Age_c			0.040***
N	2000	2000	2000

*Notes:_c – centered; B value reported; Robust (Huber–White) standard errors; *** p < 0.001, ** p < 0.05.*

Table 4 Regression Results (Linear regression models with Huber–White robust standard errors)

Further, Model 2 examines H4 by estimating the associations of economic patriotism and normative support with consumers' willingness to bear reshoring-related costs. Both variables are positively and significantly associated with economic willingness. Economic patriotism is associated with higher economic willingness (B = 0.471, $p < 0.001$), while normative support also shows an independent positive association (B = 0.256, $p < 0.001$).

Finally, Model 3 examines H5 by introducing age as a moderator of the relationship between normative support and economic willingness. The interaction term is positive and statistically significant (B = 0.040, $p = 0.002$), supporting H5 and indicating that the positive association between normative support and economic willingness becomes stronger with increasing age. Although age is negatively associated with baseline economic willingness (B = −0.033, $p = 0.031$), the positive interaction suggests that older respondents exhibit a stronger relationship between normative support and economic willingness than younger respondents. Overall, the regression results suggest that consumers' declared willingness to bear reshoring-related costs is associated with both economic patriotism and normative support, while the strength of this association varies across age groups.

4 Conclusion

This study examined the discrepancy between consumers' normative support for reshoring and their willingness to bear its economic consequences. The results consistently show that normative support for domestic production exceeds consumers' declared willingness to accept the associated costs, indicating a legitimacy–cost discrepancy in declared consumer attitudes toward reshoring. Moreover, paradox-type attitude profiles represent the most prevalent response pattern, suggesting that support for reshoring is frequently conditional rather than unconditional. The regression analysis further indicates that economic patriotism is positively associated with both normative support and economic willingness. In addition, the positive association between normative support and economic willingness becomes stronger with increasing age, suggesting that the relationship between these attitudinal dimensions differs across age groups. From a practical perspective, the findings suggest that declared support for domestic production should not automatically be interpreted as a corresponding willingness to accept higher prices. The proposed gap typology may therefore provide a useful segmentation framework for distinguishing consumer groups according to the consistency between their declared support for reshoring and their declared willingness to bear its economic consequences. The study is subject to several limitations. First, the analysis relies on cross-sectional survey data and therefore does not permit causal inference. Second, all measures are based on self-reported attitudes rather than observed purchasing behaviour. Future research could extend the proposed framework by incorporating additional attitudinal dimensions and complementing self-reported measures with behavioural indicators, such as experimentally elicited willingness to pay or observed purchasing decisions. Such evidence would contribute to a better understanding of the conditions under which declared support for reshoring is accompanied by a willingness to accept its economic costs.

Acknowledgements

This paper is financed by Student Grant Competition of the Faculty of Economics, VSB-Technical University of Ostrava; project's registration number is SP2026/071. All support is greatly acknowledged and appreciated.

References

Cronbach L J (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334. https://doi.org/10.1007/BF02310555

Fratocchi L, Ancarani A, Barbieri P, Di Mauro C, Nassimbeni G, Sartor M, Vignoli M, Zanoni A (2016) Motivations of Manufacturing Reshoring: An Interpretative Framework. International Journal of Physical Distribution & Logistics Management 46:98–127. https://doi.org/10.1108/IJPDLM-06-2014-0131

Fratocchi L, Di Stefano C (2019) Manufacturing Reshoring: A Research Agenda. International Journal of Physical Distribution & Logistics Management 49:105–127. https://doi.org/10.1108/IJPDLM-02-2018-0066

Gillani A, Kutaula S, Budhwar P (2022) Heading Home? Reshoring and Sustainability Connectedness from a Home-Country Consumer Perspective. British Journal of Management 34:1117–1137. https://doi.org/10.1111/1467-8551.12658

Grappi S, Romani S, Bagozzi R P (2015) Consumer Stakeholder Responses to Reshoring Strategies. Journal of the Academy of Marketing Science 43:453–471. https://doi.org/10.1007/s11747-015-0432-y

Grappi S, Romani S, Bagozzi R P (2018) Reshoring from a Demand-Side Perspective: Consumer Reshoring Sentiment and Its Market Effects. Journal of World Business 53:194–208. https://doi.org/10.1016/j.jwb.2017.11.001

Henkel, M, Boffelli, A, Olhager, J, Kalchschmidt, M (2022). A Case Survey of Offshoring-Backshoring Cases: The Influence of Contingency Factors. Internal Journal of Production Economics 249: 108615. https://doi.org/10.1016/j.i.jpe.2022.108615

Siamagka N T, Balabanis G (2015) Revisiting Consumer Ethnocentrism: Review, Reconceptualization, and Empirical Testing. Journal of International Marketing 23:66–86. https://doi.org/10.1509/jim.14.0085

Su L, Huang Y (2023) Willingness to Pay Premium for Green Agricultural Products. Rajabhat Chiang Mai Research Journal 24:13–30. https://doi.org/10.57260/rcmrj.2023.264981

Appendix

Appendix A Pearson correlations among the principal constructs

Construct	1	2	3
1. Economic patriotism	1		
2. Normative support	0.653***	1	
3. Economic willingness	0.682***	0.597***	1

*Notes: Pearson correlation coefficients. *** $p < 0.001$ (two-tailed). $N = 2000$*

The Impact of Competition and Quality on the Efficiency of Integrated Transportation Systems

Eva Štichhauerová[1], Miroslav Žižka[2]

Abstract. This article examines the technical efficiency and service effectiveness of 14 integrated transport systems in Czechia. The transport systems were evaluated using a non-oriented Network Data Envelopment Analysis (NDEA) model, assuming variable returns to scale, in two variants: fixed-link and free-link. The model consists of two phases. The first phase measures how efficiently regions purchase services from transport operators. The second phase assesses whether transport services are used effectively by passengers. The study also evaluated the impact of competition in the selection of transport operators on technical efficiency and the impact of transport system quality on service effectiveness. The research results showed that the average technical efficiency score is higher than the average service effectiveness score. Transportation systems in three regions (Central Bohemia, Prague, and the Karlovy Vary region) were identified as benchmarks. The fixed-link model has lower discriminatory power and is better suited to operational management of transportation systems. The free-link model also shows the desired change in intermediates and is suitable for future planning of the scope of transportation services.

Keywords: integrated transport system, technical efficiency, service effectiveness, slacks-based measure model, network data envelopment analysis

JEL Classification: C61, H54, R42

1 Introduction

This paper builds upon an earlier research (Štichhauerová and Žižka, 2025), which examined the efficiency of regional transport systems in Czechia. Unlike the cited paper, the model was extended to include factors characterizing the degree of competition in operator selection and the evaluation of the quality of regional integrated transport systems. At the same time, the latest available data for integrated transport systems (ITS) in 2024 were used. A network DEA model was used instead of a simple DEA model.

The research aimed to determine the overall efficiency of individual ITS from two interrelated perspectives. First, we examined the technical efficiency (TE) of the regional transport system. This refers to how efficiently individual regions procure scheduled bus or regional rail services. We assume that greater competition among transport operators when bidding for public service contracts to provide bus or rail transport services could positively affect this efficiency. Second, we examined the service effectiveness (SE) of the transport system, which reflects how passengers utilize the offered transport services. In this case, we consider that the quality of the integrated transport system positively influences the rate of public transport service usage. Finally, we examined whether competition in operator selection and the quality of the transport system influence the levels of efficiency and effectiveness.

2 Literature review

Research on the efficiency of transportation systems is a relatively common topic in the literature. In essence, the evaluation of transportation efficiency can be divided into two categories of problems. The first group involves single-stage evaluations of the efficiency of individual carriers or transportation systems. The second group encompasses evaluations of interconnected multi-stage systems. Our article addresses the second group of problems. This group includes, for example, situations in which a public or private entity invests capital in specific transportation infrastructure or fixed assets, and we subsequently examine the degree of utilization of that infrastructure or those assets. Given the limited space, we will focus on articles that addressed a similar problem

[1] Technical University of Liberec, Faculty of Economics, Department of Business Administration and Management, Studentska 1402/2, Liberec 1, Czechia, eva.stichhauerova@tul.cz.
[2] Technical University of Liberec, Faculty of Economics, Department of Business Administration and Management, Studentska 1402/2, Liberec 1, Czechia, miroslav.zizka@tul.cz.

using data envelopment analysis. Yu and Lin (2008) examined efficiency and effectiveness in 20 railway transport systems across various countries. A characteristic feature of railway transport is that certain inputs can be shared between passenger and freight transport (e.g., infrastructure and staff). Other inputs (e.g., railcars) are, however, unique to each mode. In the first phase, they assessed how effectively rail operators use inputs to generate transport performance (train-kilometers in passenger and freight transport). These two outputs were then used as inputs for evaluating service effectiveness in the second phase. The outputs of the second phase were transport performance (passenger-kilometers and ton-kilometers), which characterize the utilization of the offered transport services. They addressed the problem using a multi-activity network data envelopment analysis (NDEA) model with shared inputs. They also considered population density and GNI per capita as factors that could influence service effectiveness.

Similarly, Benga et al. (2025) evaluated the efficiency of 21 railway systems between 2016 and 2018 using various models (CRS-DEA, CRS-NDEA, VRS-DEA, VRS-NDEA) in two phases. For the first phase, which measured technical efficiency, they used track length, number of employees, and the rolling stock as inputs. The outputs - which also served as intermediate outputs and inputs for the second phase - were train-kilometers and gross hauled ton-kilometers. The final outputs of the second phase, referred to as service effectiveness, were passenger-kilometers and ton-kilometers. They further defined several factors that could influence efficiency and effectiveness. They assessed the impact of these factors using Tobit regression. They found that only GNI has a significant impact on technical efficiency. None of the listed factors had a significant impact on service effectiveness. The authors conclude that other factors, such as service quality, may influence service effectiveness, which represents a research gap we aim to address in our study.

Fitzová et al. (2018) investigated the determinants of efficiency for 19 urban transport companies in the Czech Republic using time-series data from 2010 to 2015. They employed standard CRS and VRS DEA models with three inputs (fleet, energy costs, staff) and one output (number of passengers). They then used a Tobit regression model to determine the impact of selected demographic, technical, and economic indicators on efficiency. The research results showed that the proportion of drivers among total employees, the average age of the fleet, the presence of a tram service, and population density positively impact efficiency. Conversely, the average fare, the share of subsidies in revenue, and the existence of a two-city system reduced the efficiency rate. Žižka (2017) reached similar conclusions, though he evaluated urban transit agencies from two perspectives: efficiency and effectiveness, using traditional CRS and VRS DEA models. In this study as well, the age of the fleet, population density, network density, low-floor vehicles, and the history of the transport system all had a positive impact on both indicators.

In analyzing efficiency and effectiveness, there are two main approaches. The first approach uses traditional DEA methods, such as radial CCR and BCC models or non-radial slacks-based models, to analyze efficiency or effectiveness in each phase separately. This approach is simple and identifies the main sources of inefficiency. However, it does not take into account that a change in the volume of intermediates (i.e., outputs of the first phase and inputs to the second phase) can increase the overall effectiveness of the system. Each phase is evaluated and optimized separately, without any link between them. The second approach is based on Network DEA (NDEA) models, which view the system as a whole. They reveal situations in which local optimization of a single phase can lead to global inefficiency (Kao and Hwang, 2008).

To analyze efficiency and effectiveness in our research, we selected the NDEA model, which, unlike the traditional "black box" DEA approach that does not account for the transformation process (Färe and Grosskopf, 2000), considers interrelationships across the entire system. Specifically, the Network Slack-Based Measure (SBM) model, as defined by Tone and Tsutsui (2009), was applied. This model was preferred over the radial model proposed by Kao and Hwang (2008) because it better captures slacks. Another reason was the ability to work with an unoriented model, which better addresses the potential conflict between minimizing inputs in the first phase and maximizing outputs in the second phase (Liang et al., 2008).

3 Data and methodology

For our research, we used data from all 14 regional ITS in Czechia for 2024. The model consists of two stages. The first stage (TE) expresses how efficiently the regions procure transport services from rail and bus operators. Regions procure transport services based on long-term contracts (10–15 years), while new contracts must be awarded through public tenders since 2019 (bus transport) and 2023 (rail transport), respectively. The inputs to the first stage reflect the financial amounts the regions spend on public service obligations, as passenger fare revenue covers only a portion of the costs. We used the regions' expenditures on providing regional bus and rail transport as inputs. This data was obtained from public service obligation reports, which the regions are required

to publish. However, for the year 2024, these reports were available for only 7 regions. Additional data sources included the regions' transport service plans and the annual reports of public transport operators in the regions. In one case, we had to request the data under the Freedom of Information Act. The outputs of the first stage are transport performance, expressed in train-kilometers for rail transport and vehicle-kilometers for bus transport. These variables serve as intermediate outputs, as they simultaneously function as inputs to the second stage. The model's outputs are then the volumes of passengers transported in both modes of transport. Data on public transport performance and the number of passengers transported in bus and regional rail services were obtained from the transport yearbook (Slabá and Houšť, 2025). The second stage (SE) assesses how the region-subsidized rail and bus lines are used by passengers, i.e., whether the volume of transport services offered is sufficient to meet the region's public service obligations. Both stages are logically closely linked. A limited transport supply may appear cost-efficient from the region's perspective, but it will not meet passenger needs, leaving connections semi-empty. Conversely, if the supply is extensive and connections are frequent, the passenger response is likely to be positive. However, a large volume of transport performance will require high financial costs. In practice, this involves balancing supply and demand for transport services while accounting for regions' financial capacity. The model structure is shown in Figure 1. Basic descriptive statistics for all variables are presented in Table 1. The table reveals significant variability in the data.

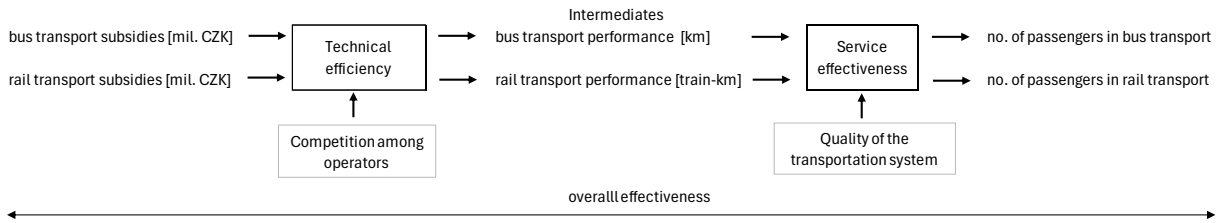


Figure 1 Model structure

To calculate the overall effectiveness (OE) score and the sub-scores for TE and SE, we used a non-oriented slack-based NDEA model with the VRS assumption (Tone and Tsutsui, 2009). Initially, a fixed-link model was used because the volume of transport services (intermediates) is contracted, typically for a long period. The contracts allow only a small margin for changes in the scope of services (e.g., by 5%), which is related to the high fixed costs carriers incur in maintaining an adequate fleet. However, this assumption proved to be very limiting, so a free-link model was also tested. Both phases - operational (technical) and service - are equally important for achieving overall effectiveness. For this reason, we have assigned them equal weight in our calculations.

The SBM-NDEA model can be formulated as follow: The number of DMUs is n (in our research, 14), the number of phases is K (k = 1, …, K), in our case K = 2, (k, h) is the link between nodes k and h, Z is the matrix of intermediates and w^k is the relative importance of phase k. For the non-oriented model, we minimize the objective function (1) subject to the constraints (2) and (3), for example, the article by Tone and Tsutsui (2009).

$$\rho_0 = min\left(\frac{\sum_{k=1}^{K} w^k\left[1-\frac{1}{m_k}\left(\sum_{i=1}^{m_k}\frac{s_i^{k-}}{x_{i0}^k}\right)\right]}{\sum_{k=1}^{K} w^k\left[1+\frac{1}{r_k}\left(\sum_{r=1}^{r_k}\frac{s_r^{k+}}{y_{r0}^k}\right)\right]}\right) \tag{1}$$

subject to

$$\begin{gathered} \boldsymbol{x}_0^k = X^k\boldsymbol{\lambda}^{\boldsymbol{k}} + \boldsymbol{s}^{k-} \quad (k=1,\dots,K) \\ \boldsymbol{y}_0{}^k = Y^k\boldsymbol{\lambda}^{\boldsymbol{k}} - \boldsymbol{s}^{k+} \quad (k=1,\dots,K) \\ \mathrm{e}\boldsymbol{\lambda}^k = 1 \quad (k=1,\dots,K) \\ \boldsymbol{\lambda}^k \geq \boldsymbol{0}, \boldsymbol{s}^{k-} \geq \boldsymbol{0}, \boldsymbol{s}^{\boldsymbol{k}+} \geq \boldsymbol{0}, \boldsymbol{w}^k \geq 0 \end{gathered} \tag{2}$$

In the model, the individual phases are connected either by a fixed link (3) or by a free link (4).

$$\begin{gathered} \boldsymbol{z}_o^{(k,h)} = Z^{(k,h)}\boldsymbol{\lambda}^h \\ \boldsymbol{z}_o^{(k,h)} = Z^{(k,h)}\boldsymbol{\lambda}^k \end{gathered} \tag{3}$$

$$Z^{(k,h)}\boldsymbol{\lambda}^h = Z^{(k,h)}\boldsymbol{\lambda}^k \text{ where } Z^{(k,h)} = \left(z_1^{(k,h)},\dots,z_n^{(k,h)}\right) \tag{4}$$

The solution yields an overall effectiveness score, which can be broken down into sub-scores for efficiency and effectiveness using equation (5).

$$\rho_k = \frac{1 - \frac{1}{m_k}\left(\sum_{i=1}^{m_k} \frac{s_i^{k-*}}{x_{i0}^k}\right)}{1 + \frac{1}{r_k}\left(\sum_{r=1}^{r_k} \frac{s_r^{k+*}}{y_{r0}^k}\right)} \tag{5}$$

where $\boldsymbol{s}^{k-*}$and $\boldsymbol{s}^{k+*}$are the optimal input and output slack vectors for (1).

In the final phase of the study, we examined the impact of competition and the quality of the transport system on TE and SE scores. Regions can procure transport services in several ways. Launching open tenders for public service obligations in both rail and bus transport is a way for regions to reduce their costs for these services. Open tenders lead to more competitors bidding for contracts, which should result in lower unit transport costs. However, quality standards for public service obligations (e.g., maximum permissible vehicle age, equipment) must be clearly defined. The use of open tenders generally results in more operators providing public transport services in a region. We measured the level of competition among carriers using the Herfindahl-Hirschman Index (HHI) (6), which measures market concentration. We obtained the data for the HHI calculation from an overview of contracts concluded between regions and operators, which is included in public service obligation reports or is published in regional transport service plans.

$$HHI = \sum_{i=1}^{n} \left(\frac{S_i}{S}\right)^2 \tag{6}$$

where S_i is the market share of the i-th operator in bus or rail transport. We calculated the HHI separately for both modes of transport, and $S = \sum S_i$. The HHI ranges from 0 (perfect competition) to 10,000 (monopoly). In our study, the HHI ranged from 1,100 (bus transport in the CB Region) to 10,000 (rail transport in the OL Region).

To evaluate the quality of transport systems, we used results from a study by Kostrej (2025). This study, based on the evaluation methodology (Jareš, 2016), includes a unique dataset on the quality of all ITS in Czechia. ITS were evaluated using a point-based system based on five criteria: level of integration, fare integration, passenger ticketing, transfer connections, and institutions. In each category, the system was rated on a scale of 1 to 3, where 1 represents the worst possible state and 3 the best. An integrated transport system could score up to 15 points. Based on the total number of points, ITS were divided into three categories: fully developed (12-15 points), developing (9-11 points), and underdeveloped (fewer than 8 points). The number of points was used as the independent variable. A Tobit regression model with a latent variable was used to determine the impact of competition and quality (variables f_{ij}) on the TE and SE scores (Iqbal et al., 2025). The Tobit model (7) uses a censored dependent variable, which in our case ranges from 0 to 1.

$$\rho_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j f_{ij} + \varepsilon_i$$
$$\rho_i = \begin{cases} \rho_i^* & \text{if } 0 < \rho_i^* < 1 \\ 0 & \text{if } \rho_i^* \leq 0 \\ 1 & \text{if } \rho_i^* \geq 1 \end{cases} \tag{7}$$

Variable	Min	Max	Mean	SD	CV
Bus subsidies	309.097	21,246.329	2,363.445	5,263.182	222.691
Rail subsidies	495.023	1,288,294.000	93,233.606	331,450.800	355.506
Bus performance	7,414,671.000	205,501,561.000	41,354,760.024	48,748,783.920	117.879
Rail performance	3,069,900.000	15,700,000.000	6,516,820.014	2,883,320.266	44.244
No. of pass. (bus)	6,212.000	64,051.900	22.852.879	16,645.014	117.879
No. of pass. (rail)	3,494.300	58,501.800	17,129.036	14,544.852	72.836
HHIBUS	1,100.274	9,985.899	4,345.238	2,424.142	55.788
HHIRAIL	3,066.020	10,000.000	7,582.315	2,423.504	31.963
Quality	6	15	10.500	2.584	24.612

Table 1 Descriptive statistics for the variables

4 Research results

The TE, SE, and OE scores are presented in Table 2. First, we used a fixed-link model, assuming that the volume of transport services could not be changed. The table shows that the average TE score is relatively high, exceeding the average SE score. Of the 14 regions, 10 were on the TE efficiency frontier. In the case of SE, this applied to 9 regions. Thus, the model distinguishes relatively little between systems. Regions that did not achieve full efficiency can be divided into service-focused (MS, SM) or operationally-oriented (HK, LB, SB, VY). The SB and VY transport systems achieved very low overall effectiveness scores. We then applied the model with a free link.

Region	**Fixed link**			**Free link**		
	TE	**SE**	**OE**	**TE**	**SE**	**OE**
Central Bohemian (CB)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Hradec Kralove (HK)	1.0000	0.6747	0.6747	1.0000	0.6717	0.6717
Karlovy Vary (KV)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Liberec (LB)	1.0000	0.6610	0.6610	1.0000	0.6610	0.6610
Moravian-Silesian (MS)	0.7876	1.0000	0.7876	0.7956	0.8126	0.6465
Olomouc (OL)	1.0000	1.0000	1.0000	0.6713	0.8204	0.5507
Pardubice (PA)	1.0000	1.0000	1.0000	0.8062	0.7274	0.5864
Pilsen (PL)	1.0000	1.0000	1.0000	0.4192	0.6512	0.2730
Prague (PR)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
South Bohemian (SB)	0.7458	0.4826	0.3599	0.7680	0.4169	0.3202
South Moravian (SM)	0.6740	1.0000	0.6740	0.6740	1.0000	0.6740
Usti and Labem (UL)	1.0000	0.3753	0.3753	0.9450	0.3652	0.3451
Vysocina (VY)	0.8772	0.5292	0.4642	0.7109	0.4322	0.3073
Zlin (ZL)	1.0000	0.6609	0.6609	0.8216	0.5421	0.4454
Mean	**0.9346**	**0.8131**	**0.7613**	**0.8297**	**0.7213**	**0.6058**

Table 2 Efficiency and effectiveness scores of integrated transport systems

In this case as well, the average TE score is higher than the average SE score. This second model discriminates significantly better. In the operational phase, only five systems were identified at the efficient frontier; in the second phase, four were identified. Only three regions achieved an overall score of 1.0 (CB, PR, and KV regions). Most transport systems can be classified as operationally oriented (in the HK, LB, PA, SB, UL, VY, ZL regions). There are three customer-oriented systems in terms of operational efficiency (OL, SM, and MS regions). The transport system in the PL region achieved very low scores in both phases.

