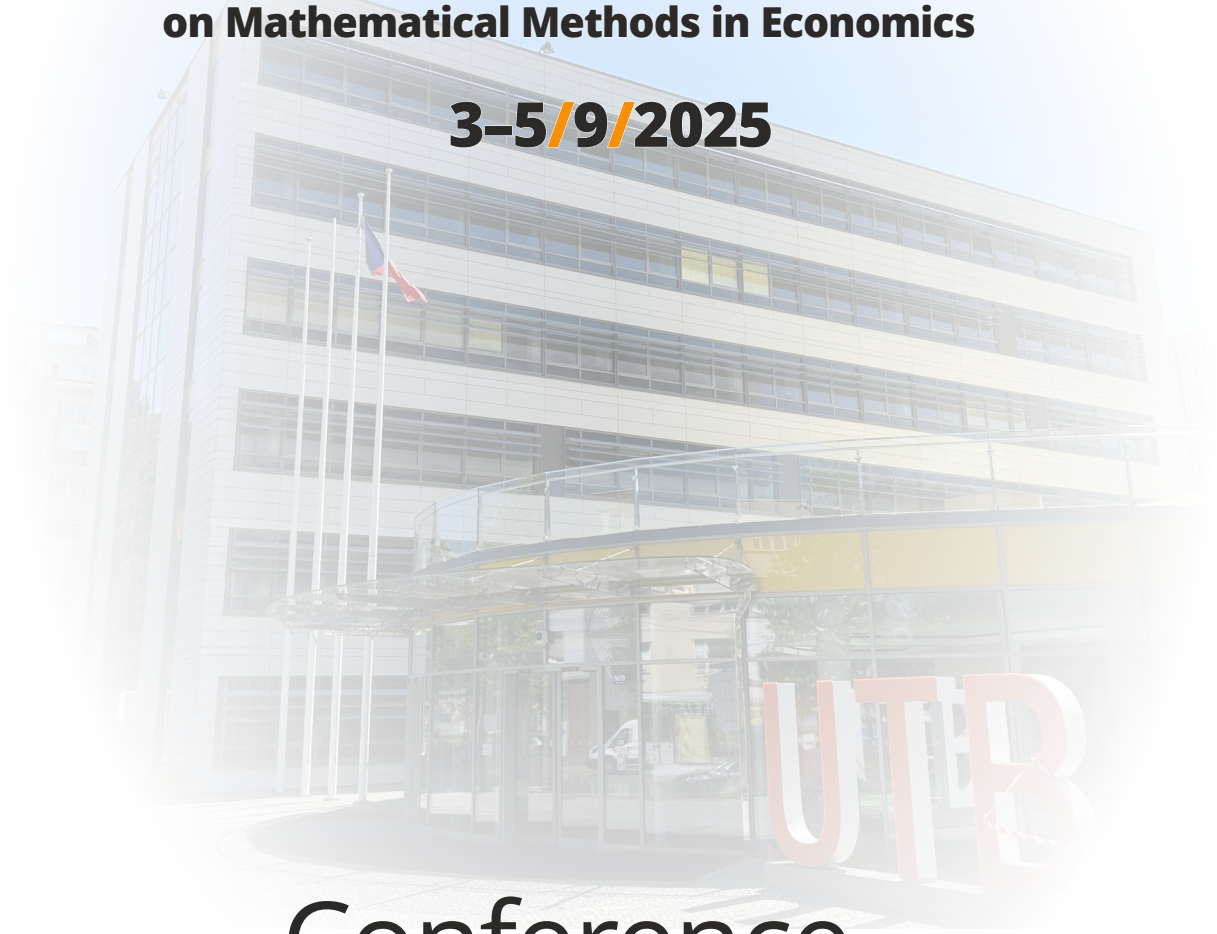
MME 2025

The 43rd International Conference on Mathematical Methods in Economics

3–5/9/2025



Conference Proceedings

Tomas Bata University in Zlín

Tomas Bata University in Zlín
Faculty of Applied Informatics

Tomas Bata University in Zlín
Faculty of Management and Economics

ČSOV

SOLVERTECH

Conference Proceedings

The 43rd International Conference on Mathematical Methods in Economics (MME 2025)

September 3 – 5, 2025

Tomas Bata University in Zlín

Zlín, Czech Republic

43rd International Conference on Mathematical Methods in Economics (MME 2025)
September 3-5, 2025.
Hosted by Tomas Bata University in Zlín, Faculty of Management and Economics, Mostní 5139, 760 01 Zlín
https://mme2025.utb.cz/

Proceedings of the 43rd International Conference on Mathematical Methods in Economics
Publisher: The Czech Society for Operations Research
Winston Churchill Square 1938/4, 130 67 Prague 3, Czechia
http://www.csov.eu/
Editor: Dušan Hrabec
Edition: 1st edition
Publication year: 2025
This publication is not subject to a language check.
All papers have passed a blind peer review process.
All papers adhere to the ethical guidelines of the publisher.
Typeset by LaTeXwith (modified) Book of Abstracts template © José Pedro Costa.
ISBN: 978-80-11-07486-9
ISSN: 2788-3965
© Authors of papers

Contents

Conference Organisation **6**

Welcome **8**

Ethical Guidelines **9**

Contributions **12**

Diana Bílková: *Visegrad Four from Crisis to Crisis* 13

Juraj Pekár, Ivan Brezina and Marian Reiff: *Portfolio Selection in the Return and Absolute Deviation Space Incorporating ESG Indicators* 20

Petr Krautwurm: *Revisiting the Almost Ideal Demand System with Respect to Substitutability and Complementarity: Suitability for Simulating Consumer Behavior* 25

Eva Štichhauerová and Miroslav Žižka: *Efficiency of Regional Transport Systems* 31

František Koblasa and Miroslav Vavroušek: *Assignment Mutation in Evolution Algorithm to Improve Solutions in Flexible Job Shop Scheduling* 37

Robert Hlavatý and Helena Brožová: *On the Efficient Design of a Network under Multiple Criteria* 43

Erik Kočandrle: *Mean-CVaR Portfolio Choice Problem with Endogenous Randomness* 49

Petra Vašaničová and Martina Košíková: *Cluster Patterns and Coopetition Potential Among EU Countries: An Analysis Based on the Travel and Tourism Development Index* 55

Markéta Matulová and Martin Flegl: *A Dynamic Two-Stage DEA Assessment of Energy Transition Progress in the European Union* 61

Vojtěch Vávra: *Chaotic Optimization with Different Chaos Maps for Tuning XGBoost: Application to Gas Demand Forecasting* 67

Josef Kunhart, Jiří Pilný, Simona Huguet and Jan Bartoška: *Lyrical Themes in Contemporary Heavy Metal Albums: A Case Study* 72

Zuzana Neničková, Lucie Waleczek Zotyková and Radmila Krkošková: *Consumer Preferences and Utility in Milk Chocolate Selection: Conjoint Analysis and TOPSIS Method* 78

Aleš Olejníček, Jakub Odehnal and Jiří Neubauer: *Relationship Between Military and Social Expenditures in NATO Countries* 84

Pavla Jindrová, Hana Bohová and Dominik Miškář: *Using Multidimensional Comparative Analysis to Assess the Health Risks of the Elderly Population* 90

Jana Heckenbergerová and Simona Tučková: *Determinants of the Herfindahl-Hirschman Index Affecting Market Power and Concentration in the Czech Insurance Sector* 96

Alena Pozdílková, Martin Pozdílek and Martina Hedvičáková: *Modeling Capital-Labor Substitution in the Era of Industry 4.0: A Sectoral Analysis of Czech Manufacturing* 106

Dominik Kavřík and Jan Rejthar: *Detecting Regime Shifts in the Taylor Curve: A Comparison of Time-Varying Parameter and Markov Switching Models* 113

Jan Bartoška, Josef Kunhart, Jiří Pilný and Simona Huguet: *Extension of the WSJF Method with the Concept of Efficiency for Scrum Team* 119

Qian Gao: *News Sentiment Aware Portfolio Optimization Strategy Based on Crisis Regime Switching* 125

Petr Ondráčková, Jakub Odehnal and Jiří Neubauer: *Economic Effects of Military Expenditures in the Czech Republic* 131

Jaroslav Janáček and Marek Květ: *Heterogeneous Region and Its Servicing by Distributed Queuing System* 137

Marek Květ and Jaroslav Janáček: *Accuracy and Complexity of the Radial Approach to Ambulance Location in Emergency Medical System* 143