Figure 2 shows the desired changes in inputs and outputs to achieve full effectiveness. The left side illustrates a situation in which changes to intermediate outputs are not possible. It is clear from the figure on the left that the SM and MS regions should focus on reducing the costs of providing transport services, particularly in bus transport. The SB region should focus primarily on attracting additional passengers in both bus and rail transport, while also partially reducing the costs of providing transport services. The UL region has a very low public transport usage rate; the desired increase in passenger volume for both modes of transport is enormous (and practically unrealistic). An increase in passenger volumes is also necessary in the HK, LB, VY, and ZL regions, with the first two regions requiring roughly the same increase in both bus and rail transport. In the VY and ZL regions, passenger volumes should increase primarily in rail transport. The change in intermediate products is zero because we used a fixed-link model, as explained above. The figure on the right indicates changes in intermediate results. This figure shows that the UL region likely has an excessive level of transport services; the model suggests reducing contracted bus performance by approximately 26% and rail performance by approximately 10%. A slight reduction in the scope of transport services (up to 2%) is also proposed in the HK Region. In the PA and PL regions, the model considers it desirable to reduce the ordered performance for rail transport slightly and, conversely, to increase the ordered volume for bus transport. An increase in transport volumes across both modes is particularly desirable in the OL, SB, and VY regions.

In addition to changes in intermediates, changes in the volume of inputs are also desirable - the costs of providing transport services should be reduced primarily in the SM, MS, SB, UL, and VY regions. An increase in passenger numbers is particularly desirable in the SB, UL, PL, and LB regions.

Table 2 highlights the lower discriminatory power of the model in the first stage (TE). At the same time, Table 1 shows high heterogeneity of the variables, confirming the suitability of using DEA for evaluating transport systems. The units can be considered homogeneous in terms of their operations and objectives – all perform the same activity (ordering, financing, and organizing regional public transport) under the same legislation. Naturally, the model's discriminatory power is influenced by the number of units and variables, where, according to the 'rule of thumb', the number of DMUs should be double, or preferably triple, the number of inputs and outputs (Zhu, 2014). This implies at least 12 to 18 DMUs. From this perspective, 14 DMUs is a threshold number, but since we are working with all existing regions in the Czech Republic, the number of units cannot be increased. To partially mitigate this limitation, we applied a free-link model, which identified only three regions as overall efficient and only five regions in the first stage (TE). It turns out that the insufficient discriminatory power of the first model was not due to a lack of data variability, but rather to the strict constraints on the links between the two stages. Alternatively, a super-efficiency model could be used in combination with the fixed-link model; however, due to space constraints, we do not present the detailed formulation and results of this model. The super-efficiency model reduces the number of efficient units to four in the first stage (CB, HK, LB, PR) and to three in the second stage (KV, PR, SM).

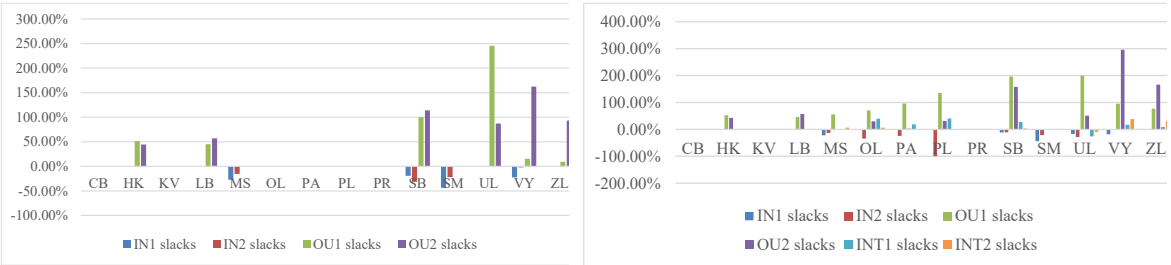


Figure 2 Desired change in slacks on the input and output sides (fixed-link model on the left, free-link model on the right)

In the final stage, we examined the potential impact of competition and quality on the TE and SE scores for both model variants. The results are presented in Table 3. The regression results indicate that we were unable to demonstrate a significant impact of competition on the TE score. This partially surprising result can be explained by a transition period, during which most regions are still procuring transport services based on old contracts concluded, particularly in rail transport, via direct award. We assume that the effect of open tenders will only be reflected in cost data after a longer time lag, once historical contracts with incumbent dominant operators expire. In the case of SE, for the free-link model, we do not reject the hypothesis at the 5% level that the quality of the integrated transport system has a positive effect on service effectiveness. Therefore, it can be stated that investments in the soft parameters of integrated transport systems (fare integration, easy ticketing, transfer connections) increase passenger demand and thereby enhance the effectiveness of the transport services offered. Each additional point in the quality assessment increases the effectiveness score by approximately 5.7 percentage points (free-link model). For the fixed-link model, a significant effect of transport system quality on SE was not confirmed at the 5% level; it was significant only at the 10% level. We must, however, note that we have a relatively small number of observations for the regression.

	Fixed link model				**Free link model**			
	TE		**SE**		**TE**		**SE**	
	Coeff.	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**	**Coeff.**	**Sig.**
(Intercept)	1.285	0.005	0.043	0.935	1.065	<0.001	0.174	0.564
HHIBUS	0.000	0.293			0.000	0.413		
HHIRAIL	0.000	0.295			0.000	0.659		
FAQuality			0.094	0.081			0.057	0.048
Log(scale)	-1.381	<0.001	-0.928	0.005	-1.477	<0.001	-1.379	<0.001
Wald st.	2.357	0.308	3.042	0.081	0.780	0.677	3.911	0.048

Table 3 Tobit regression

5 Conclusion

The research findings indicate that three regional transit systems in Czechia can serve as benchmarks for others: CB, PR, and KV. The first two systems cover large areas and serve a significant portion of the population. At the same time, they are closely interconnected. In contrast, the integrated transport system in KV is small and cost-effective, operating primarily in peripheral areas. Other regional transport systems generally have reserves either in operations or in service utilization. Some systems have reserves in both areas.

We tested two models. The fixed-link model was significantly less effective at distinguishing between well-functioning systems and underperforming ones. From a practical perspective, however, it better reflects reality, as regions can only significantly change the volume of transport services over the long term - that is, when entering into new contracts with transport operators. The fixed-link model can be used for the operational management of the transport system. It provides information on whether a region is purchasing transport services at too high a cost or by how much it should increase passenger volumes, which is related to the quality of the services offered. The free-link model is suitable for the strategic management of regional transport. Compared to the previous model, it additionally provides information on the desired adjustment of the scope of transport services. It appears that, particularly in the UL region, the scope of transport is likely overstated. In other regions (OL, SB, and VY), on the other hand, strengthening the supply side is desirable and could subsequently trigger a positive response on the demand side, thereby increasing service effectiveness. To increase service effectiveness, it is also necessary to further deepen the temporal and fare integration of transport systems and improve their user-friendliness and quality. It may be somewhat surprising that the impact of competition on technical efficiency was not demonstrated. The reason may be that, in most regions, rail transport services in particular were not yet subject to competition during the period under review. The effects of competition may therefore only become apparent in the future.

A limitation of this study is the small number of units analyzed, despite examining all existing transportation systems. A possible solution and a topic for further research would be to track the evolution of transportation systems over time and create a pooled dataset of the same units spanning a longer time period.

Acknowledgements

Supported by institutional funds for long-term conceptual development of the research organization of the TUL.

References

Benga A, Delgado Rodríguez MJ, De Lucas Santos S, Zeneli G (2025) Efficiency analysis of European railway companies and the effect of demand reduction. Journal of Rail Transport Planning & Management 34:100516. https://doi.org/10.1016/j.jrtpm.2025.100516

Färe R, Grosskopf S (2000) Network DEA. Socio-Economic Planning Sciences 34:35–49. https://doi.org/10.1016/S0038-0121(99)00012-9

Fitzová H, Matulová M, Tomeš Z (2018) Determinants of urban public transport efficiency: case study of the Czech Republic. European Transport Research Review 10:42. https://doi.org/10.1186/s12544-018-0311-y

Iqbal M, Ma J, Mushtaq Z, et al (2025) Energy efficiency evaluation of construction projects using data envelopment analysis and Tobit regression. Sci Rep 15:11444. https://doi.org/10.1038/s41598-025-90671-3

Jareš M (2016) Integrovaná doprava v praxi: jedna jízdenka, jeden tarif, jeden jízdní řád, jedna síť [Integrated Transport in Practice: One Ticket, One Fare, One Timetable, One Network]. Česká technika - nakladatelství ČVUT, Prague.

Kao C, Hwang S-N (2008) Efficiency decomposition in two-stage data envelopment analysis: An application to non-life insurance companies in Taiwan. European Journal of Operational Research 185:418–429. https://doi.org/10.1016/j.ejor.2006.11.041

Kostrej P (2025) Efektivnost systémů veřejné dopravy v ČR, [Efficiency of Public Transport Systems in Czechia]. Master's thesis, Masaryk University, Faculty of Economics and Administration, Brno. https://is.muni.cz/th/n94gq/Efektivnost_systemu_verejne_dopravy_v_CR.pdf

Liang L, Cook WD, Zhu J (2008) DEA models for two-stage processes: Game approach and efficiency decomposition. Naval Research Logistics 55:643–653. https://doi.org/10.1002/nav.20308

Slabá R, Houšť R (2025) Transport Yearbook 2024. https://www.sydos.cz/cs/rocenka-2024/index.html. Accessed 22 Feb 2026.

Štichhauerová E, Žižka M (2025) Efficiency of regional transport system. In: 43rd International Conference on Mathematical Methods in Economics (MME 2025). Conference Proceedings. Czech Society for Operations Research, Praha, pp 31–36.

Tone K, Tsutsui M (2009) Network DEA: A slacks-based measure approach. European Journal of Operational Research 197:243–252. https://doi.org/10.1016/j.ejor.2008.05.027

Yu M-M, Lin ETJ (2008) Efficiency and effectiveness in railway performance using a multi-activity network DEA model. Omega 36:1005–1017. https://doi.org/10.1016/j.omega.2007.06.003

Zhu J (2014) Quantitative Models for Performance Evaluation and Benchmarking. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-06647-9

Žižka M (2017) An Assessment of the efficiency and effectiveness of the services of urban transport operators in the Czech Republic. Transformations in Business & Economics 42:134–152

Comparison Between Logistic Regression and Machine Learning Methods Using Monte Carlo Simulations

Kryštof Tomsa[1]

Abstract. There have been many benchmarks of classification algorithms over the years, which primarily focused their evaluation on real-world datasets. A less common approach to benchmarking classification methods has been a simulation study, which this paper aims to rectify. Monte Carlo simulations are performed by simulating 500 artificial datasets across different scenarios and varying numbers of observations, on which the classification methods are trained and evaluated. The chosen methods for comparison are logistic regression, random forest, XGBoost, neural network and Gaussian naive Bayes.

Keywords: machine learning, binary classification, comparison, simulation study, benchmark

JEL Classification: C44

1 Introduction

There are many methods that can be applied to the binary classification problem. Often, practitioners choose several models that they train on a given dataset, compare them, and then choose the best one for their purpose. Although there has been increasing access to computing power, training many classification models can still be expensive. This raises the question of whether a practitioner can choose a method for a given dataset that will have a higher probability of being better than the rest of the methods and thereby saving time and resources.

There have been many benchmarks, which tried to determine this. Fernández-Delgado et al, 2014 compared 179 classifiers across 121 datasets and identified random forest as the best-performing model, though its performance was not significantly greater than the second-best model (support vector machines). The authors also determined classifiers that were not competitive at all in their analysis, such as some specific types of neural networks, nearest neighbour models and Bayesian models. This benchmark was somewhat criticised by Wainberg et al, 2016, who questioned the methodology and conclusions of Fernández-Delgado et al, 2014 regarding random forests. The authors mainly criticised the absence of a held-out test set, the exclusion of trials with errors, and the reasoning for choosing random forest as the best classifier. Nevertheless, the authors do not provide their own benchmark in their paper. Another comparison was conducted by Caruana and Niculescu-Mizil, 2006, who identified calibrated boosted trees as the best-performing model. However, the authors acknowledge that other models, like random forest, support vector machines and neural networks, also performed very well. Lessmann et al, 2015 conducted a benchmark of classification methods focused specifically on credit risk scoring. In their work, they argue that although a complex classifier does not necessarily improve overall performance over an individual classifier (such as logistic regression), some of them perform significantly better, like random forest, neural network, and particularly hill-climbing ensemble selection with bootstrap sampling, which was identified as the best model.

There have not been many benchmarks performed on artificial datasets and many studies focus on a particular method and its comparison to similar methods rather than a full-scale benchmark. Scholz and Wimmer, 2021 investigated the performance of classifiers across 6 different scenarios (data generating processes) with 2 different settings for both the number of observations and the number of features. In their work, the authors determined the heterogeneous ensemble classifier to be the best-performing model, though they also conclude that homogeneous ensembles often do not outperform individual classifiers. However, it is important to mention that the authors excluded some of the ensembles and other classifiers in some of the scenarios due to high training time, which biases the results, but can also give some notion of training costs. Orzechowski and Moore, 2021 proposed a novel approach to benchmarking by generating mathematical functions that maximise the difference in performance of chosen machine learning methods and are then used to generate the artificial datasets. From their analyses it appears that boosting algorithms are the best-performing models. However, the authors do not discuss the results in much detail and do not describe how the data generating processes influence the algorithms. Beram and El-Kotory, 2024 published another benchmark of classifiers on artificial datasets, which consisted of 84 scenarios

[1] Vysoká škola ekonomická, nám. W. Churchilla 1938/4, 13067 Praha 3, tomk09@vse.cz

in total. Logistic regression and probit models were determined to perform the best in this study, though support vector classifiers were better for some scenarios. Interestingly, the performance of some models, including logistic regression, deteriorated in certain scenarios with increasing sample size. Nevertheless, the authors did not include some popular models, mainly ensembles.

This study aims to expand the research on simulation-based benchmarks by comparing popular classification models using Monte Carlo simulations. The reviewed benchmarks often did not include the most popular or state-of-the-art models, or had limited or poorly interpretable data generating processes. Therefore, this study includes the most popular classification models and also has a clearly defined data generating process that can include a number of modifications, which are evaluated in isolation. In contrast to evaluations on real-world datasets, simulation-based benchmarking allows isolating the possible effects on predictive power of models, making the results more interpretable.

2 Data Generating Process

The base data generating process (DGP) for this simulation study closely mimics the latent variable formulation of logistic regression. Let X_j be a random variable, where j represents a predictor and $x_{ij}, i = 1, 2, ..., N$ be its *i-th* realisation. Then the latent variable Z is calculated as

$$Z = 0.8X_1 - 0.6X_2 + 0.6X_3 - 0.9X_4, \qquad \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\mu}$ is a vector of zeros with the same length as $\mathbf{X}$ and $\boldsymbol{\Sigma}$ is the defined covariance matrix, that is calculated from the given correlation matrix, which can be either a unit matrix (meaning without multicollinearity) or one with multicollinearity defined as

$$\begin{bmatrix} 1 & 0.2 & 0.7 & 0.1 \\ 0.2 & 1 & 0.2 & 0.1 \\ 0.7 & 0.2 & 1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 1 \end{bmatrix}. \tag{2}$$

Let z_i and ε_i be the *i-th* realisations of Z and $\mathcal{E} \sim Logistic(0, 1)$ respectively. The DGP further continues as

$$y_i^* = z_i - Q_{0.7}(\mathbf{z}) + \varepsilon_i, \tag{3}$$

where $Q_p(\cdot)$ is used for obtaining the empirical p-quantile, which is used to make the data unbalanced. Finally, the predicted value is obtained as $y = 1[y^* > 0]$.

A number of modifications are then added to the base DGP, which, in the order they are introduced, are:

- nonnormally distributed predictors, which changes the distribution of the first two predictors to $X_1 \sim \chi_1^2 - 1$ and $X_2 \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$ (and the correlations defined for these two predictors are disregarded),
- normally distributed noise, which changes the distribution of noise to $\mathcal{E} \sim \mathcal{N}(0, \pi^2/3)$,
- heteroskedasticity, defined as $\varepsilon_i^* = x_{i4} \times \varepsilon_i$,
- endogeneity, defined as $x_{i,3}^* = x_{i,3} + 0.2\varepsilon_i$, after which the feature is standardised as $x_{i,3}^{**} = (x_{i,3}^* - \overline{x^*}_3)/\sigma_{x_3^*}$ (where $\sigma_{\mathbf{x}_3^*}$ is the standard deviation of $\mathbf{x}_3^*$),
- nonlinearity, defined as $Z = |1.3X_1| + \sin(-1.3X_2) + 0.1X_3 - 0.5X_4$,
- omitted variable, which modifies the function to $Z^* = Z + 1.15X_5$ (but it is not added as a predictor to the final dataset),
- nonlinear predictor, which further modifies the function to $Z^* = Z - 0.1X_4^2$ and adds a new predictor $X_6 = X_4^2$,
- time dependence, defined as $z_i^* = 0.5z_{i-1} + z_i,\ i > 1$, in which case a new predictor $x_{i,5} = y_{i-1}$ is introduced (while the first value $x_{1,5} = y_1$),
- noncausal predictor, which is correlated with X_4 by a coefficient of 0.7 and adds the predictor $X_7 \sim \mathcal{N}(0, 1)$ to the generated dataset.

Any of these modifications can be combined with one another in the DGP. The scenarios for the simulation study can be seen in Table 1. For every scenario there are 3 possible settings for the number of observations for the generated dataset (N): 50, 500, 5 000. For every setup 500 datasets are generated, resulting in 18 000 artificially generated datasets.

3 Methods and training setup

A total of 5 classifiers were chosen for the simulation study: logistic regression (LR), XGBoost, random forest (RF), Gaussian naive Bayes (GNB) and neural network (NN). Logistic regression with the weight of evidence

Scenario	Multicollinearity	Modification of DGP
1	×	-
2	✓	-
3	✓	Time dependence
4	✓	Normally distributed noise
5	✓	Omitted variable
6	✓	Endogeneity
7	✓	Heteroskedasticity
8	✓	Nonnormally distributed predictors
9	✓	Nonlinear predictor
10	✓	Nonlinearity
11	✓	Noncausal predictor
12	✓	All possible modifications

Table 1 Overview of scenarios for the simulation study.

(WOE) transformation of the predictors is also included. All models are implemented in Python through the Scikit-Learn package Pedregosa et al, 2018, except for XGBoost, which is implemented through a package of the same name Chen and Guestrin, 2016. Training is done in the following steps:

1. each dataset is split into a training and a testing set using stratified random sampling and 80:20 ratio (prior to splitting, the data is shuffled),
2. for each model (except for logistic regression without WOE and GNB), 100 random combinations of hyperparameters are sampled,
3. models are trained using the sampled combinations of hyperparameters on the training set, which is further split into a smaller training set and a validation set using stratified random sampling and a 65:35 ratio,
4. the best combinations of hyperparameters are chosen based on evaluation on the validation set by maximising the area under the ROC curve Finlay, 2012,
5. models are retrained on the entire training set,
6. evaluation metrics are calculated.

The evaluation metric chosen for comparing the algorithms is the Gini coefficient (GC), which is calculated from the area under the ROC curve Finlay, 2012. Details of the implementation can be viewed in the project repository Tomsa, 2026.

4 Results

The median Gini coefficients are reported in Table 2. Across all scenarios there can be observed a substantial difference between the models for $N = 50$, which, except in scenario 10, diminishes with an increasing value of N. It is also apparent that the overall performance of the models is very similar across scenarios, apart from scenarios 6, 7, 10 and 12. In scenario 6, the overall increased predictive power is certainly caused by correlating the predictor with the noise. In scenario 7, it is the added "confidence" of making the variance of the noise dependent on x_{i4} (which diminishes the noise around zero values of x_{i4}). In scenario 10, it is caused by using the nonlinear function for the generation of the predicted value, which is not necessarily comparable to the linear function used in other scenarios. The performance in scenario 12 is affected by the combination of all of these modifications to the DGP and also perhaps by the other modifications. In all of the scenarios for $N = 50$, except for scenario 10, the median is highest for LR. The medians then only increase very slowly with increasing N compared to other models, on average by approx. 0.044 between $N = 50$ and $N = 5000$, but they still remain the highest for the scenarios with linear function, excluding scenario 7. In scenarios 7, 10 and 12 for $N = 5000$ it can be seen that the performance is worse compared to some or all the other models, especially in scenario 10, where the difference between the median of LR and the highest median is approx. 0.16. For $N = 50$, GNB usually has the second-highest median and even the highest median of 0.417 in scenario 10. GNB is then followed by RF and then the other models. With increasing N, NN and XGBoost see a substantial increase in their medians, on average by approx. 0.2 between $N = 50$ and $N = 5000$, which for $N = 5000$ are either just under the medians of LR for scenarios with the linear function or they are the highest for scenarios with the nonlinear function in the DGP. With increasing N, RF also sees a substantial increase in its median, and although for the most part it is worse than NN and XGBoost, it

reaches comparable results in the presence of heteroskedasticity. LR with WOE has the lowest median for $N = 50$ in all scenarios, except for scenario 10. Nevertheless, its median gets closer to other classifiers with increasing N, even for scenarios with the nonlinear function in the DGP.

Model	N	1	2	3	4	5	6	7	8	9	10	11	12
GNB	50	0.500	0.583	0.500	0.524	0.583	0.667	0.619	0.520	0.500	0.417	0.554	0.429
	500	0.606	0.599	0.575	0.576	0.599	0.687	0.673	0.600	0.569	0.469	0.599	0.592
	5000	0.619	0.601	0.591	0.581	0.601	0.689	0.673	0.613	0.597	0.487	0.601	0.613
LR	50	0.583	0.583	0.524	0.583	0.600	0.750	0.714	0.619	0.583	0.333	0.583	0.500
	500	0.619	0.640	0.604	0.614	0.633	0.748	0.723	0.635	0.634	0.353	0.638	0.594
	5000	0.620	0.636	0.622	0.612	0.634	0.751	0.722	0.639	0.637	0.367	0.636	0.607
LR (WOE)	50	0.333	0.375	0.333	0.347	0.417	0.476	0.429	0.333	0.375	0.286	0.333	0.327
	500	0.542	0.566	0.540	0.540	0.561	0.680	0.671	0.539	0.552	0.447	0.566	0.573
	5000	0.599	0.614	0.600	0.592	0.612	0.728	0.723	0.615	0.615	0.512	0.614	0.659
NN	50	0.417	0.434	0.429	0.429	0.500	0.592	0.583	0.440	0.429	0.238	0.440	0.333
	500	0.588	0.621	0.572	0.598	0.612	0.732	0.730	0.614	0.606	0.438	0.613	0.603
	5000	0.615	0.631	0.617	0.609	0.630	0.746	0.779	0.636	0.633	0.524	0.629	0.690
RF	50	0.458	0.500	0.429	0.440	0.500	0.619	0.619	0.500	0.500	0.333	0.500	0.429
	500	0.550	0.578	0.533	0.556	0.575	0.700	0.709	0.575	0.579	0.473	0.576	0.626
	5000	0.594	0.615	0.598	0.593	0.612	0.733	0.759	0.617	0.616	0.515	0.613	0.689
XGBoost	50	0.360	0.429	0.399	0.429	0.440	0.583	0.583	0.458	0.429	0.333	0.440	0.417
	500	0.532	0.563	0.513	0.537	0.555	0.693	0.691	0.558	0.556	0.439	0.557	0.605
	5000	0.602	0.621	0.605	0.599	0.620	0.739	0.755	0.624	0.621	0.527	0.620	0.698

Table 2 Median Gini coefficient across scenarios and models.

The average ranks based on Gini coefficients are reported in Figure 1. It is evident that although the medians saw the biggest differences for $N = 50$, the average ranks see the biggest differences for $N = 5000$. This indicates that for $N = 5000$, even though the difference in Gini coefficients is minimal, the ranks stay consistent across datasets. On the other hand, for $N = 50$, it could suggest both greater variance and many tied ranks. For GNB, it is visible that its rank deteriorates with the inclusion of multicollinearity, which violates its key assumption James et al, 2021. The figure also shows that XGBoost is better in nonlinear scenarios than NN, which was also true in the analysis of medians. However, the difference in ranks between the models seems small, just as was the difference in medians. Otherwise, the average ranks confirm the analysis of the medians.