Natália Vážošová, Martin Baťka and David Hampel: *Estimation of Price Elasticity of Demand for Energy in the Czech Republic* . . . 149
Martin Baťka: *Partial Least Squares Structural Equation Modeling in Economics: An Empirical Assessment* . . . 155
Jakub Neugebauer: *A Vector-Based Weight Adjustment Algorithm for Portfolio Optimization Boosting Genetic Algorithm Efficiency* . . . 161
Elif Garajová and Miroslav Rada: *Convex Enclosures for Interval Linear Programming* . . . 167
Pavel Pražák: *On Discrete Dynamical System of New Keynesian Model* . . . 173
Adam Borovička: *Portfolio Selecting Under Vague Preferences* . . . 179
Jindřich Lacko and Michal Černý: *Minimizing Elevation Difference in Route Construction: A Geospatial and Computational Analysis* . . . 186
Martin Flegl, Kristýna Vltavská and Andres Acero: *Towards Evaluation of the Czech Primary Education and Its Effect on Civic Engagement and Governance* . . . 192
Martina Kunčová and Roman Fiala: *Efficiency Dynamics in the Czech Brewing Industry: A Post-Pandemic Comparison of Small, Medium, and Large Breweries* . . . 198
Petr Brzobohatý and Aliaksei Sidoryk: *Automated Trading Systems as a Modern Tool in Citizen Investing* . . . 206
Jaromír Zahrádka: *The Solving of Optimal Routes for Multi-Warehouses Distribution by Returning Vehicles* . . . 212
Michaela Chocholatá: *Modeling CO2 Emissions and Economic Growth: A Spatial Panel Approach to the N-Shaped EKC* . . . 218
Michal Konopa and Jana Klicnářová: *When Does Reoptimization Pay Off? A Hierarchical Approach to Vehicle Route Planning* . . . 224
Petr Fiala: *New Negotiation Mechanism in Solving Double Auctions* . . . 232
Igor Kohút: *Dynamic Stochastic General Equilibrium Model with Heterogeneous Expectations for Small Open Economy* . . . 238
Jialei Xiong, Diana Barro and Martina Nardon: *Integrating Market Attention into Portfolio Optimization* . . . 245
Michaela Sedláková: *Coordination of Expectations: A Case Study in Multi-Asset Market* . . . 251
Anshu Anshu and Anshu Anshu: *The Impact of Social Media on Consumer Behaviour and Its Effect on GDP Growth in India* . . . 258
Mohamed Atwy: *The Impact of Healthcare Expenditures on Pharmaceutical Sales and GDP in Czech Republic* . . . 265
Minh Vinh Vo and Phuc Anh Le: *Carbon and Care: Drivers of Sustainable Social Entrepreneurship in the Digital Age* . . . 270
Phuc Anh Le and Minh Vinh Vo: *Strengthening Employee Retention in SMEs: Insights from Motivation, Satisfaction, and Healthcare System Support* . . . 278
Jan Saro: *Predictive Modelling of Dairy Herd Health Using Markov Chains and Neural Networks* 285
Roman Smirnov: *Deriving Production Functions in Economics Through Data-Driven Dynamical Systems* . . . 291
Anton Hovana and Peter Trebuňa: *A Note on Characterization of Comonotonicity* . . . 297
Bahate Maidiya: *The Industry-Level Heterogeneous Impacts of Investor Sentiment on Return, Volume, and Volatility: Evidence from the S&P500* . . . 303

Conference Organisation

Organizing Committee:

- Dušan Hrabec, Tomas Bata University in Zlín, Czech Republic
- David Tuček, Tomas Bata University in Zlín, Czech Republic
- Boris Popesko, Tomas Bata University in Zlín, Czech Republic
- Zbyněk Cerman, Tomas Bata University in Zlín, Czech Republic
- Zuzana Vaculčíková, Tomas Bata University in Zlín, Czech Republic
- Alex Ntsiful, Tomas Bata University in Zlín, Czech Republic
- Lubor Homolka, Tomas Bata University in Zlín, Czech Republic

Program Committee:

- Dušan Hrabec (chair), Tomas Bata University in Zlín, Czech Republic
- Helena Brožová (chair), President of the Czech Society for Operations Research, Czech University of Life Science Prague, Czech Republic
- Miloš Kopa (chair), Vice-president of the Czech Society for Operations Research, Charles University, Czech Republic
- Michal Černý (chair), Prague University of Economics and Business, Czech Republic
- Adam Borovička, Prague University of Economics and Business, Czech Republic
- Martin Branda, Charles University, Czech Republic
- Tomasz Brzeczek, Poznań University of Technology, Poland
- Petr Fiala, Prague University of Economics and Business, Czech Republic
- Martin Flégl, School of Engineering and Sciences, Tecnológico de Monterrey, Mexico
- Richard F. Hartl, University of Vienna, Austria
- Milan Hladík, Charles University, Czech Republic
- Arild Hoff, Molde University College, Norway
- Vladimír Holý, Prague University of Economics and Business, Czech Republic
- Lubor Homolka, Tomas Bata University in Zlín, Czech Republic
- Lars M. Hvattum, Molde University College, Norway
- 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, Prague University of Economics and Business, Czech Republic
- Francesca Maggioni, University of Bergamo, Italy
- Daniel P. Mertens, Jacksonville State University, USA
- Vieroslav Molnár, Technical University of Košice, Slovakia

- Jorge L. Oyola, University of Córdoba, Colombia
- Vít Procházka, Charles University, Czech Republic
- Miroslav Rada, Prague University of Economics and Business, Czech Republic
- Yury Redutskiy, Molde University College, Norway
- Marian Reiff, Bratislava University of Economics and Business, Slovakia
- Jana Sekničková, Prague University of Economics and Business, Czech Republic
- Ondřej Sokol, Prague University of Economics and Business, Czech Republic
- Kenneth Sörensen, University of Antwerp, Belgium
- Petra Tomanová, Prague University of Economics and Business, Czech Republic
- Miroslav Vavroušek, Technical University of Liberec, Czech Republic
- Gerhard-Wilhelm Weber, Poznan University of Technology, Poland
- Miroslav Žižka, Technical University of Liberec, Czech Republic

Foreword

It is my great pleasure to welcome you to the 43rd International Conference on Mathematical Methods in Economics (MME 2025), hosted by Tomas Bata University in Zlín. This Book of Abstracts brings together the contributions accepted after a peer-review process and presented at the conference. It reflects the breadth of current research in operations research, econometrics, and mathematical modeling in economics, and serves as a valuable record of the scientific exchange taking place here.

In addition to its core focus on operations research, econometrics, and mathematical modelling in economics, the conference also welcomes high-quality applied research drawing on quantitative methods in management, logistics, public policy, healthcare, energy systems, and related areas where mathematical and computational approaches provide analytical insight or support decision-making. This broader scope reflects the diverse applications presented at MME 2025 and highlights the role of quantitative modelling across disciplines represented in this year's proceedings

We are proud to host two world-renowned plenary speakers, Professor Stein W. Wallace and Professor Milan Hladík, whose lectures will provide inspiration and valuable insights into the state of the art of mathematical modeling. They will be complemented by Mr. Tomáš Pajonk, an expert from practice who applies advanced methods closely related to the topics of this conference in real-world logistics and decision-support systems. This combination of academic excellence and practical experience will enrich the discussions and strengthen the bridge between theory and application.

As in previous years, the conference also includes the PhD student competition, providing a platform for young researchers to present their work and engage with the scientific community. Beyond the academic program, we invite you to explore the unique industrial and cultural heritage of Zlín. The social events — including the conference excursion to Baťa's landmarks and the traditional conference dinner at Baťa's Villa – will provide excellent opportunities to strengthen professional ties and enjoy the atmosphere of our host city.

I would like to sincerely thank all authors, reviewers, and members of the organizing and program committees for their invaluable contributions. Special thanks go to the Czech Society for Operations Research for their continued support.

I wish you an inspiring conference, fruitful discussions, and a pleasant stay in Zlín.

On behalf of the Organizing Committee,
Dušan Hrabec
General Chair of MME 2025
Zlín, September 2025

Ethical Guidelines

The Czech Society for Operations Research as a publisher of the Proceedings of the International Conference on Mathematical Methods in Economics is committed to the highest ethical standards. All authors, reviewers, and editors must comply with the following ethical principles. In case of any doubts regarding the Ethical guidelines, do not hesitate to contact the editors of the proceedings.