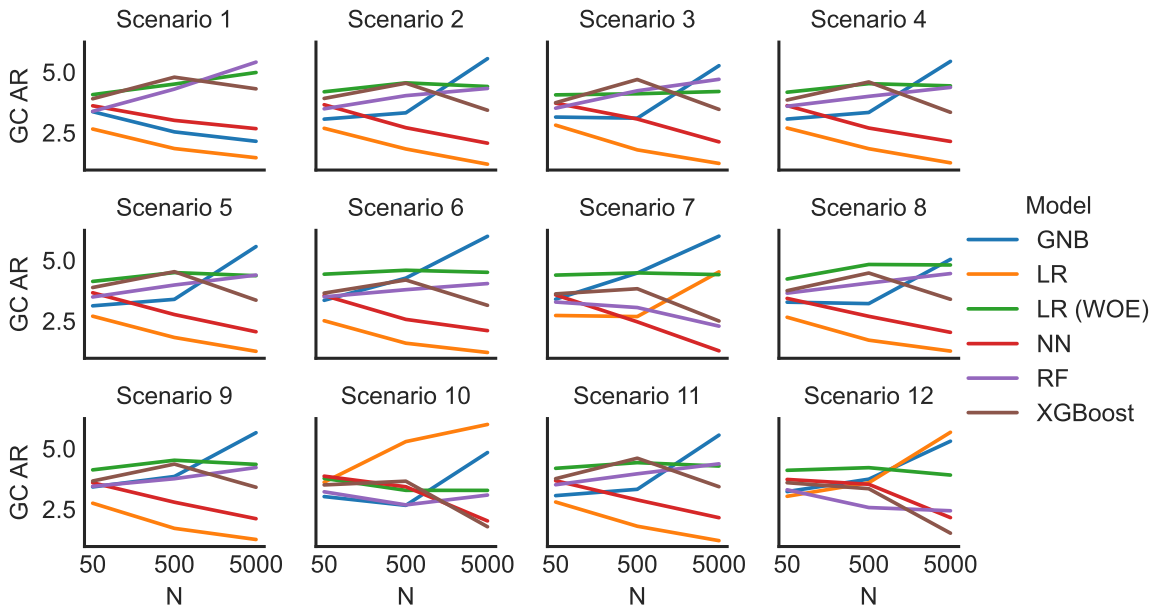


Figure 1 Average ranks (AR) based on GC. Tied ranks are averaged.

Figure 2 shows the percentages of achieving the highest Gini coefficient for every model. LR achieves the highest percentage for $N = 50$ in almost all scenarios, except for scenarios 10 and 12, where GNB achieves the highest percentage. In scenarios with the linear function, excluding scenario 7, the ratio of LR rises even more with increasing N and except for scenario 1, where both NN and GNB can somewhat compete with LR, only NN remains competitive. In scenarios with the nonlinear function, both NN and XGBoost have the highest percentages. In scenario 7, the highest ratio is achieved by NN. Other models usually achieve very low ratios.

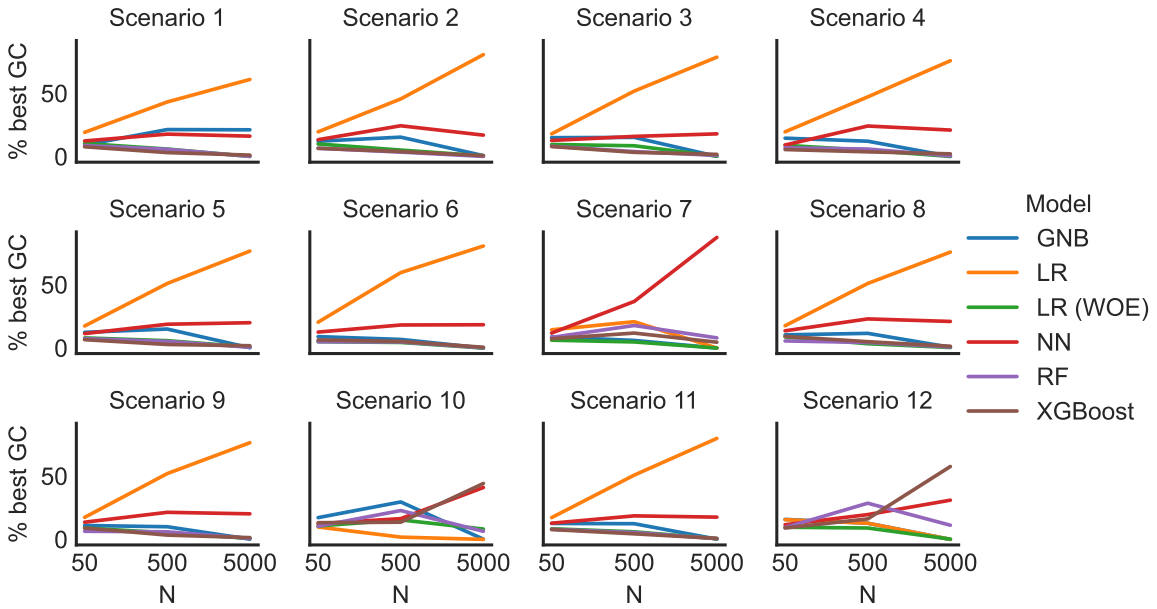


Figure 2 Percentages of achieving the best GC. Ties are counted towards all the respective models.

Finally, the models' performances were evaluated in every scenario using the Friedman test Sprent and Smeeton, 2001. For all scenarios and values of N, the p-values were lower than 1×10^{-10} and therefore show strong evidence against the null hypothesis that the models have the same predictive power, which further confirms the previous analysis.

5 Conclusion and discussion

This research aimed to compare LR with 4 other common classifiers through Monte Carlo simulations. It should be noted that the results are biased in favour of the LR due to the chosen DGP, which was selected for its interpretability. The results indicate that the main factors influencing the differences in predictive power between the classifiers are the number of observations, the inclusion of the (non)linear function for generating the predicted variable in the defined DGP and the presence of heteroskedasticity. The traditional methods, LR and GNB, seem to perform better than the other models when the linear function is used in the DGP and in the presence of homoskedasticity, or when the number of observations is low. The traditional methods also perform well with the inclusion of the variable with the nonlinear transformation in both the DGP and model training. On the other hand, the results show that XGBoost, NN and possibly RF perform better than the other models when there is a high number of observations, in the presence of heteroskedasticity and when the nonlinear function is used in the DGP. NN shows good performance, even when the linear function is used with high N, and can sometimes outperform LR. The results also give strong evidence that some models perform consistently better than the others, though the differences in median GC between the models are relatively low for high N.

It is hard to establish a common ground between the presented and the existing research due to different methodological approaches. Some benchmarks Scholz and Wimmer, 2021; Lessmann et al, 2015 identified heterogeneous ensemble classifiers as the best, though those were not included in this study. Fernández-Delgado et al, 2014 identified RF as the best algorithm, which does not correspond to the results of this study. A common conclusion from most of the reviewed benchmarks was that more complex classifiers, like RF, NN, and boosted trees, outperform classical statistical methods, such as LR Scholz and Wimmer, 2021; Lessmann et al, 2015; Fernández-Delgado et al, 2014; Caruana and Niculescu-Mizil, 2006. Although LR achieved very good results on scenarios with linear DGP and on datasets with low N, complex methods will most likely perform better on

real-world datasets with a larger sample size, since real-world problems are hardly going to be so simple (or strictly linear). However, at the very least, LR remains a useful baseline, which can be further improved using WOE for more complex problems. Some benchmarks Fernández-Delgado et al, 2014; Lessmann et al, 2015 advocate for using RF as the baseline algorithm for comparison against newly proposed methods. Based on the presented results, that is not recommendable in my opinion. Although RF achieved good results in a variety of scenarios, it did not necessarily perform the best. Furthermore, models behave differently under different conditions, as was shown in the results. Therefore, relying on only one model is greatly insufficient.

Future research should focus on adding more scenarios and including more models in the comparison. It is also desirable to research methods that would enable one to analyse datasets and choose the best model a priori if possible. Lastly, it would be beneficial to compare the costs associated with the training times of the models and to make methods of comparing models to each other more uniform and interpretable.

Acknowledgements

The work was supported by the Prague University of Economics and Business project IGS F4/29/2026.

References

Beram R, El-Kotory A (2024) Simulation-Based Assessment of Classification Methods: Statistical Models vs. Machine Learning Algorithms. The Egyptian Statistical Journal 68(1):91–124. https://doi.org/10.21608/esju.2024.260404.1025

Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery, New York, NY, USA, ICML '06, pp 161–168, https://doi.org/10.1145/1143844.1143865

Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp 785–794, https://doi.org/10.1145/2939672.2939785, 1603.02754

Fernández-Delgado M, Cernadas E, Barro S, et al (2014) Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research 15(1):3133–3181

Finlay S (2012) Credit Scoring, Response Modeling, and Insurance Rating, 2nd edn. Palgrave Macmillan UK, London, https://doi.org/10.1057/9781137031693

James G, Witten D, Hastie T, et al (2021) An Introduction to Statistical Learning: With Applications in R, 2nd edn. Springer Texts in Statistics, Springer US, New York, https://doi.org/10.1007/978-1-0716-1418-1

Lessmann S, Baesens B, Seow HV, et al (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 247(1):124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Orzechowski P, Moore JH (2021) Generative and reproducible benchmarks for comprehensive evaluation of machine learning classifiers. https://doi.org/10.48550/arXiv.2107.06475, 2107.06475

Pedregosa F, Varoquaux G, Gramfort A, et al (2018) Scikit-learn: Machine Learning in Python. https://doi.org/10.48550/arXiv.1201.0490, 1201.0490

Scholz M, Wimmer T (2021) A comparison of classification methods across different data complexity scenarios and datasets. Expert Systems with Applications 168:114217. https://doi.org/10.1016/j.eswa.2020.114217

Sprent P, Smeeton NC (2001) Applied Nonparametric Statistical Methods, 3rd edn. Chapman & Hall/CRC Texts in Statistical Science Series, Chapman & Hall/CRC, Boca Raton

Tomsa K (2026) Synthetic-classifier-benchmark. https://github.com/KTomsa/synthetic-classifier-benchmark

Wainberg M, Alipanahi B, Frey B (2016) Are Random Forests Truly the Best Classifiers? Journal of Machine Learning Research 17(110):1–5

Predicting Realized Volatility of Bitcoin and Ethereum with Sample Entropy

Quang Van Tran[1], Karel Brůna[2]

Abstract. This study uses hourly price data for Bitcoin and Ethereum from 2018 to 2025 to compute sample entropy over rolling windows, which is then employed to forecast next-day realized volatility. We find a positive and statistically significant relationship between sample entropy and realized volatility for both cryptocurrencies, indicating that higher market complexity is associated with increased future volatility. The predictive power of sample entropy varies depending on the choice of rolling window and varies over time, reflecting evolving market conditions. Overall, our findings demonstrate that chaos-based measures can be dynamically utilized to predict realized volatility and suggest potential applications to other economic and financial variables.

Keywords: Sample entropy, Realized volatility, Cryptocurrencies, Prediction

JEL Classification: C530, G170

1 Introduction

The volatility of financial asset returns has been modeled using a wide range of approaches, including GARCH models (Van Tran et al, 2026), stochastic volatility models (Heston, 1993), implied volatility measures (Black and Scholes, 1973), machine learning–based methods (Donaldson and Kamstra, 1997), and realized volatility frameworks (Andersen et al, 2003). However, there is relatively little systematic evidence on the use of chaos-based measures, such as entropy or Lyapunov exponents, as predictive inputs for the volatility of financial asset returns. This study helps fill this gap by exploring sample entropy as a predictor of daily realized volatility in cryptocurrency markets, which are characterized by pronounced nonlinearities, frequent regime shifts, and potentially chaotic dynamics. Using hourly price data for Bitcoin and Ethereum - the two most actively traded cryptocurrencies - from 2018 to 2025, we compute sample entropy over rolling intraday windows and examine its ability to forecast next-day realized volatility. The use of high-frequency intraday data provides sufficient granularity to support reliable entropy estimation, while the focus on daily realized volatility aligns with economically meaningful risk horizons for investors and risk managers. This multi-scale framework allows us to assess whether chaos-based metrics can serve as valuable complementary tools in volatility forecasting for digital asset markets.

2 Sample entropy

In statistics and information theory, entropy measures the degree of unpredictability, uncertainty, or randomness associated with the outcomes of a random variable, abstracting from the temporal ordering of observations. Sample entropy (SAE) is designed for time-series data and captures structural complexity and predictability (Richman and Moorman, 2000). Lower values indicate greater regularity, while higher values correspond to increased randomness. Sample entropy is less sensitive to sample length and mitigates bias by excluding self-matches. Sample entropy of a time series x of length n $SAE(m, r, n)$ is defined as:

$$SAE(m, r, n) = -\log\left(\frac{A}{B}\right), \tag{1}$$

where $m < n$ is the embedding dimension, $r > 0$ is the tolerance parameter. The quantities A and B are given by

$$A = \frac{(n-m-1)(n-m)}{2} A^m(r) \tag{2}$$

$$B = \frac{(n-m-1)(n-m)}{2} B^m(r) \tag{3}$$

[1] Prague University of Economics and Business, Faculty of Finance and Accounting, Department of Monetary Theory and Policy, Nam W. Churchilla 1, Praha 3, Czech Republic, tran@vse.cz

[2] Prague University of Economics and Business, Faculty of Finance and Accounting, Department of Monetary Theory and Policy, Nam W. Churchilla 1, Praha 3, Czech Republic, bruna@vse.cz

where $A^m(r)$ denotes the probability that two sequences match in $m+1$-dimensional space, and $B^m(r)$ is the probability that two sequences match in m-dimensional space. To compute these probabilities, the one-dimensional time series $x$is embedded into an m-dimensional space using phase space reconstruction technique (Kantz and Schreiber, 2003), yielding $n-m$ vectors $X_1, ..., X_{n-m}$. The distance between two vectors X_i and X_j is defined as

$$d(X_i, X_j) = \max_{k=1,...,m} (|X_i^{(k)} - X_j^{(k)}|). \tag{4}$$

For each vector X_i, the proportion of vectors $X_j, j \neq i$ satisfying $d(X_i, X_j) \leq r$ is

$$B_i^m(r) = \frac{1}{n-m-1} \sum_{j=1, j\neq i}^{n-m} \mathbf{1}_{(d \leq r)}, \tag{5}$$

where $\mathbf{1}_{(.)}$ is the indicator function. The empirical probability of a match in m-dimensional space is then

$$B^m(r) = \frac{1}{n-m} \sum_{i=1}^{n-m} B_i^m(r). \tag{6}$$

Repeating the procedure in $m+1$-dimensional space yields $A^m(r)$. By construction, $A^m(r) \leq B^m(r)$.

3 Data

This study uses hourly cryptocurrency price data obtained from the CryptoCompare API, covering the period from 1 January 2018 to 15 July 2025. Hourly returns capture intraday dynamics that are essential for modeling volatility persistence and nonlinear behavior. At the same time, they are less affected by market microstructure noise than ultra-high-frequency data, such as tick-by-tick prices, which are often problematic in cryptocurrency markets due to fragmented liquidity across exchanges. The two largest cryptocurrencies by market capitalization, Bitcoin (BTC) and Ethereum (ETH), are used in this analysis. As cryptocurrency markets operate continuously, hourly data are very suitable for high-frequency volatility analysis. Both assets exhibit long and highly liquid trading histories, ensuring reliable intraday price observations and limiting potential microstructure effects. The dataset is complete, with no missing observations. The descriptive statistics of the original price series and their corresponding return series are shown in Table 1. The price series in U.S. dollars are transformed into hourly logarithmic return series.

	BITCOIN		**ETHEREUM**	
Characteristic	**Price Series**	**Return series**	**Price Series**	**Return series**
Mean	32963.08	3.23E-05	1576.96	2.12E-05
Median	25793.23	6.52E-05	1600.56	7.31E-05
Mode	3919.13	0	136.20	0
Minimum	3183.27	-0.167	82.48	-0.221
Maximum	122751.24	0.184	4848.03	0.170
1st Quartile	9247.82	-0.002	289.58	-0.003
3rd Quartile	48595.65	0.003	2523.57	0.004
Std. deviation	27868.53	0.007	1219.56	0.009
Skewness	1.05	-0.101	0.40	-0.374
Kurtosis	3.26	37.096	2.03	28.350
Num of Obs	66061	66060	66061	66060

Table 1 Descriptive statistics of Bitcoin and Ethereum

We use the hourly returns to generate the daily realized volatility series RV_t according to formula

$$RV_t = \sum_{i=1}^{24} r_{i,t}^2, \tag{7}$$

where $r_{i,t}$ is the hourly return of a cryptocurrency in hour i of day t. From the dataset available in the form OHLCV we choose the highest and lowest values of the day (H_t and L_t) to compute the proxy volatility (SP_t) defined as follows (Parkinson, 1980):

$$SP_t = \frac{[\log(H_t/L_t)]^2}{4\log(2)}. \tag{8}$$

We apply the algorithm for computing sample entropy described in the previous section to the hourly return series of the two cryptocurrencies, thereby generating corresponding sample entropy series. In addition, we incorporate trading volume data denominated in U.S. dollars. As trading volume is several orders of magnitude larger than realized volatility, we apply a double logarithmic transformation to reduce its scale (llVolume). All series are tested for stationarity using the Augmented Dickey–Fuller (ADF) test. The test results indicate that all series are stationary.

4 Results and Discussion

We compute sample entropy from the hourly return series using the algorithm described in Section 2. As $SAE(m, r, n)$ depends on three parameters m, r, n, they must be specified appropriately. The embedding dimension is selected according to (Kantz and Schreiber, 2003), yielding $m = 3$. This approach ensures that the reconstructed phase space captures the essential nonlinear dynamics of the return process without over-embedding. The tolerance parameter is set to the commonly used value recommended by (Lake et al, 2002), $r = 0.2\sigma$, where σ denotes the sample standard deviation of returns. The window length is fixed at $n = 168$, corresponding to a weekly trading horizon. This choice balances the need for reliable estimation of nonlinear dependence with the preservation of local market dynamics.

We use a rolling-window approach to compute sample entropy for the hourly return series of Bitcoin and Ethereum, allowing for time-varying measures of market complexity. The window is advanced by 24 observations at each step, corresponding to one trading day and ensuring alignment with daily realized volatility. This procedure yields overlapping entropy estimates. All computations were performed in MATLAB.

Following the heterogeneous autoregressive realized volatility (HAR-RV) framework (Corsi, 2009), we model daily realized volatility as a function of its own lagged components and additional explanatory variables. Specifically, realized volatility on day t is predicted using lagged realized volatility, sample entropy, proxy volatility, and transformed trading volume, all measured at $t - 1$ to avoid look-ahead bias. To capture potential leverage and downside-risk effects, we allow for asymmetric responses of the explanatory variables, excluding proxy volatility, conditional on negative returns on the previous day. Accordingly, the general regression specification is given by:

$$RV_t = c + (\alpha_1 + \kappa\alpha_2)SAE_{t-1} + \beta SP_{t-1} + (\gamma_1 + \kappa\gamma_2)RV_{t-1} + (\delta_1 + \kappa\delta_2)V_{t-1} + e_t, \tag{9}$$

where

$$\kappa = \begin{cases} 1 \text{ if } r_{t-1} < 0 \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

and

$$r_t = \sum_{i=1}^{24} r_{i,t}. \tag{11}$$

We use the prepared data to estimate the parameters of model (9). The results indicate no statistically significant asymmetric effect associated with sample entropy, while the transformed trading volume exhibits a statistically significant asymmetric response conditional on negative returns. For brevity, we report only the most statistically significant results for Bitcoin in Table 2 and for Ethereum in Table 3[1].

Variable	Coefficient	S.E.	t-stat	p-value
c	-5.45E-04	2.82E-04	-1.930	0.054
SAE168(-1)	1.75E-03	4.24E-04	4.120	3.91E-05
SP(-1)	1.150	0.074	15.622	0
RV(-1)	0.006	0.025	0.237	0.812
RVa(-1)	0.609	0.030	20.578	0
llVolumea(-1)	-2.42E-04	4.03E-05	-5.999	2.24E-09
	R = 0.4140		DW = 1.9348	

Table 2 Estimation results with window length $l = 168$ for Bitcoin

[1] All partial estimation results are available upon request.

Variable	Coefficient	S.E.	t-stat	p-value
c	-1.05E-03	4.86E-04	-2.168	0.030
SAE168(-1)	2.81E-03	6.92E-04	4.065	4.94E-05
SP(-1)	0.894	0.073	12.173	0
RV(-1)	0.105	0.028	3.707	2.14E-04
RVa(-1)	0.421	0.031	13.527	0
llVolumea(-1)	-1.53E-04	5.81E-05	-2.633	8.51E-03
	R =0.3552		DW = 1.9559	

Table 3 Estimation results with window length $l = 168$ for Ethereum

The estimation results reported in Tables 2 and 3 show that the coefficients on sample entropy are highly statistically significant for both Bitcoin and Ethereum. Their magnitudes are relatively small, reflecting the fact that sample entropy is several orders of magnitude larger than realized volatility. The estimated coefficients are positive, indicating that higher return instability on the previous day, captured by sample entropy, is associated with higher realized volatility today.

The lagged realized volatility exhibits strong persistence in both markets, with economically and statistically significant effects on current volatility, although its average impact is weaker for Bitcoin. Proxy volatility also displays a large and significant effect, underscoring its importance as a predictor of realized volatility.

The asymmetric component of transformed trading volume enters with a negative coefficient, suggesting that following negative returns, subsequent volatility tends to decline. The adjusted R-squared ranges from 0.35 to 0.42, indicating that a substantial fraction of volatility variation is explained by the model. Finally, the Durbin-Watson statistics are close to two in both cases, suggesting no evidence of residual autocorrelation. Robust standard errors may therefore not materially differ from conventional estimates. Taken together, the results indicate that time-varying informational complexity and market activity contribute meaningfully to the dynamics of cryptocurrency volatility, complementing traditional persistence-based volatility effects.

Robustness diagnostics

First, we examine the robustness of our choice of the rolling window length used to estimate sample entropy. As alternatives to our baseline length $l = 168$, we consider $l = 72$ and $l = 336$, corresponding to a three-day and a two-week trading span, respectively. For $l = 72$, sample entropy is computed using the algorithm described in Section 2, with the same embedding dimension $m = 3$ and tolerance parameter $r = 0.2\sigma$. In the case of Bitcoin, we identify five periods in which sample entropy takes an infinite value, while for Ethereum this occurs in 17 periods (out of a total of 2,752 periods). An infinite value of sample entropy indicates that the analyzed data sequence is completely random and unpredictable, exhibiting no repetitive patterns. These results suggest that when the rolling window is too short, the estimation may yield unusable entropy values, thereby limiting its applicability in subsequent analysis. Consequently, a shorter rolling window appears to be inappropriate for computing sample entropy in this context.

Next, we compute sample entropy using a longer rolling window of $l = 336$, denoted as SAE336, while maintaining the same parameter values for m and r. The estimation proceeds without computational issues. On average, SAE336 values are lower than those obtained using SAE168. This outcome is intuitive, as longer time windows increase the likelihood of observing recurring patterns, thereby reducing measured irregularity. The evolution of SAE over time, calculated using two different window lengths, is shown in Figure 1. We then replace SAE168 with SAE336 in estimating the relationship between realized volatility and its predictors as specified in equation (9). The estimated coefficients on SAE336 are positive and strongly statistically significant for both cryptocurrencies, indicating that greater return irregularity on the previous day is associated with higher realized volatility today. However, replacing SAE168 with SAE336 alters the estimated coefficients of the model and, most importantly, leads to a reduction in predictive power as measured by the R^2 statistic. These findings support our baseline choice of $l = 168$ as the optimal rolling window length.

Next, we investigate the time-varying nature of the effect of the return irregularity measured by the sample entropy on today's realized volatility. As the time span of the dataset is large, covering two major events, such as the COVID pandemic and the war in Ukraine, we want to investigate how the impact of the sample entropy on realized volatility changes in the two subperiods with the COVID pandemic and the Ukrainian war. For this purpose, we

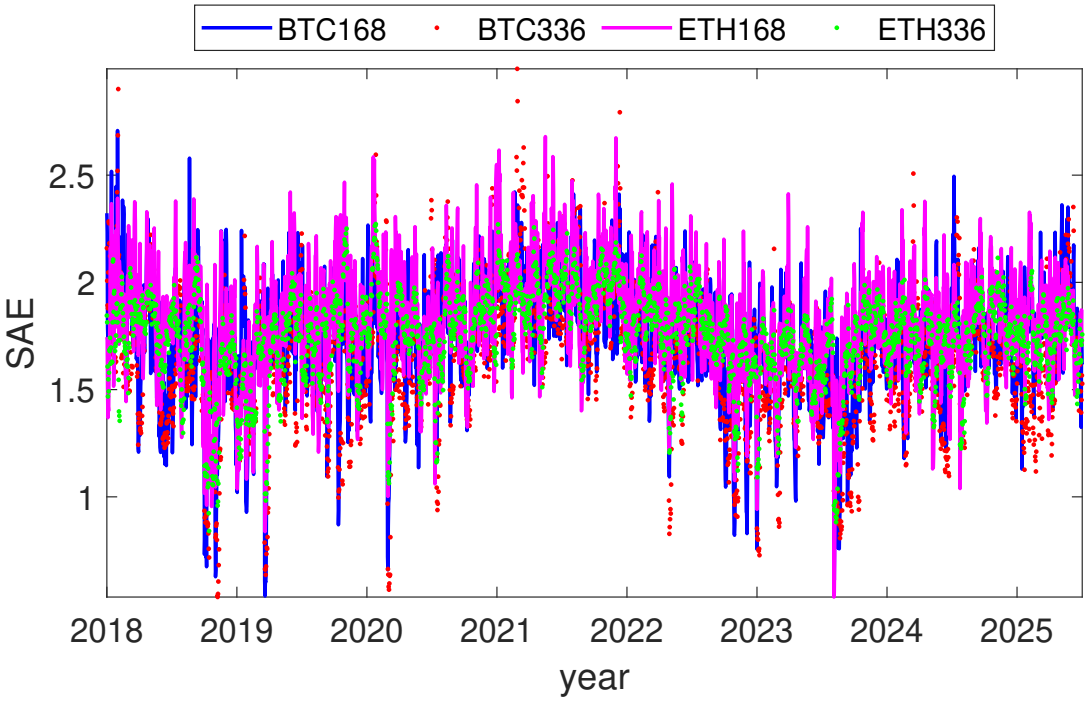


Figure 1 The evolution of sample entropy of two series over time

set the term with SAE168 as

$$RV_t = c + (\alpha_0 + \sum_{i=1}^{2} \kappa_i \alpha_i) SAE_{t-1} + ... + e_t, \tag{12}$$

where

$$\kappa_i = \begin{cases} 1 \text{ if the event is present} \\ 0 \text{ otherwise,} \end{cases} \tag{13}$$

$i = 1, 2$, $i = 1$ denoting the COVID-19 subperiod and $i = 2$ denoting the Ukrainian war subperiod. The subperiod with COVID pandemic is from 11 March 2020 to 31 December 2021[2] and the Ukrainian war subperiod is from 24 February 2022 to 15 July 2025 (the end date of the dataset). Using SAE168 and the remaining predictors, we estimate the modified model. The inclusion of the interaction term does not materially affect the estimated impacts of the other explanatory variables and slightly improves the model's predictive performance as measured by the R^2. More importantly, the results reveal a statistically significant and negative change in the effect of sample entropy on realized volatility during crisis periods. This reduction is more pronounced during the Ukrainian war subperiod than during the COVID-19 pandemic. These findings suggest that in times of heightened market stress, the influence of return irregularity on realized volatility is weakened.

Finally, to ensure valid statistical inference, we apply heteroskedasticity and autocorrelation consistent (HAC) standard error corrections on the extended specification of model (9). Given that regression residuals in time-series models often exhibit serial correlation and heteroskedasticity, this procedure improves the robustness of inference without altering coefficient estimates. The parameters of the prewhitening procedure are specified as follows: length lags $k = 24$, Quadratic-Spectral kernel, and bandwidth h determined using the Andrews automatic selection method (Andrews, 1991). The HAC-corrected results indicate a weakened impact of transformed trading volume and of realized volatility persistence in the absence of price growth. Importantly, even after applying the correction, the effect of sample entropy remains strong and statistically significant, reinforcing the robustness of our main findings.

Overall, these robustness and subperiod analyses highlight the dual role of sample entropy as both a stable and context-dependent predictor of realized volatility in cryptocurrency markets. While the positive relationship between return irregularity and future volatility remains statistically robust across alternative specifications and estimation procedures, its magnitude depends critically on the choice of the rolling window and on prevailing

[2] The World Health Organization declared the outbreak a Public Health Emergency of International Concern on 30 January 2020, and it was characterized as a pandemic on 11 March 2020. Although the WHO declared the public health emergency caused by COVID-19 to have ended in May 2023, we adopt a more conservative approach and consider the pandemic period to have ended on 31 December 2021.

market conditions. The superior predictive performance of SAE168 reflects a trade-off between capturing sufficient temporal structure and avoiding excessive smoothing, while the attenuated impact during crisis periods suggests that extreme market-wide shocks may dominate endogenous measures of return complexity. Taken together, these findings underscore the importance of carefully calibrating entropy-based measures and accounting for regime-dependent dynamics when modeling volatility in highly turbulent and evolving financial markets.

5 Conclusion

This study examines whether sample entropy, hourly returns derived from Bitcoin and Ethereum prices from 2018 to 2025, can predict their daily realized volatility. Our results show that sample entropy is both a statistically and economically significant predictor of future realized volatility. We find a robust positive relationship between sample entropy and next-day volatility, indicating that higher market complexity and irregularity are associated with increased volatility risk. The predictive power of sample entropy depends on the length of the rolling window used for its estimation and varies over time, reflecting changing market conditions. A one-week rolling window provides the best balance between information content and smoothing, yielding the strongest forecasting performance. These findings demonstrate that chaos-based measures can be dynamically incorporated into volatility forecasting frameworks for cryptocurrency markets. This approach complements traditional volatility predictors and suggests that sample entropy contains valuable information relevant for risk management. Moreover, the framework may be extended to other financial assets and economic variables characterized by pronounced nonlinear dynamics.

Acknowledgements

The authors acknowledge the Institutional Support IP 100040 of the Prague University of Economics and Business and the financial support of the grant 24-10008S of the Czech Science Foundation.

References

Andersen TG, Bollerslev T, Diebold FX, et al (2003) Modeling and forecasting realized volatility. Econometrica 71(2):579–625

Andrews DWK (1991) Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica pp 817–858

Black F, Scholes M (1973) The pricing of options and corporate liabilities. Journal of Political Economy 81(3):637–654

Corsi F (2009) A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics 7(2):174–196

Donaldson GR, Kamstra M (1997) An artificial neural network-GARCH model for international stock return volatility. Journal of Empirical Finance 4(1):17–46

Heston SL (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. The Review of Financial Studies 6(2):327–343

Kantz H, Schreiber T (2003) Nonlinear Time Series Analysis. Cambridge University Press

Lake DE, Richman JS, Griffin PM, et al (2002) Sample entropy analysis of neonatal heart rate variability. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology 283(3):R789–R797

Parkinson M (1980) The extreme value method for estimating the variance of the rate of return. The Journal of Business pp 61–65

Richman JS, Moorman RJ (2000) Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology 278(6):H2039–H2049

Van Tran Q, Molnár P, Sensoy A (2026) Unconditional and conditional heavy-tailed distributions for the returns of cryptocurrencies with a novel range exponential GARCH model. Borsa Istanbul Review p 100803

Convergence and Differentiation in Slovak Spa Enterprises: A Cluster Analysis of Financial Indicators

Petra Vašaničová[1], Sylvia Jenčová[2], Martina Košíková[3], Marta Miškufová[4], Kateryna Melnyk[5]

Abstract. This study examines the financial performance of spa enterprises in Slovakia using cluster analysis applied to selected financial indicators across five dimensions: profitability (Return on Assets, Return on Sales), efficiency (Assets Turnover), cost efficiency (Personnel Cost Ratio, Cost-to-Sales Ratio, Return on Costs), liquidity (Net Working Capital, Current Ratio), and indebtedness (Equity to Total Liabilities Ratio, Debt Ratio). The analysis was conducted separately for the years 2023 and 2024 in order to identify changes in the internal structure of the sector. The results reveal a reduction in the number of clusters from six in 2023 to five in 2024, indicating a partial convergence of financial characteristics among spa enterprises. A substantial proportion of spas shifted into a single dominant cluster in 2024, suggesting increasing similarity in their financial profiles. At the same time, a distinct subgroup of spa enterprises emerged in 2024, forming a separate cluster after shifting from the previously dominant cluster in 2023, indicating differentiated financial trajectories within the sector. Several entities maintained stable positions across both years, while certain spas consistently formed individual clusters, reflecting unique financial characteristics. The findings suggest a trend toward homogenization within the Slovak spa sector, accompanied by the persistence of specific outliers and emerging subgroups. These insights contribute to a better understanding of the financial dynamics and structural evolution of spa enterprises and may support strategic decision-making and sectoral policy development.

Keywords: spa enterprises, tourism, financial performance, financial indicators, cluster analysis

JEL Classification: L83, C38, G39

1 Introduction

Tourism and its related hospitality industries are among the key drivers of economic growth and the development of national economies (Agag et al., 2025). Within this sector, health and spa tourism has gained particular importance, as it represents a dynamically growing segment with a high potential for revenue generation and for enhancing the competitiveness of destinations (Torkzadeh et al., 2024). Spa enterprises face increasing demands for quality, innovation, and financial stability, while their performance is also influenced by external shocks and changing industry conditions (Torres-Pruñonosa et al., 2022; Milojević et al., 2025).

The financial performance of tourism enterprises is a multidimensional concept that requires the simultaneous consideration of profitability, liquidity, efficiency, and indebtedness (Tutar et al., 2025). Given the heterogeneity of enterprises, cluster analysis is an appropriate analytical tool, as it enables the identification of homogeneous groups based on financial characteristics (Grande et al., 2024). Despite the growing interest in financial performance in tourism, less attention has been paid to the dynamics of changes in financial structures over time. Most existing studies focus on static evaluation, while the evolution of cluster structures, convergence, and differentiation of firms' financial profiles remain less explored (Milojević et al., 2025). Particularly in the post-pandemic

[1] University of Prešov, Faculty of Management and Business, Department of Finance, Accounting and Mathematical Methods, Konštantínova 16, 080 01 Prešov, Slovakia, petra.vasanicova@unipo.sk.
[2] University of Prešov, Faculty of Management and Business, Department of Finance, Accounting and Mathematical Methods, Konštantínova 16, 080 01 Prešov, Slovakia, sylvia.jencova@unipo.sk.
[3] University of Prešov, Faculty of Management and Business, Department of Finance, Accounting and Mathematical Methods, Konštantínova 16, 080 01 Prešov, Slovakia, martina.kosikova@unipo.sk.
[4] University of Prešov, Faculty of Management and Business, Department of Finance, Accounting and Mathematical Methods, Konštantínova 16, 080 01 Prešov, Slovakia, marta.miskufova@unipo.sk.
[5] University of Prešov, Faculty of Management and Business, Department of Finance, Accounting and Mathematical Methods, Konštantínova 16, 080 01 Prešov, Slovakia, kateryna.melnyk@smail.unipo.sk.

period, it is important to examine whether the financial characteristics of enterprises are converging or, conversely, whether distinct groups are emerging (Wang et al., 2026; Wu et al., 2025). This paper addresses this gap by extending the analysis of the financial performance of spa enterprises to include a temporal dimension. It focuses on the period 2023–2024 and examines whether convergence of financial profiles is occurring in the Slovak spa sector, or whether differentiated groups of enterprises persist. The aim of this study is to identify and compare groups of spa enterprises in Slovakia based on their financial performance using cluster analysis, and to examine changes in the structure of these groups between 2023 and 2024. This study builds on the methodological approach of Vašaničová et al. (2025), who applied cluster analysis to 20 spa enterprises in Slovakia using key financial indicators for the years 2018 and 2023. In line with this study, the same set of financial indicators is adopted, covering profitability, efficiency, cost efficiency, liquidity, and indebtedness dimensions. This ensures methodological comparability and allows for a consistent interpretation of financial structures within the spa sector over time.

2 Material and Methods

2.1 Data

Currently, the Slovak Republic has 25 spa locations, comprising a total of 31 spa enterprises. This study focuses on those operating as joint-stock companies, limited liability companies, or state-owned enterprises. The term spa enterprise is used consistently, as it more accurately reflects the legal and economic diversity of the sector. Unlike the term company, which typically refers to privately owned entities, enterprise encompasses the full range of organizational and ownership structures present in the Slovak spa industry. Out of the 31 active spa entities, only 19 had complete data available for the period 2023–2024 and were therefore included in the analysis. These are Bardejov Spa (Bardejov), Natural Iodine Spa Číž (Číž), Spa Dudince (Dudince), Spa Kováčová (Kováčová), Spa Bojnice (Bojnice), Spa Horný Smokovec (H. Smokovec), Spa Lúčivná (Lúčivná), Spa Lúčky (Lúčky), Spa Nový Smokovec (N. Smokovec), Spa Vyšné Ružbachy (Ružbachy), Specialized Medical Institute Marína (Marína), Spa Nimnica (Nimnica), Slovak Health Spa Piešťany (Piešťany), Slovak Health Spa Rajecké Teplice (R. Teplice), Spa Sliač (Sliač), Slovthermae Spa Diamant Dudince (D. Dudince), Spa Štós (Štós), Spa Trenčianske Teplice (Tre. Teplice), and Slovak Health Spa Turčianske Teplice (Tur. Teplice).

The selected financial indicators were intended to capture the key dimensions of corporate performance, namely profitability, efficiency, cost efficiency, liquidity, and indebtedness. Profitability was represented by return on assets (ROA) and return on sales (ROS), while efficiency was measured using asset turnover (AT). Cost efficiency was assessed through personnel cost ratio (PCR), defined as personnel costs to total costs, cost-to-sales ratio (CSR), and return on costs (ROC). Liquidity was evaluated by net working capital (NWC) and the current ratio (CR). Indebtedness was expressed by the equity to total liabilities ratio (E/TL) and the debt ratio (DR). This framework captures the essential aspects of financial performance while aligning with established approaches in the literature (Vašaničová et al., 2025), thereby supporting the validity and interpretability of the results. The analysis is based on data for the years 2023 and 2024, obtained from publicly available financial statements recorded in the Register of Financial Statements maintained by the Ministry of Finance of the Slovak Republic.

2.2 Cluster Analysis

From a methodological perspective, this study applies cluster analysis to reduce data complexity by grouping objects based on multiple variables. A hierarchical approach is used, specifically Ward's method, which forms relatively homogeneous clusters of similar size. The results are visualized using a dendrogram. The optimal number of clusters is determined using the Duda–Hart $Je(2)/Je(1)$ index together with the pseudo-T^2 statistic. These measures identify the point at which further merging no longer improves cluster quality. The preferred solution is indicated by a high Duda–Hart index and a low pseudo-T^2 value, reflecting well-separated clusters.

Prior to the hierarchical cluster analysis, all variables were standardized using z-scores (mean = 0, standard deviation = 1) in order to eliminate differences in measurement scales across indicators. Moreover, correlation analysis is performed prior to clustering to identify highly correlated variables, which may introduce redundancy and distort distance calculations, potentially biasing the resulting cluster structure. All analyses are performed in Stata.

3 Results

Correlation analysis revealed several strong relationships among the variables, particularly between profitability indicators (ROA, ROS, ROC), with correlation coefficients reaching high values (e.g., $r \approx 0.9–1.0$). Similarly, strong positive relationships were identified among financial structure variables (liquidity and indebtedness) such

as E/TL, NWC, and CR (r ≈ 0.8). In contrast, the cost efficiency indicator (CSR) shows strong negative correlations with performance measures (e.g., r ≈ −0.7 to −1.0). These findings indicate the presence of multicollinearity and suggest that some variables capture similar underlying dimensions. Despite this, all variables were retained in the cluster analysis. The decision was based on their distinct economic interpretation and the aim to capture the multidimensional financial profile of spa enterprises. Retaining correlated variables allows the clustering procedure to reflect the combined effect of related financial characteristics, even if certain dimensions exert a stronger influence on the resulting cluster structure.

As Ward's method resulted in a relatively fragmented cluster structure, the Duda–Hart *Je*(2)/*Je*(1) index was applied to identify the optimal number of clusters. The index was evaluated for solutions ranging from one to ten clusters. The preferred solution is indicated by a high value of the index in combination with a low pseudo-T^2 statistic. Based on these criteria (see Table 1), six clusters were identified for 2023 and five clusters for 2024. The corresponding dendrograms are shown in Figure 1. This approach ensures that the final cluster solution is supported by statistical evidence rather than being determined arbitrarily or solely by the clustering algorithm, while also maintaining interpretability of the resulting groups.

	Number of clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2023	***Je*(2)/*Je*(1)**	0.705	0.755	0.582	0.613	0.000	**0.565**	0.542	0.146	0.531	0.442
	pseudo-T^2	7.13	4.88	6.47	5.05	.	**3.85**	3.38	5.85	3.53	2.53
2024	***Je*(2)/*Je*(1)**	0.639	0.595	0.588	0.394	**0.630**	0.450	0.165	0.510	0.545	0.493
	pseudo-T^2	9.61	4.08	6.31	4.61	**4.70**	2.44	5.07	2.89	2.50	2.06

Note: Values in bold indicate the most significant breakpoint.

Table 1 Duda-Hart Index and Pseudo-T^2

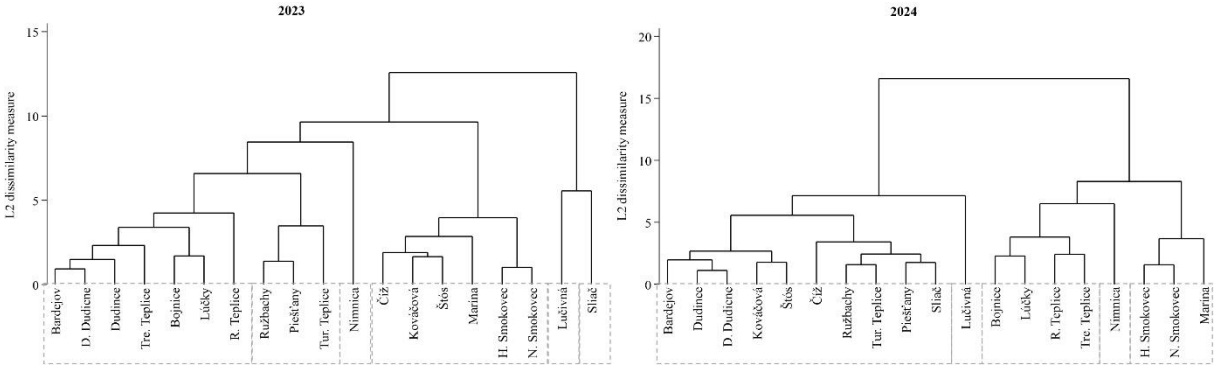


Figure 1 Dendrograms for Analyzed Years

Table 2 presents cluster assignments of Slovak spa resorts for 2023 and 2024. Since the clustering procedure was conducted separately for each year, cluster labels are not directly comparable across periods. To improve interpretability, color shading is used to highlight spa resorts that remain grouped together across both years, even if their assigned cluster labels differ.

Spa	**Bardejov**	**D. Dudince**	**Dudince**	**Tre. Teplice**	**Bojnice**	**Lúčky**	**R. Teplice**
2023	1	1	1	1	1	1	1
2024	1	1	1	3	3	3	3
Spa	**Ružbachy**	**Piešťany**	**Tur. Teplice**	**Nimnica**	**Číž**	**Kováčová**	**Štós**
2023	2	2	2	3	4	4	4
2024	1	1	1	4	1	1	1
Spa	**Marína**	**H. Smokovec**	**N. Smokovec**	**Lučivná**	**Sliač**		
2023	4	4	4	5	6		
2024	5	5	5	2	1		

Table 2 Cluster Membership Changes and Stability of Slovak Spa Enterprises (2023–2024)

The results of the cluster analysis indicate a change in the structure of spa enterprises between 2023 and 2024. While six clusters were identified in 2023, their number decreased to five in 2024, suggesting a partial consolidation and convergence of the financial characteristics of the analyzed entities. This trend is most evident in the emergence of a dominant cluster 1 in 2024, which includes more than half of the spa enterprises. In addition to those that retained their original membership in this cluster (Bardejov, D. Dudince, Dudince), several entities transitioned into it from different 2023 clusters, specifically from clusters 2, 4, and 6. This shift suggests that the differences among these spas diminished over time and that their financial profiles became more similar.

At the same time, a distinct group of spa enterprises emerged in 2024, separating from the previously dominant cluster 1 of 2023. This group includes Tre. Teplice, Bojnice, Lúčky, and R. Teplice, which had formed a single cluster in the previous year but subsequently constituted a new homogeneous group (cluster 3). This development may reflect specific changes in their financial performance or strategic orientation that differentiated them from other entities.

Some spa enterprises maintained relatively stable positions despite formal changes in cluster numbering. For example, the group of spas located in the High Tatras (H. Smokovec, N. Smokovec), and Marína remained together within the same homogeneous cluster in 2024 (cluster 5). Similarly, spas such as Nimnica exhibit only minor shifts between clusters (from 3 to 4), without substantial changes in their relative position. Particular attention should be paid to Lučivná, which forms a separate cluster in both observed years. This indicates a specific financial profile that significantly differs from other analyzed entities and remains stable over time.

The results suggest a reduction in diversity among spa enterprises between 2023 and 2024, with the majority converging toward a more similar financial model. At the same time, smaller groups or individual spas with distinct development patterns persist, highlighting the continued heterogeneity within the sector.

To gain a clearer understanding of the financial characteristics of spa enterprises within the five clusters (six in 2023), descriptive statistics (specifically the mean and standard deviation) were analyzed for ten selected financial indicators (see Table 3). These indicators reflect key dimensions of performance, including profitability (ROA, ROS), efficiency (AT), cost efficiency (PCR, CSR, ROC), liquidity (NWC, CR), and indebtedness (E/TL, DR).

Cluster		ROA	ROS	AT	E/TL	NWC	DR	CR	PCR	CSR	ROC
						2023					
1	**Mean**	0.094	0.158	0.626	4.656	2,724,734	0.078	2.504	0.424	0.881	0.179
	S.D.	0.043	0.072	0.185	2.419	1,841,781	0.023	0.654	0.082	0.058	0.093
2	**Mean**	0.042	0.075	0.572	1.753	-237,687	0.124	0.794	0.363	1.202	0.032
	S.D.	0.063	0.101	0.146	1.293	1,140,588	0.048	0.560	0.069	0.197	0.063
3	**Mean**	0.126	0.200	0.631	8.596	10,100,000	0.050	14.206	0.476	0.782	0.290
	S.D.	.	.	.	.	.	.	.	.	.	.
4	**Mean**	0.087	0.068	1.282	1.921	503,091	0.131	2.802	0.536	0.974	0.056
	S.D.	0.086	0.064	0.136	1.725	504,963	0.042	1.123	0.090	0.079	0.074
5	**Mean**	-0.081	-0.188	0.431	0.409	-2,124,406	0.429	0.118	0.216	1.294	-0.168
	S.D.	.	.	.	.	.	.	.	.	.	.
6	**Mean**	-0.295	-0.395	0.748	0.505	1,676,217	0.097	3.597	0.459	1.449	-0.284
	S.D.	.	.	.	.	.	.	.	.	.	.
						2024					
1	**Mean**	0.045	0.053	0.889	2.166	361,993	0.111	1.875	0.501	0.995	0.033
	S.D.	0.039	0.045	0.330	1.704	742,321	0.047	1.413	0.074	0.044	0.044
2	**Mean**	0.024	0.051	0.467	0.469	-3,346,851	0.666	0.052	0.303	1.037	0.026
	S.D.	.	.	.	.	.	.	.	.	.	.
3	**Mean**	0.098	0.177	0.567	5.831	4,209,738	0.076	2.603	0.415	0.844	0.210
	S.D.	0.038	0.039	0.196	2.640	3,575,770	0.026	0.833	0.079	0.040	0.050
4	**Mean**	0.097	0.160	0.604	10.226	11,800,000	0.041	18.005	0.484	0.810	0.243
	S.D.	.	.	.	.	.	.	.	.	.	.
5	**Mean**	0.161	0.112	1.440	5.202	832,559	0.125	3.670	0.605	0.907	0.121
	S.D.	0.029	0.015	0.179	0.746	450,853	0.039	1.298	0.196	0.015	0.016

Note: Cluster numbers correspond to those shown in the dendrograms (Figure 1). S.D. denotes standard deviation.