1) Ethical expectations and responsibilities

Editors' responsibilities

- To acknowledge receipt of submitted manuscripts within two working days and to ensure an efficient, fair, and timely review process.
- To identify manuscripts that are fully within the scope and aim of the conference. To treat all submissions fairly without any favour of prejudice regarding authors' gender, religious or political beliefs, ethnic or geographical origin.
- To only authorise the review and publication content of the highest quality.
- To recuse himself or herself from processing any manuscript if he or she has any conflict of interest with any of the authors or institutions related to the manuscripts.
- To provide advice to the authors during the submission process when necessary.
- Editors and organisers are responsible for ensuring that all accepted submissions fall within the scientific scope of the conference and contribute meaningfully to its thematic areas. They must verify that each paper meets the required academic standards in terms of methodological soundness, relevance, and use of quantitative or analytical approaches consistent with the aims of MME. This responsibility includes evaluating whether the submitted work appropriately matches the conference focus and maintaining the coherence and academic integrity of the proceedings
- To be transparent with regards to the review and publication process with an appropriate care that individuals will not be identified when it is inappropriate to do so.
- To not use any parts or data of the submitted manuscript for his or her own future research as the submitted manuscript is not published yet.
- To respond immediately and take reasonable action when ethical problems occur concerning a submitted or published manuscript.
- The review process for this proceedings volume was conducted through the EasyChair conference management system. Each submission was assigned to an editor responsible for overseeing the evaluation and coordinating the review. Manuscripts were reviewed by one or two independent reviewers in a multi-round process, allowing authors to revise and improve their papers based on constructive feedback. The conference employed a blind peer-review procedure to ensure fairness, impartiality, and academic integrity throughout the evaluation of all contributions

Authors' responsibilities:

- To carefully read the Instructions for authors published on the website of the conference.
- To claim that the submitted manuscript is not under consideration or accepted for publication elsewhere. To acknowledge and cite accordingly all content overlaps with already published content.
- To ensure that the submitted manuscript is original, prepared to a high scholarly standard and fully referenced using the prescribed referencing convention.

- To obtain permission to reproduce any content from other published sources. To clarify that all data used in the manuscript has been acquired following ethical research standards.
- To accurately acknowledge funding sources related to the submitted manuscript.
- To carefully read all the conditions included in the copyright form and to accept the copyright form during the submission process.
- To declare any potential conflicts of interest (real or apparent) at any stage during the publication process.
- To recognise that the editor has the final decision to publish the submitted, reviewed and accepted manuscript.
- To immediately inform the editors whenever any obvious error in his or her published manuscript is identified. To cooperate with the editors in the retraction or correction of the manuscript if necessary.

Reviewers' responsibilities:

- To immediately notify the editors whenever the reviewer feels unqualified to review the assigned manuscript or sees difficulties in meeting the deadline for completion of the review.
- To agree to review a reasonable number of manuscripts at the same time.
- To inform the editors if there is any possible conflict of interest related to the assigned manuscript. Specifically, to avoid reviewing any manuscript authored or co-authored by a person with whom the reviewer has an obvious personal or academic relationship.
- To treat the manuscript in a confidential manner and to not use any parts or data of the reviewed manuscript for his or her own future research as the reviewing manuscript is not published yet.
- To assist in improving the quality of the manuscript by reading the manuscript with appropriate care and attention, reviewing the manuscript objectively and being constructively critical.
- To immediately notify the editors of any similarities between the reviewing manuscript and another article either published or under consideration by another journal.

2) Procedures for dealing with unethical behaviour

Dealing with possible misconduct:

- Editors have a duty to act if they suspect any misconduct or if a claim of misconduct has been reported by anyone. This duty applies to both published and unpublished articles.
- Editors should not simply reject an article that raises concerns about possible misconduct. Editors are ethically obliged to pursue alleged cases.
- Whoever informs the editors of such conduct should provide sufficient information and evidence to initiate an investigation.
- Editors must take all allegations seriously and treat them similarly until a successful decision or conclusion is reached.
- Editors should first seek a response from those suspected of misconduct. If the editors are not satisfied with the response, they should ask the relevant employer, institution, or some appropriate body to investigate.

- Editors should make all reasonable efforts to ensure that a proper investigation into alleged misconduct is conducted. If this does not happen, editors should make all reasonable attempts to persist in obtaining a resolution to the problem.