Table 3 Cluster mean and standard deviation for each financial indicator

Table 3 shows clear differences in financial profiles across clusters. In 2023, clusters 3 and 1 represent the most profitable and financially stable spa enterprises, while clusters 5 and 6 are characterized by negative profitability and weaker financial positions. In 2024, the structure becomes more consolidated, with most enterprises grouped in cluster 1 with moderate performance, while clusters 3 and 4 remain the strongest and cluster 2 represents the weakest financial position. Overall, the results indicate a convergence toward average performance, accompanied by the persistence of a few extreme groups.

Table 3 shows clear differences in financial profiles across clusters. In 2023, clusters 3 (Nimnica) and 1 (Bardejov, D. Dudince, Dudince, Tre. Teplice, Bojnice, Lúčky, and R. Teplice) represent the most profitable and financially stable spa enterprises, while clusters 5 (Lúčivná) and 6 (Sliač) are characterized by negative profitability and weaker financial positions. In 2024, the structure becomes more consolidated, with most enterprises grouped in cluster 1 with moderate performance, while clusters 3 (Tre. Teplice, Bojnice, Lúčky, R. Teplice) and 4 (Nimnica) remain the strongest and cluster 2 (Lúčivná) represents the weakest financial position. Overall, the results indicate a convergence toward average performance, accompanied by the persistence of a few extreme groups.

4 Discussion

Several spa enterprises experienced horizontal (lateral) movements between clusters, which did not necessarily reflect improvements or deteriorations in performance, but rather shifts in their financial orientation. These transitions may be associated with internal restructuring processes, diversification of activities, or changes in strategic positioning, and can be interpreted as an expression of adaptability. Existing literature highlights that the ability to adjust to changing market conditions is essential for the long-term sustainability of the spa industry (Szromek et al., 2022). Similarly, Dang et al. (2020) stress the need for a strategic transformation from traditional spa services toward more integrated health tourism concepts, supported by technological progress and interdisciplinary approaches—an evolution that may already be reflected in certain enterprises within the sector. Ngo et al. (2025) further demonstrate that spa providers increasingly adapt their services to different demographic segments, with successful establishments typically combining medical, recreational, and cultural offerings, thereby enhancing their competitiveness and market reach. In line with this diversification trend, Oghbatalab et al. (2025) identify common features of high-performing spa enterprises, including strategic marketing practices, effective digital communication with clients, and strong integration with healthcare services. These elements contribute to improved customer engagement and represent potential benchmarks for less successful providers seeking to strengthen their market position.

From a policy standpoint, Ramanauskas et al. (2023) suggest that public support for the spa sector should focus on sustainability-oriented measures, such as tax incentives, workforce development, and simplified investment procedures. Such initiatives could be particularly relevant for financially weaker spa enterprises, improving their competitiveness and long-term stability. More broadly, these findings may be useful for regional policymakers and tourism authorities. As noted by Panyadee et al. (2023), benchmarking and performance evaluation are important tools for evidence-based policy in health and wellness tourism, as they enable the identification of leading destinations and more efficient allocation of public resources toward infrastructure, innovation, and sustainable development of the sector.

5 Conclusion

The results of the analysis confirm that the financial structure of spa enterprises in Slovakia underwent noticeable changes between 2023 and 2024. The reduction in the number of clusters and the formation of a dominant group indicate a tendency toward convergence in financial performance across the sector. This suggests that a large proportion of spa enterprises are adopting similar financial patterns, potentially reflecting common market conditions or management strategies. At the same time, the persistence of smaller clusters and individually positioned enterprises highlights that not all entities follow this general trend. Some spa enterprises maintain distinct financial profiles, which may result from differences in size, specialization, or strategic orientation. These differences underline the continued heterogeneity within the sector. The findings demonstrate that while the Slovak spa sector shows signs of increasing homogenization, it still retains a degree of internal diversity. The results provide useful insights into the financial positioning of spa enterprises and may serve as a basis for further research, as well as for managerial decision-making and policy formulation in the field of health tourism.

Future research could extend this analysis by applying alternative or complementary multivariate methods to provide a more comprehensive assessment of spa enterprises' financial performance. In particular, multi-criteria decision-making techniques such as TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) or

other methods within the family of MCDA (Multi-Criteria Decision Analysis) could be used to rank enterprises according to their overall financial strength. Finally, extending the dataset to a longer time period or incorporating non-financial indicators (e.g., service quality, occupancy rates, or regional factors) could provide a more holistic view of the sector's development.

Acknowledgements

This research was supported by the Scientific Grant Agency of the Ministry of Education, Research, Development and Youth of the Slovak Republic and the Slovak Academy of Sciences, grant No. 1/0241/25–VEGA.

References

Agag G, Yousaf A, Mishra A, Amin I (2025) How and when do marketing analytics pay off in the travel and tourism industry? The role of marketing agility and fit perspective. International Journal of Hospitality Management 126:104060. https://doi.org/10.1016/j.ijhm.2024.104060

Dang H-S, Nguyen T-M-., Wang C-N, Day J-D, Dang T M (2020) Grey system theory in the study of medical tourism industry and its economic impact. International Journal of Environmental Research and Public Health 17(3):961. https://doi.org/10.3390/ijerph17030961

Grande K, Dong H, Peypoch N (2024) Successful camping operators: Performance measurement using cluster analysis and DEA methods. Journal of Outdoor Recreation and Tourism 46:100768. https://doi.org/10.1016/j.jort.2024.100768

Milojević S, Matejić T, Knežević S (2025) Analysis of dynamic financial performance of airline companies using an algorithm combining genetic algorithm and DTW clustering. TEM Journal 14(4):2895–2910. https://doi.org/10.18421/TEM144-01.

Oghbatalab H, Hamid M, Salimi B, Rabbani M (2025) Performance optimization of the an health tourism organization: An integrated approach based on resilience engineering and emotional intelligence. Human Factors and Ergonomics in Manufacturing & Service Industries 35(3):e70003. https://doi.org/10.1002/hfm.70003

Panyadee C, Krajangchom S, Sangkakorn K, Intawong K (2023) Smart wellness technology for tourism destination based on evolving tourist expectation model. TEM Journal 12:1218–1226. https://doi.org/10.18421/tem122-68

Ramanauskas J, Banevičius Š, Jerochina-Labanauskienė J (2023) Aspects of the socio-economic impact of health tourism. Baltic Journal of Economic Studies 9(5):23–32. https://doi.org/10.30525/2256-0742/2023-9-5-23-32

Szromek A R, Puciato D, Markiewicz-Patkowska J I, Colmekcioglu N (2022) Health tourism enterprises and adaptation for sustainable development. International Journal of Contemporary Hospitality Management 35(1):1–25. https://doi.org/10.1108/ijchm-01-2022-0060

Torkzadeh L, Jalilian H, Zolfagharian M, Torkzadeh H, Bakhshi M, Khodayari-Zarnaq R (2024) Market segmentation in the health tourism industry: a systematic review of approach and criteria. Journal of Policy Research in Tourism, Leisure and Events 16(1):69–88. https://doi.org/10.1080/19407963.2021.1988622

Torres-Pruñonosa J, Raya J M, Crespo-Sogas P, Mur-Gimeno E (2022) The economic and social value of spa tourism: The case of balneotherapy in Maresme, Spain. PLOS ONE 17(1):e0262428. https://doi.org/10.1371/journal.pone.0262428

Tutar H, Ay G, Şahi M (2025) Financial Performance Analysis of Companies in the Stock Exchange Istanbul Tourism Index. Journal of Tourism, Sustainability and Well-being 13(1):36–54. https://doi.org/10.34623/78wh-ax35

Vašaničová P, Košíková M, Jenčová S, Miškufová M, Korečko J (2025) Financial Performance-Based Clustering of Spa Enterprises in Slovakia. Journal of Risk and Financial Management 18(9):482. https://doi.org/10.3390/jrfm18090482

Wang H, Zhang W, Zhang K, Hu B (2026) The rebounding paradox in the hospitality and tourism industry: Why strategic ESG strengthens organizational resilience while responsive ESG undermines it. International Journal of Hospitality Management 135:104614. https://doi.org/10.1016/j.ijhm.2026.104614

Wu J, Fu X M, Song H (2025) ESG and financial performance in Asia-Pacific gaming tourism industry: Pandemic resilience and recovery. Tourism Economics. https://doi.org/10.1177/13548166251395896

Beyond Predict-then-Optimize: Decision-Focused Learning for Grid-Scale Battery Arbitrage

Vojtěch Vávra[1]

Abstract. The integration of renewable energy has introduced extreme price volatility in short term electricity markets, creating arbitrage opportunities for Grid Scale Battery Energy Storage Systems. Traditional trading relies on a two step predict then optimize method, minimizing statistical error before the final optimization step. This approach is suboptimal due to the severe disconnect between statistical accuracy and financial outcomes.

This paper proposes an End to End Optimization framework applied to the Czech day-ahead and Imbalance markets. We encode physical and operational constraints directly into a convex optimization problem acting as the final differentiable layer of a neural network. The model is trained via decision-focused learning. The loss function directly minimizes financial regret rather than standard statistical errors. Empirical evaluations on historical market data demonstrate that our integrated framework yields superior capital preservation and risk mitigation compared to traditional decoupled models, establishing a much safer bidding strategy for highly volatile environments.

Keywords: Decision-Focused Learning, Battery Storage, Energy Arbitrage, Predict-then-Optimize

JEL Classification: C45, C61, Q41

1 Introduction

The rapid shift toward renewable energy has completely changed how short term electricity markets work. The growing share of variable power sources, especially solar and wind, has caused massive price swings. We can see this clearly in the duck curve effect. During the middle of the day, solar panels produce too much power, driving spot prices to zero or even below zero. Then, when the sun goes down and evening demand hits, prices spike rapidly. In this unstable environment, large battery storage systems have become a crucial tool. By trading energy through arbitrage, which means charging when power is cheap and discharging when prices peak, battery operators help stabilize the grid and can also protect their investment capital against market shocks.

To maximize these profits, operators usually use a two stage predict then optimize method. First, a machine learning model learns to forecast future electricity prices. Next, the system treats these predictions as absolute facts and passes them to a mathematical solver. This solver, like a linear programming model, then creates the best plan for charging and discharging. The main problem with separating these steps is how the forecasting model learns. Usually, forecasting algorithms train to minimize standard statistical errors, like Mean Squared Error or Mean Absolute Error.

However, statistical accuracy does not automatically mean business success. The standard error metric assumes that all mistakes of the same size matter equally. For battery trading, this is highly misleading. Missing a large price spike could cost thousands of euros in lost revenue. On the other hand, making a prediction mistake during a flat and cheap period might not change the battery actions at all, because the system would just stay fully charged. Since the traditional machine learning model is completely blind to the physical limits of the battery and the financial impact of its choices, the standard predict then optimize method is naturally flawed as demonstrated by Bengio, 1997 and Wilder et al, 2019.

This paper fixes this disconnect by proposing an end to end optimization structure based on decision-focused learning. Instead of focusing on prediction errors, this learning approach puts the final trading problem directly into the training loop of the neural network. By coding the physical limits of the battery and the market logic as a convex optimization layer that the network can learn through Amos and Kolter, 2017, we train the model to directly minimize lost financial profit.

This paper brings three main contributions. First, we formulate a battery arbitrage strategy specifically designed for the extreme volatility of the imbalance market, incorporating strict physical degradation limits. Second, we place this exact linear program inside a Temporal Fusion Transformer using KKT mathematical conditions that

[1] Prague University of Economics and Business, Department of Econometrics, náměstí Winstona Churchilla 1938/4, 130 67 Praha 3, vojtech.vavra@vse.cz

allow for network training. Finally, we test the proposed framework on real world data from the Czech electricity market. We show that optimizing for decision quality provides significantly better capital preservation and risk management compared to traditional models that only care about forecast accuracy.

2 Methodology

The main idea of decision-focused learning is to replace the traditional two stage method. Instead of forecasting and optimization working in total isolation, we combine them into one unified structure that can learn as a whole. To do this, we need to express the operational limits of the battery as a convex mathematical layer that fits directly into the neural network.

2.1 Battery Arbitrage as a Convex Optimization Problem

The goal of the battery operator is to maximize profit by perfectly planning when to charge and discharge over a set time window. We typically use $H = 96$ periods to represent a full day broken down into 15 minute intervals.

Let $\hat{\pi}_t$ be the predicted electricity price at time step t. The decision variables P_t^{ch} and P_t^{dis} represent the charging and discharging power in megawatts. The stored energy in the battery at time t is denoted by E_t in megawatt hours. The battery operation is restricted by its physical limits. These include the maximum power capacity P_{max}, total energy capacity E_{max}, and the round trip cycle efficiency, which we distribute symmetrically as $\sqrt{\eta}$.

To ensure the model behaves realistically, we must also account for battery wear and tear. We introduce a degradation cost C_{deg} applied to every megawatt hour of discharged energy. We can write this as a linear programming model:

$$\max_{P^{ch}, P^{dis}} \sum_{t=1}^{H} \Delta t \left[P_t^{dis} (\hat{\pi}_t - C_{deg}) - P_t^{ch} \hat{\pi}_t \right] \tag{1}$$

Subject to the following operational constraints:

$$E_t = E_{t-1} + \Delta t \left(P_t^{ch} \sqrt{\eta} - \frac{P_t^{dis}}{\sqrt{\eta}} \right), \quad \forall t \in \{1, \dots, H\} \tag{2}$$

$$0 \leq E_t \leq E_{max}, \quad \forall t \in \{1, \dots, H\} \tag{3}$$

$$E_0 = 0 \tag{4}$$

$$0 \leq P_t^{ch} \leq P_{max}, \quad \forall t \in \{1, \dots, H\} \tag{5}$$

$$0 \leq P_t^{dis} \leq P_{max}, \quad \forall t \in \{1, \dots, H\} \tag{6}$$

Equation (2) tracks the energy balance over time while accounting for efficiency losses. Equations (3) and (4) define the maximum battery capacity and the starting energy level. Equations (5) and (6) make sure the system respects the maximum power limits of the hardware.

2.2 Differentiable Optimization via KKT Conditions

In a standard predict then optimize setup, the predicted prices $\hat{\pi}$ simply go into a mathematical solver to find the best actions P^{ch*} and P^{dis*}. The true market prices π are only used after the fact to see how much money was made. The neural network would simply try to minimize the standard squared error: $\mathcal{L}_{MSE} = \frac{1}{H} \sum_{t=1}^{H} (\hat{\pi}_t - \pi_t)^2$.

In contrast, our decision-focused approach treats the entire linear problem defined by equations (1) through (6) as a structured layer that the network can learn from. Let $f(\hat{\pi})$ represent the optimal solution of our linear program. The actual financial regret, which serves as the loss function for our neural network, is the negative profit. We calculate this profit using the *true* prices π combined with the actions based on the *predicted* prices:

$$\mathcal{L}_{DFL}(\hat{\pi}, \pi) = - \sum_{t=1}^{H} \Delta t \left[P_t^{dis*}(\hat{\pi})(\pi_t - C_{deg}) - P_t^{ch*}(\hat{\pi}) \pi_t \right] \tag{7}$$

To update the neural network weights during the training process, we need to calculate how the optimal decisions change when the predicted prices change. Because finding the best solution involves searching a restricted mathematical space, calculating these changes is not a simple task.

We solve this problem using the method proposed by Amos and Kolter, 2017. By ensuring the math problem has a strictly valid convex shape, we can compute the exact learning signals by working through the KKT optimality conditions of the linear program. This allows the financial regret $\mathcal{L}_{DFL}$ to send learning signals back into the Temporal Fusion Transformer. It effectively teaches the neural network to prioritize accuracy during market conditions that truly decide the profitability of the battery, like huge price spikes or negative prices.

3 Data and Experimental Setup

The performance of decision-focused learning models depends heavily on the quality of the input data. In this study, we combine highly detailed market data with physics based weather calculations to create a strong dataset for the Czech electricity market.

3.1 Market Data and Time Series Construction

The main dataset uses 15 minute electricity market data from the Czech market operator OTE, a.s., 2026, covering the period from January 2023 to December 2025. We specifically use the day-ahead market prices (π_{DA}) and the imbalance prices (π_{Imb}), along with system wide imbalance volumes. To handle forecasting over long periods, we structure the data into daily windows of $H = 96$ steps.

To capture the repeating patterns of energy markets, we encode time related data using sine and cosine math functions. We also add simple flags for weekends and public holidays in the Czech Republic, because these days have very different demand and price behaviors compared to normal working days. These seasonal dynamics and historical market patterns for the testing period are illustrated in Figure 1.

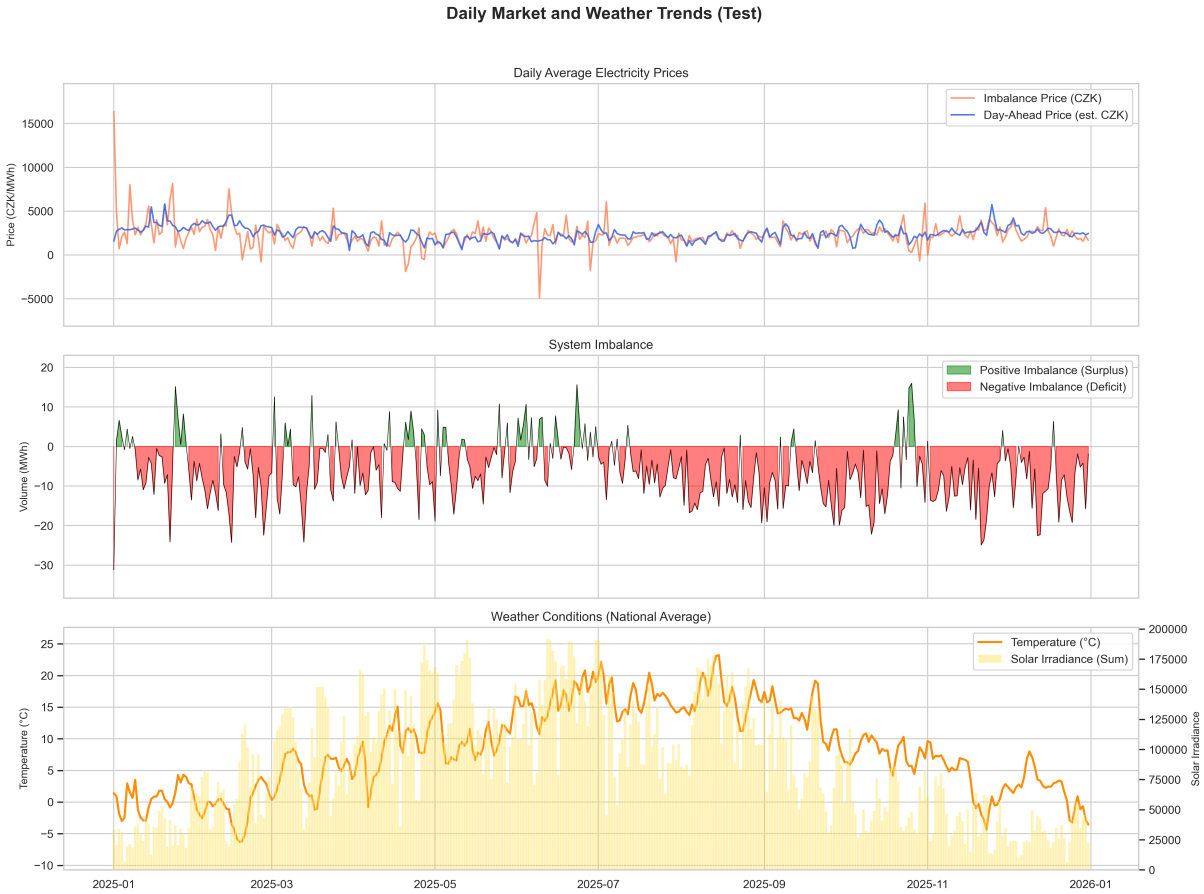


Figure 1 Market trends and seasonality profiles for the testing data

3.2 Physics Informed Feature Engineering

A major part of our data preparation is turning raw weather data from the OpenMeteo archive into simulated production numbers Zippenfenig, 2023. This approach makes learning easier for the neural network because we

manually encode the known rules of physics:

- **Simulated Solar Index:** We transform short wave radiation data into a solar production estimate. Unlike simple sun brightness numbers, our index includes a temperature penalty. This reflects the reality that solar panel efficiency drops steadily when the panel temperature goes over 25 °C.
- **Simulated Wind Potential:** Based on wind speed at 80 meters, we simulate wind power production using the rule that power grows with the cube of wind speed ($P \propto v^3$). To keep things realistic, we enforce activation limits at 3 m/s and deactivation limits at 25 m/s. This effectively models the physical limits of large scale wind turbines.
- **Heating and Cooling Demand:** To capture price changes driven by human consumption, we calculate heating and cooling needs. We use national temperature averages and compare them against 15.5 °C for heating and 22.0 °C for cooling.

Finally, we introduce history features. These include 24 hour and 7 day price histories as well as 6 hour rolling averages, which give the model a good understanding of recent market trends.

3.3 Model Architectures and Benchmarks

We build our decision-focused framework using the Temporal Fusion Transformer Lim et al, 2021. We chose this transformer for its leading performance in long term forecasting and its ability to separate static data, known future data like weather forecasts, and unknown history variables.

To compare our results, we use two benchmarks. The first is a simple strategy that charges at night and discharges in the evening. The second is a standard predict then optimize framework using a LightGBM gradient boosting model. While the LightGBM model optimizes for the lowest squared error, our transformer model trains directly on the profit objective defined in Section 2. All models train on data up to September 2024 and are evaluated on a strictly separate test set of 2025 market data.

4 Results and Discussion

We evaluate the performance of all models on the unseen data from the test year 2025. To provide a strict and fair comparison, all models send their price predictions to the exact same linear programming solver. The solver enforces the physical limits of the battery and applies a degradation cost of 1000 CZK for every discharged megawatt hour. The financial outcomes for all tested strategies are summarized in Table 1.

Strategy	Net Profit (CZK)	Degradation (MWh)
Perfect Foresight	7,502,132	1,630
Naive Time Strategy	85,958	728
LightGBM (MSE)	−410,307	895
NHiTS (MSE)	−2,566,379	2,036
TFT Baseline (MSE)	−5,717,583	4,852
TFT DFL (Ours)	−43,444	42

Table 1 Comparison of BESS Arbitrage Strategies on 2025 Test Data

The empirical results reveal a critical vulnerability in standard machine learning pipelines applied to physical energy assets. The deep learning models trained on traditional statistical error metrics, namely the NHiTS and the TFT Baseline, produced catastrophic financial losses. The degradation column provides a clear explanation for this failure. The standard TFT model aggressively cycled the battery, discharging over 4850 megawatt hours. Because these traditional models were trained solely to track price trajectories, they learned to identify and trade on hundreds of microscopic price fluctuations. However, they remained completely blind to the operational reality that every physical cycle costs real money in wear and tear. They essentially destroyed the battery through severe overtrading in pursuit of statistically accurate but economically worthless margins.

The decision-focused Temporal Fusion Transformer tells an entirely different story. Even though it shares the exact same neural architecture as the failing TFT Baseline, it achieved dramatically better financial results. By embedding the linear optimization layer directly into the neural network, the model learned the true economic cost of its actions during training. Recognizing that the extreme price volatility in 2025 was accompanied by spreads that often failed to cover the strict degradation cost, the model naturally adopted a sophisticated capital

preservation strategy. It reduced its market activity to a mere 42 megawatt hours, effectively refusing to trade during mathematically identified but economically unviable periods.

While the simple naive strategy managed to generate a small positive return by blindly capturing the most reliable daily price gap between night and evening hours, the most scientifically significant comparison lies among the machine learning models. The decision-focused framework saved over 5.6 million CZK compared to its structurally identical baseline. This behavior proves that when dealing with physical assets that suffer from degradation, upgrading the loss function to reflect actual business reality is vastly more important than upgrading the complexity of the forecasting architecture.

5 Conclusion

This paper challenged the standard predict then optimize framework traditionally used in algorithmic energy trading. We demonstrated that minimizing statistical error metrics such as Mean Squared Error is fundamentally disconnected from the actual business objective of maximizing profit and minimizing physical wear.

By formulating the operational constraints of a grid scale battery as a differentiable linear program, we successfully trained a Temporal Fusion Transformer using decision-focused learning. Tested on the highly volatile Czech electricity market, the traditional deep learning models failed severely. They overtraded on small price variations, leading to massive battery degradation and heavy financial losses. In contrast, our decision-focused model inherently learned the cost of physical degradation, effectively halting unprofitable trades and successfully protecting the investment capital.

These results point to a much needed shift in how we apply machine learning to modern energy systems. Moving forward, researchers should expand this framework to include more complex non linear models of battery degradation. It would also be highly valuable to test decision-focused learning on other active grid assets.

Furthermore, predicting the exact imbalance price remains a major challenge that requires deeper investigation. Because this specific price reflects sudden grid emergencies, unexpected weather shifts, and human behavior, it is significantly harder to forecast than the standard day-ahead market. Future work must develop more robust methods to capture these sudden market shocks. If we can improve how our models anticipate these extreme imbalance prices, the decision-focused framework will unlock even greater financial benefits for battery operators.

Acknowledgements

The research is supported by the Internal Grant Agency, Faculty of Informatics and Statistics, Prague University of Economics and Business – project no. F4/46/2025.

References

Amos B, Kolter J Z (2017) Optnet: Differentiable optimization as a layer in neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), PMLR, pp 136–145

Bengio Y (1997) Using a financial training criterion rather than a prediction criterion. International Journal of Neural Systems 8(04):433–443

Lim B, Arık S Ö, Loeff N, et al (2021) Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting 37(4):1748–1764

OTE, a.s. (2026) Yearly market report 2026. `https://www.ote-cr.cz/en/statistics/yearly-market-report?set_language=en&date=2026-01-01`

Wilder B, Dilkina B, Tambe M (2019) Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 1658–1665

Zippenfenig P (2023) Open-meteo.com weather api. `https://doi.org/10.5281/zenodo.7970649`, `https://open-meteo.com/`

Translating Online Search Behavior into Quantitative Portfolio Views: Evidence from Hong Kong

Jialei Xiong[1]

Abstract. A growing body of research confirms that investor attention, captured through Internet search activity, carries predictive information for asset returns, trading volume, and volatility. Yet attention signals have been rarely incorporated into portfolio construction frameworks. This study proposes a methodology that operationalizes Google Trends (GT) search data as quantitative investor views within the Black–Litterman (BL) model. Attention is measured as the relative deviation of current search intensity from its eight-week median baseline, consistent with the attention-driven price pressure hypothesis. Panel regression results confirm that attention changes statistically significantly predict next-week stock returns. We then embed this signal into the BL framework to construct GT-augmented portfolios. A rolling-window out-of-sample back test benchmarks the resulting BL+GT strategy against passive and mean–variance-optimized alternatives. The strategy yields higher cumulative returns and lower maximum drawdown, suggesting that search-based attention provides incremental predictive information for portfolio construction.

Keywords: Investor attention, Portfolio optimization, Behavioral finance, Google Trends, Black–Litterman model

JEL Classification: G11, G41

1 Introduction

The Efficient Market Hypothesis (Fama, 1970) is the conventional starting point of modern asset pricing. It treats financial markets as efficient information-processing mechanisms in which prices fully reflect all available information, so that no public signal can generate persistent risk-adjusted excess returns. On this foundation rest most of the standard tools of portfolio selection, asset pricing, and risk management. Over the last four decades, however, a substantial body of empirical evidence has accumulated that challenges the rationality assumption behind this framework. Documented anomalies include long-horizon reversals, short-term momentum, size effects, calendar patterns, and excess volatility that exceeds what is justified by fundamentals. The coexistence of these patterns is difficult to reconcile with the view that prices are formed only by rational, risk-based mechanisms.

Behavioral finance provides an alternative account. It argues that real investors are subject to cognitive constraints and information-processing limits, and that these constraints generate systematic, not random, departures from fully rational choice. A classical objection is that even if many individual investors are biased, rational arbitrageurs should bring prices back to fundamental values. The limits-to-arbitrage literature shows, however, that arbitrage is bounded by capital constraints, short horizons, and the risk that mispricing can widen before it corrects. Individual-level biases, therefore, do not wash out in aggregate but can persist as market-level distortions, especially in market segments with a large share of retail participants (Barber and Odean, 2008).

Within the behavioral paradigm, much work has focused on investor sentiment, a directional measure of how optimistic or pessimistic investors are about an asset. A more fundamental variable is investor attention, which captures what investors are considering before any evaluation is formed. Attention is a scarce cognitive resource (Kahneman, 1973) that must be allocated across thousands of tradable securities. Because processing capacity is limited, this allocation itself influences which assets enter investors' decision sets and, through that channel, how prices are formed. In Merton's (1987) recognition framework, stocks that investors are less aware of earn higher expected returns, making the breadth of attention, not just its direction, a determinant of equilibrium pricing. Measuring investor attention directly, however, is difficult. Early proxies relied on analyst coverage, media coverage, or trading volume, but each has limitations: analyst coverage reflects institutional information production; media coverage is driven by editorial choices and measures information supply rather than demand; trading volume is jointly produced by attention, information arrivals, and liquidity shocks, which creates

[1] VŠB - Technical University of Ostrava, Department of Finance, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic, jialei.xiong@vsb.cz.

endogeneity. Da et al. (2011) introduced a more direct alternative, Google search volume. When investors become interested in a stock, they typically search for information about it, so search volume records an active information-seeking action at a high frequency and is especially informative about retail investors, who are the primary users of general-purpose search engines. A broad empirical literature has since confirmed that Google search volume predicts short horizon returns, trading activity, and volatility across markets and asset classes.

Most of this evidence, though, remains at the asset pricing level. The literature has devoted considerable effort to testing whether attention predicts returns, but much less effort to the question of how attention signals can be turned into concrete portfolio construction rules. The few studies that have attempted this translation have used relatively informal constructions. Kristoufek (2013), for example, proposes a power-law weighting scheme in which stocks with higher search volume receive smaller weights, a heuristic approach that lacks a formal optimization structure and does not quantify the uncertainty of the attention signal. A natural way to embed a behavioral signal in a principled optimization framework is the Black–Litterman (BL) model (Black and Litterman, 1992). The BL model is a Bayesian construction in which an equilibrium prior, obtained by reverse optimization from market capitalization weights, is updated using investor views with an explicit confidence matrix. The model is well-suited to attention signals for three reasons. First, an attention change is naturally directional, its sign can set the direction of a view. Second, its magnitude can be mapped to the confidence attached to that view through Idzorek's (2007) method, so that stronger attention shifts tilt the posterior more strongly away from the prior. Third, the framework handles situations in which the signal is informative for only a subset of stocks, because assets without a view retain their equilibrium expected returns. Existing BL extensions have incorporated analyst forecasts or macroeconomic views, but to our knowledge, no prior study has systematically embedded Google Trends attention signals into the BL view structure.

This paper fills that gap. We construct an attention change indicator from Google Trends search volume, validate its predictive content for next-week returns through a fixed-effects panel regression with heteroskedasticity-robust standard errors, and translate the validated signal into investor views within the BL framework. The empirical sample consists of Hang Seng Index (HSI) constituents. The Hong Kong equity market is a natural testing ground for this design. Google is the dominant search engine in Hong Kong with a market share of approximately 92% (Statcounter, 2026), so search volume provides a representative measure of local information-seeking activity. At the same time, the market has a high share of retail participation, which is the investor segment most likely to trade on attention-based signals. A rolling-window out-of-sample back test then compares the resulting BL+GT strategy to an equal-weight benchmark. The results show that the attention signal predicts next-week returns at the 5% level after filtering low-quality search series, and that the BL+GT strategy improves both return and risk-adjusted return relative to the benchmark without raising volatility.

2 Methodology

In this section, we will provide a brief overview of the methods applied in this study.

2.1 Attention change indicator

The attention change indicator ΔGSV measures the deviation of investor search attention for a specific stock relative to its recent "normal level." The construction method follows the idea of the Abnormal Search Volume Index (ASVI) proposed by Da et al. (2011). However, the present study does not use the logarithmic difference form. Instead, it adopts the relative form proposed by Qureshi (2025). The formula is defined as:

$$\Delta GSV_{i,t} = \frac{GSV_{i,t} - Med(GSV_{i,t-1}, \dots, GSV_{i,t-8})}{Med(GSV_{i,t-1}, \dots, GSV_{i,t-8})}, \tag{1}$$

where $GSV_{i,t}$ denotes the Google Trends search data for stock i in week t, and $Med(GSV_{i,t-1}, \dots, GSV_{i,t-8})$ represents the median value over the past 8 weeks. This method enables the capture of short-term fluctuations in search interest that are likely to be associated with changes in market sentiment and trading behavior. Two design choices deserve comment. First, the median rather than the mean is used as the baseline because financial search series occasionally exhibit large spikes driven by news events or earnings releases. The mean would be heavily pulled by a single outlier and would contaminate the normalization, whereas the median remains stable under such events. Second, the eight-week window follows Da et al. (2011). It is long enough to smooth out week-to-week noise and to establish a stable baseline, yet short enough to track genuine shifts in investor interest rather than secular trends in search behavior.

2.2 Predictive validation

Before the signal can serve as a BL view, its predictive content for next-week returns must be verified. We estimate a fixed-effects panel regression as follows:

$$r_{i,t+1} = \alpha_i + \gamma_t + \beta \cdot \Delta GSV_{i,t} + \varepsilon_{i,t}, \tag{2}$$

where $r_{i,t+1}$ is the return of stock i in week $t+1$, α_i denotes the individual fixed effect, γ_t represents the time fixed effect, β is the core attention coefficient, and $\varepsilon_{i,t}$ is the error term. Stock fixed effects absorb time-invariant heterogeneity across firms, such as differences in baseline visibility or industry composition. Time fixed effects absorb common shocks such as market-wide news and aggregate sentiment swings, so β isolates the within-stock, within-week effect of attention on next-week returns. Standard errors are computed using the HC3 estimator (MacKinnon and White, 1985), which weights each squared residual by the reciprocal of its leverage and delivers reliable inference in the presence of heteroskedasticity and high-leverage observations, both of which are pervasive in weekly return panels. A positive and statistically significant β would be consistent with the attention-driven price pressure hypothesis, retail investors crowd into attention-grabbing stocks and push prices up in the short run. A negative β would instead be consistent with an overreaction reversal pattern. Either way, statistical significance establishes that ΔGSV carries information about future returns and therefore qualifies as a valid view signal in the BL framework.

This validation step is not only a specification test but a necessary precondition for the BL construction that follows. The sign of the estimated coefficient determines the direction assigned to r_{high} and r_{low} in the view vector Q, while the statistical reliability of β, established here through fixed effects and HC3 standard errors, justifies treating $|\Delta GSV|$ as a meaningful basis for the confidence calibration in Ω. Without this validation, the attention-based views would rest on an unverified premise and could inject noise rather than information into the posterior. The regression therefore serves as the empirical bridge between the raw attention signal and its formal role as a Bayesian view in Section 2.3.

2.3 Black-Litterman integration

The Black–Litterman model proposed by Black and Litterman (1992) is a Bayesian framework that combines market equilibrium returns with investor views. The core idea is to treat the market equilibrium return as the prior distribution and then update it using investor views to obtain the posterior expected return. Portfolio optimization is then performed based on this posterior expectation. The posterior expected return vector in the BL model is obtained using the Bayesian updating formula:

$$E(R) = [(\tau\Sigma)^{-1} + P^T\Omega^{-1}P]^{-1}[(\tau\Sigma)^{-1}\pi + P^T\Omega^{-1}Q]. \tag{3}$$

The posterior covariance matrix is:

$$\overline{\Sigma} = [(\tau\Sigma)^{-1} + P^T\Omega^{-1}P]^{-1} + \Sigma. \tag{4}$$

The parameter τ is a scalar parameter that reflects the degree of uncertainty in the prior estimate. This study sets τ = 0.05, indicating relatively high confidence in the prior return estimates. The value of τ influences the sensitivity of the posterior to the view signals, a larger τ implies greater uncertainty in the prior, causing the posterior to lean more toward the investor views. Representing the implied expected returns that render the current market capitalization weights optimal, the prior return vector π in the BL model is established via reverse optimization from market equilibrium：

$$\pi = \lambda \cdot \Sigma \cdot w_{mkt}. \tag{5}$$

Where λ represents the risk aversion coefficient, reflecting the overall risk preference of the market, with risk aversion λ = 2.5 (He and Litterman, 2002). The covariance matrix Σ denotes the covariance matrix of asset returns, is estimated on a rolling 52-week window. Parameter w_{mkt} is the weight vector based on actual free-float market capitalizations. Absolute views are specified by setting P to the identity matrix. The view return Q is driven by the sign of ΔGSV. Stocks with $\Delta GSV_{i,t} > 0$ are assigned the historical average next week return r_{high} conditional on positive attention changes, while stocks with $\Delta GSV_{i,t} < 0$ are assigned r_{low}. Both quantities are re-estimated within the training window at every rebalancing date, so no forward-looking information enters the strategy. It can be regarded as an investor view based on Google Trends data and integrated into the BL model.

The confidence matrix Ω follows the method of Idzorek (2007), which translates the view confidence expressed as a percentage into the diagonal elements of Ω. Confidence levels are assigned in proportion to the magnitude

$|\Delta GSV|$. Stocks in the top 25% of $|\Delta GSV|$ at a given rebalancing date receive a confidence of 75%, stocks in the bottom 25% receive 25%, and the remaining middle group receives 50%. Stronger attention shifts therefore tilt the posterior more strongly away from the equilibrium prior, while weaker shifts leave the prior largely intact. Portfolio weights are obtained by mean–variance optimization on the BL posterior under two constraints. The first is non-negative weights, which reflect the short-selling restrictions faced by most retail investors in Hong Kong. The second is weights summing to one. The constrained quadratic program is solved numerically with Sequential Least Squares Programming.

3 Data

The empirical sample covers January 2020 to December 2024 at weekly frequency and consists of Hang Seng Index constituents. For each stock, Google Trends search volume is collected using the keyword format "HKG:xxxx" (e.g., HKG:0700 for Tencent Holdings), with the Topic filter applied to exclude unrelated queries and "Worldwide" selected as the geographic scope, since the Hong Kong filter returns insufficient data. For periods longer than five years, Google Trends downgrades from weekly to monthly frequency, so the full history is retrieved in overlapping five-year sub-intervals and spliced through proportional calibration on a four-week overlap. Google Trends reports zero whenever search activity falls below an internal threshold. A high share of zeros causes the denominator in formula (1) to vanish and contaminates the indicator. We therefore exclude stocks whose zero-value ratio over the sample exceeds 20%. Applying this filter leaves 23 HSI constituents with acceptable data quality; the remaining sporadic zero observations are set to missing in the ΔGSV computation. Weekly stock returns are computed from closing prices and used as the dependent variable in regression and as the realized return in the back test.

4 Results

We first give the descriptive properties of ΔGSV. Table 1 summarizes the distribution of ΔGSV over the 23 filtered stocks. The mean is close to zero, which is consistent with attention fluctuating around a stable baseline over long horizons. The minimum of −1.000 corresponds to the theoretical floor when current search volume falls to zero; the maximum of 6.571 reflects extreme attention spikes during major market events. The sample contains a comparable number of positive and negative observations, which is sufficient to support directional BL views in both regimes.

Statistic	Value
Observations	6017
Mean	0.0279
Standard deviation	0.4193
Minimum	−1.000
Maximum	6.5714
The number of stocks-weeks when $\Delta GSV > 0$	2124
The number of stocks-weeks when $\Delta GSV < 0$	2558

Table 1 Descriptive statistics of the ΔGSV indicator

Table 2 reports the results of the panel regression in formula (2) under the two sample specifications. Columns are estimated with z-standardized variables, stock fixed effects, and HC3 robust standard errors. The first specification includes all 30 stocks initially selected and yields a positive but statistically insignificant coefficient (β = 0.0243, p = 0.118). The second specification excludes the seven stocks with severe zero-value problems and retains 23 stocks with acceptable data quality. The coefficient becomes significant at the 5% level (β = 0.0603, p = 0.020). A one standard deviation increases (increase is a noun here) in ΔGSV is therefore associated with an increase of about 0.06 standard deviations in the next-week return. The R-squared value of 0.007 is low in absolute terms but falls within the range typically reported for return prediction regressions (Da et al., 2011). The coefficient remains significant in a pooled OLS robustness check (β = 0.0609, p = 0.019). The direction of the effect is consistent with the attention-driven price pressure hypothesis of Barber and Odean (2008).

Statistic	All 30 stocks	23 filtered stocks
Coefficient β	0.0243	0.0603
p-value	0.118	0.020**
R-squared	0.004	0.007
F-statistic p	0.783	0.423
Condition number	31	24.9

Table 2 Panel regression of next-week returns on ΔGSV

Before implementing the back test, we verify the directional behavior of the view returns on a fixed training block covering the first 80% of the sample. The average next-week return is 0.3944% for weeks with $\Delta GSV > 0$ and −0.1280% for weeks with $\Delta GSV < 0$, so rising attention is indeed followed on average by positive returns and falling attention by negative returns. These two statistics are only a descriptive check. In the back test, r_{high}, r_{low}, the covariance matrix, and all BL inputs are re-estimated at every rebalancing date using a rolling 120-week training window.

The out-of-sample period spans 132 weeks. At each week, the BL view vector and confidence matrix are reconstructed from the latest ΔGSV cross-section, and the portfolio is re-optimized. Table 3 compares the BL+GT strategy against an equal-weight benchmark that serves as a standard naive-diversification reference. The BL+GT strategy delivers a higher annualized return, a slightly lower annualized volatility, and a higher Sharpe ratio. The improvement arises from a simultaneous gain in return and reduction in risk, rather than from additional risk-taking.

Metric	BL+GT	Equal weight
Annualized return	12.34%	10.97%
Annualized volatility	20.05%	20.31%
Sharpe ratio	0.6153	0.5402

Table 3 Out-of-sample performance of the BL+GT and equal-weight strategies

To assess whether the performance gap is statistically meaningful, we apply a Wilcoxon signed-rank test to the weekly return differences between the two strategies. The median weekly return difference is 0.0154%, and the mean is 0.0262%. The test statistic is W = 3611.0 with a p-value of 0.0772, which is marginally significant at the 10% level but does not reach the 5% threshold. Two factors most likely drive the limited statistical power. First, the cross-section is small (23 stocks) because of the zero-value filter on Google Trends data. Second, the attention signal is inherently noisy at the weekly horizon. Despite the marginal significance, the economic magnitude is not negligible: an annualized excess return of 1.37% is material when compounded over longer horizons, and the strategy achieves it without increasing risk.

5 Conclusion

This paper proposes a method that operationalizes Google Trends search data as investor views within the Black–Litterman framework. Three main findings emerge from the empirical analysis. First, the relative attention change indicator ΔGSV predicts next-week returns of Hang Seng Index constituents at the 5% level once low-quality search series are filtered out, with a positive sign consistent with attention-driven price pressure. Second, embedding this signal in the BL view structure, with confidence calibrated by $|\Delta GSV|$ through Idzorek's method, produces a portfolio that outperforms the equal-weight benchmark on annualized return and Sharpe ratio over a 132--week out-of-sample window, while maintaining a slightly lower volatility. Third, the improvement is driven by simultaneous gains in return and risk control rather than by additional risk-taking, which indicates that the attention signal contributes genuine predictive information beyond what is already captured in the equilibrium prior.

The results carry several implications. From a methodological perspective, they demonstrate that behavioral signals need not remain confined to reduced form asset pricing regressions but can be translated into implementable portfolio rules through a Bayesian framework that quantifies their uncertainty. From a practical perspective, the BL+GT approach is fully replicable: the attention indicator is constructed from publicly available search data, the view calibration rule is transparent, and the portfolio optimization is standard. The results should

be read against two limitations of the current implementation. The sample is limited to 23 stocks because of the zero-value filter, which reduces the statistical power of the Wilcoxon test, and transaction costs are not yet included in the back test.

Acknowledgements

The research was financially supported by the SGS project no. SP2026/003 of VSB – Technical University of Ostrava.

References

Barber B M, Odean T (2008) All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. The review of financial studies 21: 785-818. https://doi.org/10.1093/rfs/hhm079

Black F, Litterman R (1992) Global portfolio optimization. Financial analysts journal 48: 28-43. https://doi.org/10.2469/faj.v48.n5.28

Da Z, Engelberg J, Gao P (2011) In search of attention. The journal of Finance 66: 1461-1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

Fama E F (1970) Efficient capital markets: A review of theory and empirical work. The journal of Finance 25: 383-417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x

He G, Litterman R (2002) The intuition behind Black-Litterman model portfolios. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.334304

Idzorek T (2007) A step-by-step guide to the Black-Litterman model: Incorporating user-specified confidence levels. In Forecasting expected returns in the financial markets, pp 17-38. https://doi.org/10.2139/ssrn.3479867

Kahneman D (1973) Attention and Effort. Englewood Cliffs, Prentice-Hall.

Kristoufek L (2013) Can Google Trends search queries contribute to risk diversification?. Scientific reports 3: 2713. https://doi.org/10.1038/srep02713

MacKinnon J G, White H (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of econometrics 29: 305-325. https://doi.org/10.1016/0304-4076(85)90158-7

Merton R C (1987) A simple model of capital market equilibrium with incomplete information. The Journal of Finance 42: 483-510. https://doi.org/10.1111/j.1540-6261.1987.tb04565.x

Search Engine Market Share Worldwide - Statcounter Global Stats (2026). StatCounter Global Stats. https://gs.statcounter.com/search-engine-market-share

Modeling number of advertisements and asking prices of flats with regression analysis: Moravian-Silesian Region 2018–2026

Jaromír Zahrádka[1], Jaroslav Marek[2]

Abstract. Monthly time series of apartment listings and average asking prices per square meter in six districts of the Moravian-Silesian Region (08/2018–03/2026) are analyzed. To examine the influence of the number of advertisements and mortgage interest rates on offered prices, a system of simultaneous equations is estimated. This approach allows modeling the relationship between apartment supply and price.

Keywords: Simultaneous equations, real estate market, average asking prices of apartments, estimates of unknown parameters

JEL Classification: C30, C32

1 Introduction

The housing market has become an important subject of economic and regional research due to its influence on household wealth, affordability of housing, and regional development. Housing prices are affected by a combination of demographic, economic and financial factors, including labour market conditions, population growth, housing supply and mortgage interest rates. Understanding these relationships is essential for both policymakers and market participants. Recent studies have increasingly employed advanced statistical and machine learning methods to analyse housing prices and identify their determinants. Machine learning approaches have demonstrated strong predictive capabilities in real estate applications (Forradellas and Benitez, 2026), (Cellmer and Kobylińska, 2024), and (Çılgın and Gökçen, 2023). At the same time, regional analyses have shown that housing market dynamics differ substantially across economically differentiated regions and may respond differently to economic shocks (Hromada et al., 2025), and (Hromada et al., 2023). These findings highlight the importance of analysing housing markets at a regional level and considering both supply-side and financial factors.

The objective of this study is to analyse the mutual relationships among apartment asking prices, the number of advertised apartments and mortgage interest rates in selected districts of the Moravian-Silesian Region during the period 2018–2026. To account for the potential simultaneity between housing supply and prices, a simultaneous equations model is employed. The study is based on a unique dataset obtained through automated collection of advertisements from the Czech real estate portal sreality.cz. Simultaneous equations for modeling the relationships between the prices of new apartments, old apartments, unemployment rate, and municipality size were used in the article (Koťátková-Stránská, Marek, 2015).

2 Methodology

In this chapter, the simultaneous equations model is introduced. The price data used in this study were obtained through a proprietary real-estate web analytics software solution. Information on the floor area of advertised apartments and their asking prices was acquired by automated web parsing. Subsequently, the price per square meter was calculated. This procedure enabled the determination of average apartment prices for municipalities with extended powers.

A Python script was developed for collecting data from the sreality.cz website. First, the data structure was defined, and the collected daily information was subsequently stored in a MongoDB database. Details of the data acquisition methodology were described by (Pozdílková et al., 2019) and (Pozdílková and Marek, 2021). In its general form, information was obtained on the location, floor area, asking price, and type of apartment. Consequently, no straightforward algorithm was available for directly determining the average apartment price in a particular

[1] University of Pardubice, Department of Mathematics and Physics, Studentská 95, 53210 Pardubice, jaromir.zahradka@upce.cz.
[2] University of Pardubice, Department of Mathematics and Physics, Studentská 95, 53210 Pardubice, jaroslav.marek@upce.cz.

municipality, and the price per square meter had to be calculated. The specific data source consisted of apartment advertisements published on the sreality.cz website.

The analysis was conducted for the following districts of the Moravian-Silesian Region: Bruntál (BR), Frýdek-Místek (FM), Karviná (KA), Nový Jičín (NJ), Opava (OP), and Ostrava (OS). Apartment advertisements were processed on 2,708 days out of a total of 2,799 days during the monitored period from August 2018 to March 2026. On 91 days, data were not obtained due to application startup errors. For example, in the Nový Jičín district, a total of 210,836 advertisements were processed during the monitored period. The average daily number of advertisements in this district was 78. The number of days for which a particular apartment remained advertised was not monitored.

The values of the explanatory variable representing mortgage interest rates were obtained from data publicly available through the website of the Czech Statistical Office.

2.1 Regression model

The simultaneous equations framework was selected because apartment prices and housing supply are mutually dependent. Higher prices may attract additional sellers and increase the number of advertisements, while the volume of advertisements can influence market expectations and future price development. Estimating separate regression models would therefore lead to endogeneity problems and potentially biased parameter estimates. Simultaneous equation models allow these feedback mechanisms to be explicitly represented within the modelling framework.

Consider the model of two simultaneous equations

$$\begin{aligned} S_t &= \alpha_0 + \alpha_1 \cdot S_{t-1} + \alpha_2 \cdot P_{t-1} + \alpha_3 \cdot M_t + \varepsilon_{1,t} \\ P_t &= \beta_0 + \beta_1 \cdot S_t + \beta_2 \cdot P_{t-1} + \beta_3 \cdot M_t + \varepsilon_{2,t} \end{aligned} \tag{1}$$

with the parameters given in Table 1.

Parameter	Description
S	column vector of number of advertised apartments in the region
P	column vector of average prices for 1 m^2 apartment in the location. i.e., measurements of prices $y_1 \ldots, y_n$, that are realized in the region
M	column vector of mortgage interest in time (month) 1, 2…, T
t	time, index of month
$\propto, \beta$	vector of unknown parameters
ε	column vector of independent possibly heteroscedastic errors

Table 1 Parameters of the model and their description. Source: own

The first equation of the simultaneous equations model (1) was constructed to describe the dependence of the number of advertisements on the prices of apartments in the previous period and on the interest rate. The second equation was designed to model the dependence of the price at time t on the estimated number of advertisements and the mortgage interest rate. A similar model for modeling apartment prices was used in the article (Koťátková - Stránská, Marek, 2015), where other variables were used.

3 Results and discussion

Due to the scope of this article, only a brief descriptive analysis of the relationships among the number of advertised apartments, mortgage interest rates, and apartment prices is presented. See Tab. 2.
With one exception (the district of Nový Jičín, where the correlation between the number of advertisements and the price per square meter was not statistically significant), all correlation coefficients were statistically significant at the selected significance level. In most districts, the number of advertised apartments was negatively correlated with the price per square meter, indicating that a higher supply of apartments was generally associated with lower prices. The only exception was the Bruntál district, where a statistically significant positive correlation was observed, suggesting a different market dynamic compared with the other districts.
A positive correlation between the number of advertisements and mortgage interest rates was found in all districts. Similarly, apartment prices and mortgage interest rates were positively correlated throughout the region. This

somewhat counterintuitive result may reflect the strong long-term upward trend in housing prices during the observed period, which outweighed the dampening effect of higher borrowing costs.

Table 1 summarizes the correlation coefficients between the number of advertised apartments, the average price per square meter, and mortgage interest rates for the districts of the Moravian-Silesian Region.

District Correlation Coeffitient	BR	FM	KA	NJ	OP	OS
Monthly number of adveretisment – square meter price	0.41*	-0.36*	-0.33*	0.02	-0.30*	-0.28*
Monthly number of adveretisment – mortgage interest rate	0.85*	0.27*	0.33*	0.66*	0.34*	0.36*
Square meter price – mortgage interest rate	0.70*	0.57*	0.62*	0.65*	0.68*	0.60*

Table 2 The correlation coefficients between the number of advertisements, the price per square meter and mortgage interest rate in discricts of Moravian-Silesian Region

For the collected data, we first plotted several graphs: first, we showed the development of the number of advertisements, the average price per 1 m2 and the mortgage interest rate for the period 08/2018 to 03/2026. See Figure 1, 2 and 3. The six districts of the Moravian-Silesian Region are shown in color as follows: Bruntál (BR) in blue, Frýdek Místek (FM) in brown, Karviná (KA) in red, Nový Jičín (NJ) in green, Opava (OP) in yellow, Ostrava (OS) in black.

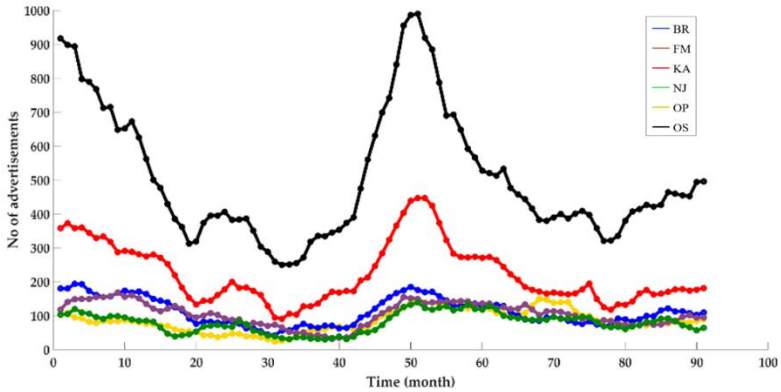


Figure 1 Time series of advertisement number w.r.t. time

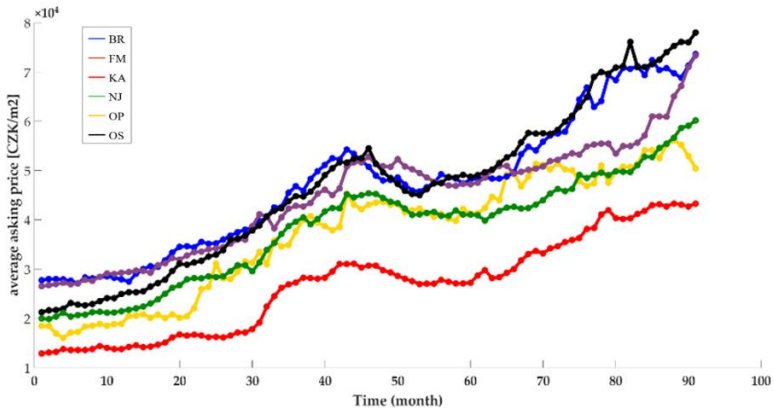


Figure 2 Time series of average price w.r.t. time

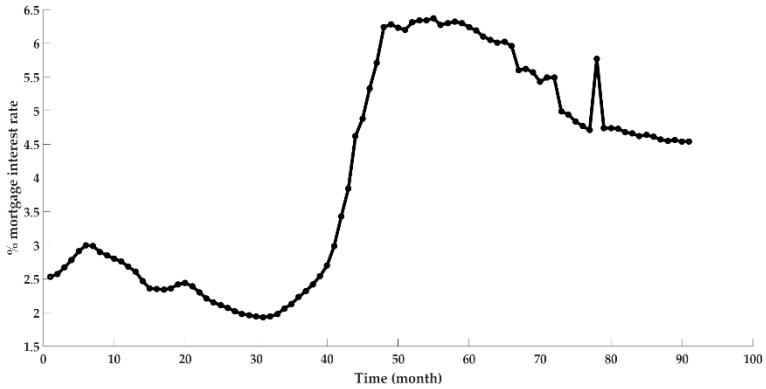


Figure 3 Time series of mortgage interest w.r.t. time

Figure 4 shows the number of advertisements on the mortgage interest rate. Figure 5 shows the dependence of the average price of 1 square meter of an apartment on the number of advertised apartments. Figure 6 shows the dependence of the average price of 1 square meter of an apartment on the mortgage interest rate. When looking at these dependencies, it is important to note the turning points. If we focus on their determination, we obtain the following months in all districts: 01-03/2021 and 01-03/22. The first turning point corresponds to the beginning of the Covid19 pandemic. The second turning point was probably determined by an increase in the interest rate.

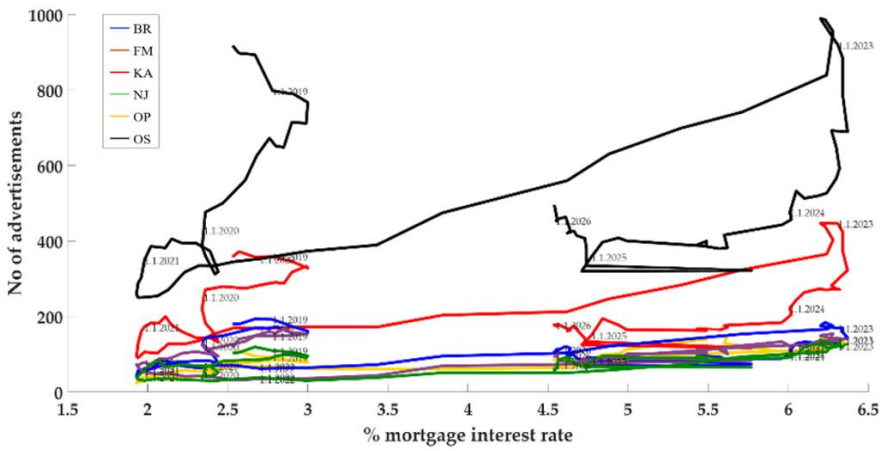


Figure 4 Time series of No of advertisement w.r.t. mortgage interest

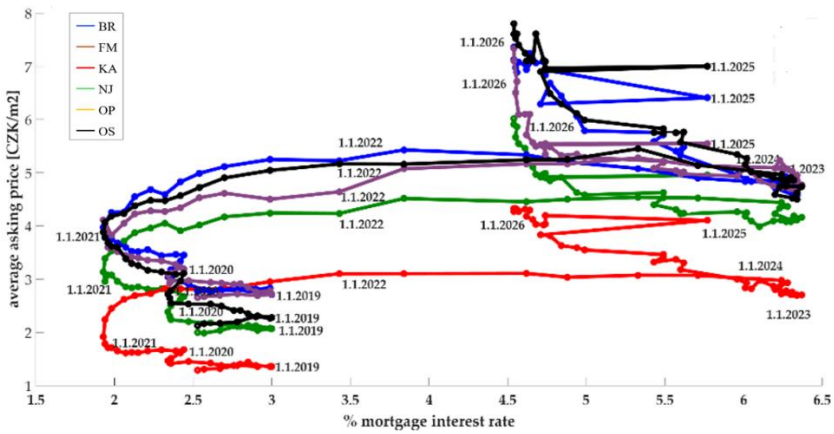


Figure 5 Time series of Price w.r.t.. mortgage interest

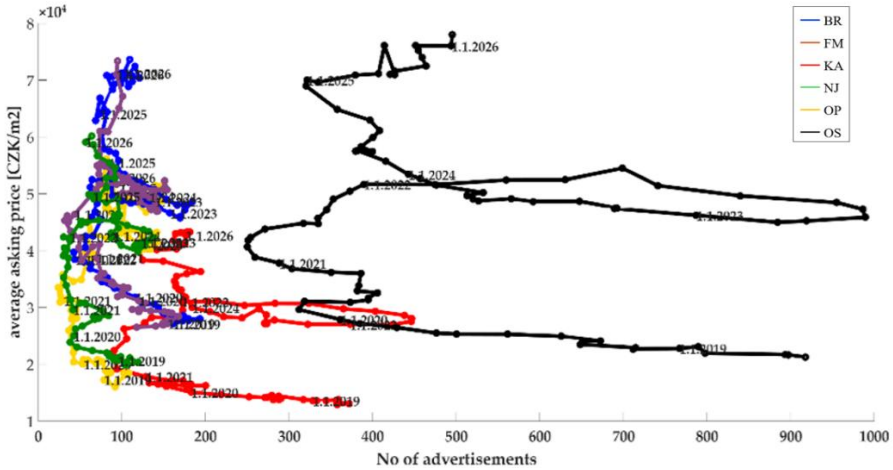


Figure 6 Time series of price w.r.t. advertisement number

From the collected data, the unknown parameters of the regression model (1) were estimated, see Table 3. Parameter estimates that were significant are marked with an asterisk in the table. The determination index was in the range of 0.95 to 0.99, which expresses the high ability of the model to explain the dependencies between variables. It is evident that the number of advertisements and the price were most influenced by the mortgage interest rate. A special situation occurred in the Bruntál and Karviná districts, where the beta coefficients were not statistically significant for either the number of advertisements or the mortgage interest rate.

Parameter / **District**	α_0	α_1	α_2	α_3	β_0	β_1	β_2	β_3	R^2
BR	2.14	0.83*	-0.0001	3.60*	1876.76*	-28.67	0.96*	562.09	0.96
FM	-6.18	0.99*	0.0002	-0.47	2530.42*	-15.06*	0.98*	125.18*	0.95
KA	-16.33	1.02*	0.0010*	-3.71	1661.93*	-5.15	0.98*	90.21	0.97
NJ	10.53*	0.79*	-0.0003*	4.54*	883.26*	-7.75	1.02*	-130.37*	0.96
OP	12.17	0.90*	-0.0003	2.61*	-1347.69*	5.72	1.08*	-516.25*	0.96
OS	-36.95*	1.01*	0.0013*	-7.30*	987.14*	-5.42	1.01*	-117.79*	0.98

Table 3 Estimated parameters. The statistically significant parameters are marked by an asterisk ($p < 0.05$)

Using the regression model (1), the expected price in the next period was estimated for several variants of mortgage interest rates in the range of 2-6 percent and several variants of the number of advertisements. See Figure 7. The relative changes of prices were shown using contours (relative changes shown for isolines are in percentages). The dependence of supply (number of advertisements) and price estimate is shown for Ostrava district and for Nový Jičín district. The red asterisk indicates the current interest rate of 4.54%. The red upward arrow shows what change in the interest rate would not lead to a change in the price w.r.t. number of advertisements.

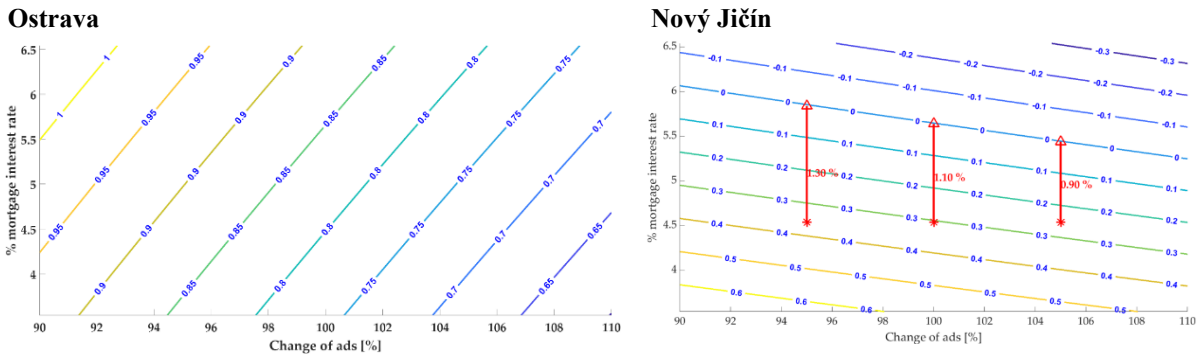


Figure 7 Izolines of Ostrava and Nový Jičín

4 Conclusions

The proposed simultaneous equations model provided a satisfactory description of the relationships among apartment prices, housing supply and mortgage interest rates in the selected districts of the Moravian-Silesian Region. The obtained coefficients of determination indicate a strong explanatory power of the model. The results suggest that apartment prices remained relatively resistant to increases in housing supply during the observed period, while mortgage rates exhibited a more complex relationship with prices.

The findings are consistent with recent studies highlighting regional heterogeneity of housing markets and the importance of economic conditions in explaining real estate price dynamics (Hromada et al., 2025), and (Hromada et al., 2023). Future research may extend the proposed framework by incorporating additional explanatory variables, such as population development, unemployment rates, household income or the number of economic entities. Moreover, machine learning approaches recently applied in housing market analysis (Forradellas and Benitez, 2026), (Cellmer and Kobylińska, 2024), and (Çılgın and Gökçen, 2023) could be compared with the econometric framework proposed in this study.

Acknowledgements

The paper was supported by institutional support of University Pardubice.

References

Cellmer R, Kobylinska K (2024) Housing Price Prediction - Machine Learning and Geostatistical Methods. Online. Real Estate Management and Valuation 33(1): 1-10. ISSN 2300-5289. Available from: https://doi.org/10.2478/remav-2025-0001. [cit. 2026-06-08]

Çilgin C, Gökċen H (2023) Machine learning methods for prediction real estate sales prices in Turkey. Online. Revista de la construccion. Journal of Construction 22(1): 163-177. Available from: https://doi.org/10.7764/rdlc.22.1.163. [cit. 2026-06-08].

Forradellas R F R, Benitez A G (2026) Development of a Predictive Tool for Real Estate Analysis Using Machine Learning Techniques. Online. International Journal of Financial Studies 14(5): 130-130. ISSN 2227-7072. Available from: https://doi.org/10.3390/ijfs14050130. [cit. 2026-06-08]

Hromada E, Čermíková K, Kureková L, Kadeřábková B (2025) On sustainable urban development - mapping real estate trends across economically differentiated regions. Online. Economics and Environment 92(1): 859-859. ISSN 2957-0387. Available from: https://doi.org/10.34659/eis.2025.92.1.859. [cit. 2026-06-08]

Hromada E, Schneiderová Heralová R, Čermáková K, Piecha M, Kadeřábková B (2023) Impacts of Crisis on the Real Estate Market Depending on the Development of the Region. Online. Buildings 13(4): 896-896. ISSN 2075-5309. Available from: https://doi.org/10.3390/buildings13040896. [cit. 2026-06-08]

Koťátková-Stránská P, Marek J (2015) Simultaneous equations and modelling of new flat demand. In: Mezinárodní vědecká konference Hradecké ekonomické dny 2015 (Jedlička, P., ed.). Univerzita Hradec Králové, Fakulta informatiky a managementu, Hradec Králové, díl I, pp 11–16

Pozdílková A, Marek J (2021) Data Mines in Real Estate Web Pages: Investigation of Changes in the Czech Real Estate Market Based on Elasticity and on Modified Price Volume Indicator. Online. In: Sanjurjo González, H., Pastor López, I., García Bringas, P., Quintián, H., Corchado, E. (eds.) 16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2021). Advances in Intelligent Systems and Computing, vol 1401. Springer Nature Switzerland AG, Cham, pp 155-164

Pozdílková A, Nedvědová M, Marek J (2019) An Approach to Web Advertising Server parsing - Apartment Price Analysis in the Czech Republic. In: 15th International Scientific Conference on Informatics INFORMATICS 2019: proceedings, pp 333-337

Slaný M, Emmer F (2024) Evidence of the ripple effect across the Czech housing market. Online. International Journal of Economic Sciences 13(2): 104-115. ISSN 1804-9796. Available from https://doi.org/10.52950/es.2024.13.2.006. [cit. 2026-06-08]

Mortgage interest rate chart. [online], [cit. 15. 1. 2020]. URL: https://www.kurzy.cz/hypoteky/graf-sazeb-hypotek/

A DEA-Bootstrap Framework for Assessing Educational Quality in Industrial Engineering Programs

Rohemi Zuluaga[1], Carlos Garcia-Yerena[2], Enrique Delahoz-Domínguez[3]

Abstract. This study evaluates the relative efficiency of industrial engineering programs in Colombia by integrating Data Envelopment Analysis (DEA) with the smoothed bootstrap, which yields bias-corrected estimates and confidence intervals. Under constant and variable returns to scale (CRS, VRS), an output-oriented model relates the generic competencies to the discipline-specific professional competencies measured by the Saber PRO national examination across 92 universities. Because both competency blocks are assessed in the same examination at the end of the program, the model is interpreted as a cross-sectional description of each program's attainment structure rather than as a transformation of entry-level inputs into outcomes. The results reveal wide dispersion in the alignment between the two competency levels, and show that a substantial share of the measured inefficiency is associated with the scale of operation rather than with purely managerial factors. A complementary clustering and principal component analysis organizes programs into two action profiles supporting tailored improvement strategies.

Keywords: educational evaluation, efficiency, Data Envelopment Analysis, bootstrap, learning analytics

JEL Classification: C44

1 Introduction

Engineering education needs operational tools to assess and assure program quality (Duque Oliva & Chaparro Pinzón, 2012), since that quality determines its capacity to produce professionals able to meet the changing demands of the industrial sector (Ameen Abdulla et al., 2020; Zuluaga-Ortiz et al., 2023). Robust program-level methodologies remain scarce, which impedes the identification of improvement opportunities and the design of targeted curricular action.

This study proposes a framework combining Data Envelopment Analysis (DEA) with the bootstrap of Simar and Wilson (1998) to evaluate the relative efficiency of Colombian industrial engineering programs from competency-based standardized test data. The contribution is threefold: (i) the first bias-corrected DEA-bootstrap assessment of these programs based on the Saber PRO national examination; (ii) an interpretation of efficiency as the degree of alignment between the generic and the discipline-specific competency levels attained by graduates; and (iii) action profiles that translate the efficiency estimates into improvement strategies for program managers.

2 Literature Review

Since Charnes, Cooper and Rhodes (1978) introduced the CCR model under constant returns to scale (CRS) and Banker, Charnes and Cooper (1984) extended it to variable returns to scale (VRS), DEA has become a standard non-parametric technique for efficiency measurement, and higher education a natural field of application, since universities combine multiple inputs and outputs in the absence of market prices. Buitrago et al. (2017) review DEA applications in higher education institutions and Salas-Velasco (2020) reports a system-level analysis for Spain. In Latin America, DEA has been applied to Colombian public institutions (Visbal-Cadavid et al., 2016) and Argentine universities (Villarreal & Tohmé, 2017), while alternative frontier approaches such as StoNED have also been explored (Arévalo et al., 2022). For engineering education specifically, earlier work combines DEA with partial least squares (Delahoz-Dominguez et al., 2022), Six Sigma tools (Ameen Abdulla et al., 2020; Delahoz-Dominguez et al., 2023) and efficiency analysis trees (Zuluaga-Ortiz et al., 2023).

[1] Universidad Tecnológico de Comfenalco, Cartagena, Colombia. zuluagaortizra@tecnocomfenalco.edu.co. ORCID: 0000-0002-7245-8761.

[2] Universidad del Magdalena, Santa Marta, Colombia. cgarciaey@unimagdalena.edu.co. ORCID: 0000-0002-9973-552X.

[3] Universidad del Magdalena, Santa Marta, Colombia. enriquedelahoz@unimagdalena.edu.co. ORCID: 0000-0003-2531-6389.

A well-known limitation of deterministic DEA is that its scores are point estimates, sensitive to sampling variation and biased towards the frontier. Simar and Wilson (1998, 2000) proposed a bootstrap that corrects this bias and yields confidence intervals, recently applied to Greek secondary schools (Kounetas et al., 2023). To the best of our knowledge, no previous study has combined bias-corrected DEA-bootstrap estimation with clustering and PCA to assess the competency-based efficiency of industrial engineering programs. This paper addresses that gap.

2.1 Data

The analysis draws on the Saber PRO national examination results of 4,976 students enrolled in industrial engineering programs at 92 Colombian universities, taken from the open dataset compiled by Delahoz-Dominguez et al. (2020) from records of the Colombian Institute for the Evaluation of Education (ICFES). The examination results correspond to the 2018 Saber PRO cycle assessment cycle.

Table 1 reports the descriptive statistics of the eight modules used in the model.

Module	Code	Mean	Median	Std. dev.	Range	Min	Max
Critical Reading	CR	53.94	51.74	13.50	59.32	25.13	84.44
Citizenship Skills	CS	52.52	50.75	12.59	64.06	17.63	81.68
English	ENG	60.08	59.48	15.28	74.61	20.00	94.61
Written Communication	WC	52.48	51.64	9.53	57.78	19.67	77.45
Quantitative Reasoning	QR	67.79	67.40	13.62	73.80	20.67	94.47
Scientific and Statistical Mathematical Thinking	MSST	133.71	131.36	13.06	59.64	114.47	174.12
Design of Logistics and Production Processes	DPLS	147.80	145.80	16.59	93.94	99.00	192.94
Formulation of Engineering Projects	FEP	150.83	148.18	14.23	78.24	109.00	187.24

Table 1 Descriptive statistics of the Saber PRO modules used in the model (student level, n = 4,976).

The structure of the examination. Saber PRO is a single, compulsory examination administered to students in the final year of their undergraduate program. It comprises a generic module common to all fields of study — Critical Reading (CR), Citizenship Skills (CS), English (ENG), Written Communication (WC) and Quantitative Reasoning (QR) — and a set of discipline-specific modules which, for industrial engineering, are Scientific and Statistical Mathematical Thinking (MSST), Design of Logistics and Production Processes (DPLS) and Formulation of Engineering Projects (FEP). It must be stressed that both blocks are measured in the same sitting and at the same point in the program. The generic scores are therefore contemporaneous attainment measures, not observations of the competency level with which students entered the program. The two blocks are reported on different score scales (Table 1); since radial DEA efficiency measures are invariant to the units of measurement of inputs and outputs, this does not affect the estimated scores.

Aggregation to the university level. The decision-making units (DMUs) are the 92 universities, not the individual students. For each university and each module, the institutional score is the unweighted arithmetic mean of the scores of its students in the dataset. Cohort sizes are unequal, ranging from 6 to 42 students per university, with a median of 25.

2.2 DEA-Bootstrap model specification

The model considers five inputs, corresponding to the generic competencies (CR, CS, ENG, WC, QR), and three outputs, corresponding to the discipline-specific professional competencies (MSST, DPLS, FEP). The sample size satisfies the usual discrimination rule of thumb, since the number of DMUs (92) exceeds three times the sum of the numbers of inputs and outputs (3 × 8 = 24).

Let $x_k \in \mathbb{R}^5$ and $y_k \in \mathbb{R}^3$ denote the input and output vectors of university k, for $k = 1, \ldots, n$ with $n = 92$. The output-oriented radial DEA problem is

$$F_k = max \{ \varphi : \varphi\, y_k \leq \Sigma_j \lambda_j y_j,\ \ \Sigma_j \lambda_j x_j \leq x_k,\ \ \lambda_j \geq 0 \} \tag{1}$$

where the CRS specification (Charnes et al., 1978) imposes no further restriction on the intensity vector λ, while the VRS specification (Banker et al., 1984) adds the convexity constraint $\Sigma_j \lambda_j = 1$. The comparison between the two allows technical inefficiency to be decomposed into pure managerial inefficiency and scale inefficiency.

Score convention. The optimal value $F_k \geq 1$ in (1) is the Farrell output-expansion factor. Throughout this paper, and consistently with the output of the *Benchmarking* package used for estimation (Bogetoft & Otto, 2011), efficiency is reported in the Shephard convention

$$\theta_k = 1 / F_k \in (0, 1], \tag{2}$$

so that $\theta_k = 1$ identifies a program on the estimated frontier. Under this convention the feasible proportional expansion of all outputs, holding inputs fixed, is

$$\Delta_k = (F_k - 1) \times 100\% = (1 / \theta_k - 1) \times 100\%. \tag{3}$$

It is important to note that Δ_k is *not* equal to $(1 - \theta_k) \times 100\%$; the latter is the correct reading of a score of θ_k only under an input-oriented specification, where it measures the feasible proportional reduction of inputs. Equation (3) is used for all output-expansion statements in Section 4.

Bootstrap. To correct the well-known finite-sample bias of the DEA estimator, the homogeneous smoothed bootstrap of Simar and Wilson (1998, 2000) is applied with $B = 2{,}000$ replications. Denoting by $\hat{\theta}_k$ the deterministic estimate and by $\theta^*_{k,b}$ its value in replication b, the estimated bias and the bias-corrected score are

$$bias_k = B^{-1} \Sigma_b \theta^*_{k,b} - \hat{\theta}_k, \quad \theta_k^{bc} = \hat{\theta}_k - bias_k. \tag{4}$$

Percentile confidence intervals at the 95% level, $[L_k, U_k]$, are obtained from the bootstrap distribution. The associated measure of estimation uncertainty used in this paper is the interval width

$$w_k = U_k - L_k. \tag{5}$$

We distinguish this quantity carefully from bias_k in (4): the bias measures the systematic distortion of the point estimator, whereas w_k measures the dispersion of its sampling distribution. The two are related, since both grow when the estimated frontier is sensitive to the resampled sample, but they are not interchangeable and are not used interchangeably below. All computations were performed in R with the *Benchmarking* package (Bogetoft & Otto, 2011). An output-oriented model is adopted because the discipline-specific modules are the outcomes that curricular and pedagogical decisions are most directly intended to affect, whereas the generic competencies are transversal and are shaped to a large extent by prior schooling and by the cognitive foundations that students bring with them. The orientation reflects the direction in which program-level action is meaningful; it is not a claim that generic competencies are exogenous inputs observed before enrolment.

Because generic and professional competencies are measured in the same examination and at the same moment, the model does not identify a transformation over time, and we do not claim that universities *transform* entry-level competencies into professional ones. What the model measures is the extent to which a program's discipline-specific attainment is high *relative to* that of programs whose graduates display comparable generic-competency profiles. A low score identifies a program whose professional-competency results fall short of what peer programs achieve at similar generic-competency levels. This is a cross-sectional, associational statement about the structure of attainment, and it should not be read as evidence about value added, which would require measurements taken at the beginning of the study program. We note that the source dataset (Delahoz-Dominguez et al., 2020) also contains Saber 11 end-of-secondary results for the same students.

3 Results

3.1 Efficiency analysis

Table 2 summarises the bias-corrected efficiency estimates. Under the CRS specification only two units (2.17%) lie on the estimated frontier and the mean score is low; the VRS specification identifies the same number of frontier units but yields a substantially higher average level of efficiency.

Statistic	CRS	VRS	Scale efficiency
Frontier DMUs (%)	2 (2.17%)	2 (2.17%)	2 (2.17%)
Mean	0.62	0.81	0.78
Standard deviation	0.11	0.11	0.16
Minimum	0.46	0.61	0.48
Median	0.60	0.79	0.80
Third quartile	0.66	0.88	0.94

Table 2 Summary of bias-corrected DEA-bootstrap efficiency estimates (Shephard convention, $\theta \in (0,1]$; n = 92). Scale efficiency is computed as the ratio of the CRS to the VRS score

In educational terms, these results indicate substantial room for improvement. Applying equation (3), a program whose bias-corrected CRS score equals the sample mean ($\theta = 0.62$) could expand all three professional-competency outputs by a factor of 1/0.62 = 1.61, that is, by approximately **61%**, while holding its students' generic-competency levels unchanged. The corresponding figure under VRS, at the mean score of 0.81, is approximately 24%. Because the transformation in (3) is non-linear, these values are expansion factors evaluated *at* the mean score and are not the mean of the individual expansion factors; the latter is necessarily larger by Jensen's inequality. Only two programs (2.17%) define the best-practice frontier and act as benchmarks for the remaining institutions. Moreover, the gap between the mean CRS score (0.62) and the mean VRS score (0.81) reveals that an important share of the measured inefficiency is associated with the scale of operation rather than purely managerial factors, suggesting different improvement levers for different types of institutions.

Table 3 reports the results for a sample of 10 DMUs. The columns show, for each returns-to-scale specification, the deterministic estimate $\hat{\theta}$, the bias-corrected estimate $\hat{\theta}^{bc}$, the estimated bias from equation (4), and the lower and upper limits of the 95% percentile confidence interval.

DMU	CRS model					VRS model				
	$\hat{\theta}$	$\hat{\theta}$ bc	Bias	CI low	CI up	$\hat{\theta}$	$\hat{\theta}$ bc	Bias	CI low	CI up
U1	0.76	0.71	0.04	0.65	0.75	0.78	0.74	0.04	0.69	0.78
U2	0.60	0.55	0.05	0.50	0.59	0.63	0.59	0.05	0.55	0.63
U3	0.89	0.83	0.06	0.76	0.89	0.94	0.90	0.04	0.85	0.94
U4	0.75	0.68	0.06	0.62	0.74	0.79	0.74	0.05	0.70	0.79
U5	0.63	0.53	0.10	0.48	0.62	0.77	0.70	0.07	0.64	0.76
U6	0.66	0.60	0.05	0.55	0.65	0.69	0.64	0.05	0.60	0.68
U7	0.67	0.61	0.06	0.56	0.66	0.83	0.79	0.04	0.75	0.83
U8	0.66	0.61	0.05	0.55	0.65	0.66	0.62	0.05	0.57	0.66
U9	0.63	0.55	0.09	0.50	0.62	0.65	0.58	0.06	0.54	0.64
U10	0.57	0.53	0.04	0.49	0.57	0.88	0.83	0.06	0.78	0.88

Table 3 Sample of results for 10 DMUs. $\hat{\theta}$ is the deterministic estimate, $\hat{\theta}$ bc the bias-corrected estimate, and CI low / CI up the limits of the 95% bootstrap percentile interval

3.2 Exploratory analysis: clustering and PCA

Silhouette analysis gives a maximum average width of 0.486 at k = 2, the number of clusters retained. The two groups differ little in mean CRS efficiency and only slightly in VRS efficiency, but Cluster 1 shows greater dispersion in the bias corrections of both models. A PCA of the four efficiency variables (SC, BC, SV, BV) was then computed on the correlation matrix — all variables centred and scaled to unit variance beforehand, since the

scores and the bias values differ by an order of magnitude in dispersion. The first two components retain 57.9% and 26.0% of the total variance (83.9% jointly).

Table 4 reports each variable's percentage contribution to these dimensions, that is, the share of the component's variance attributable to the variable, so that columns sum to 100 by construction; these are not loadings, which would be variable–component correlations and would not sum to a fixed total. Dimension 1 is dominated jointly by BC and BV, Dimension 2 mainly by SC and, to a lesser extent, SV.

Variable	Contribution to Dim. 1 (%)	Contribution to Dim. 2 (%)
Score efficiency CRS (SC)	14.28	51.58
Bias CRS (BC)	32.46	7.75
Score efficiency VRS (SV)	20.67	31.13
Bias VRS (BV)	32.60	9.54
Total	**100.00**	**100.00**

Table 4 Percentage contribution of each variable to the first two principal components. Columns sum to 100 by construction; these are contributions, not loadings

The Dimension 1 is dominated by the bootstrap bias of the two models and is accordingly interpreted as a bias-magnitude axis: it separates programs whose deterministic scores required a large downward correction from those whose scores were already close to their bias-corrected values. We deliberately avoid describing this dimension as an uncertainty dimension. The bootstrap bias measures a systematic distortion of the point estimator, not the dispersion of its sampling distribution; the corresponding uncertainty measure is the confidence-interval width w_k of equation (5). Dimension 2 is driven mainly by the CRS and VRS scores and represents an efficiency-level axis.

Figure 1 shows the two-dimensional representation. The observations in quadrants 1, 2, 3 and 4 are distributed as 17 (18.48%), 26 (28.26%), 8 (8.70%) and 41 (44.56%), respectively. Quadrant 2 groups programs that combine high efficiency with small bias corrections, so that their position near the frontier is stable across bootstrap replications; these constitute the most reliable benchmarks. Quadrant 1 contains programs with high efficiency but large bias corrections, whose apparent performance rests on a frontier segment that is sensitive to the composition of the sample and should be confirmed with additional data. Quadrant 3 corresponds to programs whose low scores are robust to resampling. Finally, quadrant 4 combines low efficiency with large bias corrections and concentrates the priority cases for intervention.

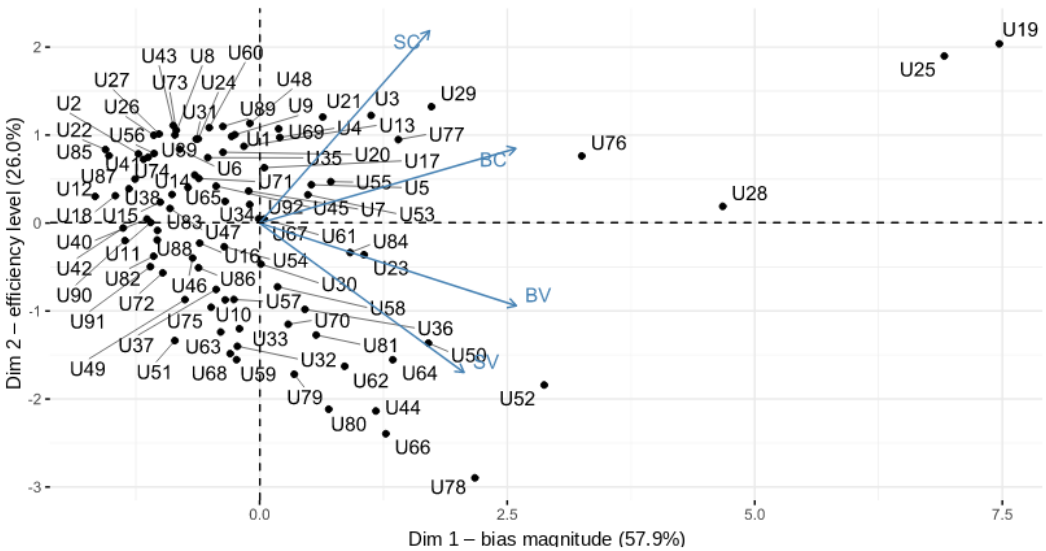


Figure 1 Graphical representation of the PCA in two dimensions. Dimension 1 is the bias-magnitude axis; Dimension 2 is the efficiency-level axis

3.3 Action profiles

Combining the cluster analysis with the PCA quadrants allows the 92 programs to be organized into two action profiles that translate the efficiency results into improvement paths.

Profile 1 — consolidation. Programs located in quadrants 1 and 2, which are efficient or close to the frontier, should focus on sustaining their performance: documenting and disseminating their pedagogical practices, acting as reference institutions for benchmarking, and targeting the specific competencies in which residual slack remains (for instance English, which exhibits the widest dispersion among the generic competencies in Table 1).

Profile 2 — structural improvement. Programs in quadrants 3 and 4, with low bias-corrected scores, require deeper interventions: reviewing the curricular articulation between the generic and the discipline-specific components, reinforcing quantitative reasoning as a driver of the professional modules, and evaluating scale-related decisions in the light of the CRS–VRS gap.

The bootstrap intervals provide a useful frame of reference, but the upper bound Uk is not a performance target: it bounds the plausible values of a program's true efficiency given sampling variation in the estimated frontier and carries no information about what managerial action can achieve, nor any guarantee of attainability. We use it only as an exploratory reference value, indicating that apparent improvements remaining below Uk cannot be distinguished from estimation noise. Operational targets should instead come from the bias-corrected scores of the peers in the reference set (the λ-weighted peers of equation (1)) of each inefficient DMU, since these correspond to observed and therefore demonstrably attainable practice.

4 Discussion and Conclusion

Combining DEA with the Simar–Wilson bootstrap yields bias-corrected efficiency estimates together with an explicit measure of their sampling variability, and estimating both CRS and VRS allows the measured inefficiency to be decomposed. The results reveal wide dispersion in the alignment between the generic and the discipline-specific competency levels attained by graduates, together with a substantial scale-related component.

The two action profiles identify where intervention is most likely to matter and provide a basis for improvement plans tailored to each program. Relative to deterministic DEA (Salas-Velasco, 2020), the bootstrap approach makes the reliability of each estimate explicit (Kounetas et al., 2023), which is what allows the reliable benchmarks of quadrant 2 to be distinguished from the sample-sensitive cases of quadrant 1.

There are four limitations in the study. First, both competency blocks are measured in the same end-of-program examination, so the results describe the cross-sectional structure of attainment rather than value added; a longitudinal specification using the Saber 11 entry scores available in the same source dataset is the priority extension. Second, institutional scores are means over cohorts of unequal size, and the bootstrap intervals do not incorporate the sampling error of those means. Third, the model uses test results only, omitting institutional resources, faculty qualifications and student socioeconomic composition, so part of the measured inefficiency may reflect these unobserved factors. Fourth, a single examination cycle permits no inference about trends.

References

Ameen Abdulla M S, Mohammed Navas O P, Amal M S, Nizam B, Kavilal E G (2020) Quality assurance in education based on six sigma tool. Proceedings of the International Conference on Industrial Engineering and Operations Management, pp 2347–2360

Arévalo A, Giménez V, Prior D (2022) Análisis de eficiencia en educación: Una aplicación del método StoNED. Revista Desarrollo y Sociedad 92, https://doi.org/10.13043/DYS.92.2

Banker R D, Charne, A, Cooper W W (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30(9): 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078

Bogetoft P, Otto L (2011) Benchmarking with DEA, SFA, and R. Springer. https://doi.org/10.1007/978-1-4419-7961-2

Buitrago O, Espitia A, Molano L (2017) Análisis envolvente de datos para la medición de la eficiencia en instituciones de educación superior: Una revisión del estado del arte. Revista Científica General José María Córdova 15(19): 147–173. https://doi.org/10.21830/19006586.84

Charnes A, Cooper W W, Rhodes E (1978) Measuring the efficiency of decision making units. European Journal of Operational Research 2(6): 429–444. https://doi.org/10.1016/0377-2217(78)90138-8

Clayman C L, Srinivasan S M, Sangwan R S (2020) K-means clustering and principal components analysis of microarray data of L1000 landmark genes. Procedia Computer Science 168: 97–104. https://doi.org/10.1016/j.procs.2020.02.265

Delahoz-Dominguez E, Zuluaga R, Fontalvo-Herrera T (2020) Dataset of academic performance evolution for engineering students. Data in Brief 30: 105537. https://doi.org/10.1016/j.dib.2020.105537

Delahoz-Dominguez E, Zuluaga-Ortiz R, Camelo-Guarín A, Suarez-Sánchez M (2022) Performance evaluation of mechanical engineering degrees using partial minimum squares and data envelopment analysis. International Journal of Innovation and Learning 32(4): 397–413. https://doi.org/10.1504/IJIL.2022.126635

Delahoz-Dominguez E, Zuluaga-Ortiz R, Suarez-Sánchez M (2023) A Six Sigma and DEA approach for learning outcomes assessment at industrial engineering programs. International Journal of Six Sigma and Competitive Advantage 14(3): 279–299. https://doi.org/10.1504/IJSSCA.2023.130283

Duque Oliva E J, Chaparro Pinzón C R (2012) Medición de la percepción de la calidad del servicio de educación por parte de los estudiantes de la UPTC Duitama. Criterio Libre 10(16): 159–192

Granato D, Santos J S, Escher G B, Ferreira B L, Maggio R M (2018) Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. Trends in Food Science & Technology 72: 83–90. https://doi.org/10.1016/j.tifs.2017.12.006

Kounetas K, Androulakis G, Kaisari M, Manousakis G (2023) Educational reforms and secondary school's efficiency performance in Greece: A bootstrap DEA and multilevel approach. Operational Research 23(1): 9. https://doi.org/10.1007/s12351-023-00764-y

O'Flaherty J, Liddy M (2018) The impact of development education and education for sustainable development interventions: A synthesis of the research. Environmental Education Research 24(7): 1031–1049. https://doi.org/10.1080/13504622.2017.1392484

Salas-Velasco M (2020) The technical efficiency performance of the higher education systems based on data envelopment analysis with an illustration for the Spanish case. Educational Research for Policy and Practice 19(2): 159–180. https://doi.org/10.1007/s10671-019-09254-5

Simar L, Wilson P W (1998) Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. Management Science 44(1): 49–61. https://doi.org/10.1287/mnsc.44.1.49

Simar L, Wilson P W (2000) A general methodology for bootstrapping in non-parametric frontier models. Journal of Applied Statistics 27(6): 779–802. https://doi.org/10.1080/02664760050081951

Villarreal F, Tohmé F (2017) Análisis envolvente de datos. Un caso de estudio para una universidad argentina. Estudios Gerenciales 33(144): 302–308. https://doi.org/10.1016/j.estger.2017.06.004

Visbal-Cadavid D, Mendoza Mendoza A, Causado Rodríguez E (2016) Eficiencia en las instituciones de educación superior públicas colombianas: Una aplicación del análisis envolvente de datos. Civilizar Ciencias Sociales y Humanas 16(30): 105–118

Zuluaga-Ortiz R, Camelo-Guarin A, Delahoz-Domínguez E (2023) Efficiency analysis trees as a tool to analyze the quality of university education. International Journal of Electrical and Computer Engineering 13(4): 4412–4421. https://doi.org/10.11591/ijece.v13i4

Authors	Collective of authors
Editors	Radek Němec, Lucie Chytilová
Department	Department of Systems Engineering
Title	Proceedings of the 44th International Conference on Mathematical Methods in Economics 2026
Place, Year, Edition	Ostrava, 2026, 1st ed.
Number of pages	401
Publisher	VŠB-Technical University of Ostrava, Faculty of Economics, Department of Systems Engineering and Informatics, Czech Republic
Publication form	On-line; published as publicly accessible through URL
Production	Department of Systems Engineering and Informatics
Cover design	Markéta Šindlerová
ISBN 978-80-248-4880-8 (on-line)	
ISSN 2788-3965	

The proceedings publication, in all its forms, may not be sold separately. Duplication of the proceedings content and/or public sharing of the proceedings media-carrier content is a subject of the Copyright.

Unauthorized duplication and/or public sharing may be treated as a direct infringement of the Copyright law!

44TH INTERNATIONAL CONFERENCE

MME2026

Mathematical Methods in Economics

FACULTY OF ECONOMICS

VSB – TUO

OSTRAVA