Retraction of an article if:

- Editors have clear evidence that the findings are unreliable, either as a result of misconduct or honest error.
- The findings have previously been published elsewhere without proper cross referencing, permission or justification (i.e. redundant publication). The journal that first published the article may issue a notice of redundant publication, but should not retract the article unless the findings are unreliable. Any journals that subsequently publish a redundant article should retract it and state the reason for the retraction.
- The article constitutes plagiarism and reports unethical research.
- The article should be retracted as soon as possible after the editors are convinced that the article is seriously flawed and misleading.

Outcomes of unethical behaviour in increasing order of severity:

- Informing or educating the author or reviewer where appears to be a misunderstanding or misapplication of acceptable standards.
- A more strongly worded letter to the author or reviewer covering the misconduct and as a warning to future behaviour.
- Retraction or withdrawal of a publication from the proceedings, in conjunction with informing the head of the author or reviewer's department, as well as Abstracting & Indexing services.
- Imposition of a formal embargo on contributions from an individual in the proceedings for a defined period.

Contributions

Received: 2025-03-18 Last revision: 2025-07-04 Accepted: 2025-07-23

Visegrad Four from Crisis to Crisis

Diana Bílková[1]

Abstract. The purpose of this paper is to quantify the impact of the sequence of crises, which began with the COVID-19 pandemic crisis (in Europe in early 2020) followed by the energy crisis and then the crisis related to the war in Ukraine, on the development of the gross domestic product of the Visegrad Four countries. An important objective is also to compare this impact with the consequences of the financial crisis that began in the fall of 2008, which transformed into an economic crisis.
Data on annual gross domestic product per capita at current prices (in USD) come from the official website of the Organization for Economic Cooperation and Development and cover the period from the beginning of the 3rd millennium to 2023. Using data on annual inflation for individual countries, these data were converted to constant prices with a base in 2000.
Exponential smoothing was deliberately used in the construction of the models of the obtained time series. Based on the models, predictions of the mentioned time series for the next three years were constructed, simulating the situation that would occur under unchanged conditions if the economies of these countries had not rebounded from the bottom.

Keywords: Brown's linear exponential smoothing, Brown's quadratic exponential smoothing, Holt's linear exponential smoothing, Visegrad Four, gross domestic product, recession, and economic crises

JEL Classification: C22, E01, E31

AMS Classification: 60H30, 62M10

1 Introduction

An economic crisis, by definition, occurs when the economy declines year-on-year for at least a full year. If the economy declines for at least two consecutive quarters, it is a recession. The accompanying phenomenon of the economic slowdown and the transition to a downward cycle are the economic problems of companies and households.

The Visegrád Four (V4) is a regional grouping of four Central European states: the Czech Republic, Hungary, Poland, and Slovakia. It was established in 1991 with the signing of a declaration by the V4 member states on close cooperation on their path to European integration.

The research uses data downloaded from the official website of the Organization for Economic Cooperation and Development (OECD). This is the annual gross domestic product (GDP) per capita in USD at current prices and current PPPs. Data on annual inflation (consumer price index) is also downloaded from the official OECD website. Using this indicator, nominal GDP values were converted to real GDP at constant prices as of 2000. All further calculations in this text are then performed for the real GDP values thus obtained. The subject of the research is the development of GDP per capita in the Visegrad Four countries in the period 2000–2023, including predictions for the years 2024–2026, assuming unchanged conditions. The purpose is to simulate the situation that would have occurred if the economies of these countries had not rebounded. For this reason, exponential smoothing was used for this purpose.

In the early 2020s, the global economy was hit by a succession of crises. It all started with the COVID-19 pandemic crisis, followed by an energy crisis, followed by an inflationary crisis, and this process was further exacerbated by the crisis associated with the start of the war in Ukraine in February 2022 and the associated refugee crisis. The aim of this research is to determine the impact of the crises that have affected the world economy since the beginning of the 20s of the third millennium on the development of GDP in the V4 countries. The aim is also to compare the impact of the consequences of these crises with the impact of the financial crisis that began in the autumn of 2008 and which later grew into an economic crisis, also on the GDP of the V4 countries.

[1] Prague University of Economics and Business/Faculty of Informatics and Statistics, Department of Statistics and Probability, Sq. W. Churchill 1938/7, 130 67 Prague 3, Czechia, bilkova@vse.cz.

Similarly, Krkošková (2021) addresses the issue of causality between energy consumption and economic growth in the V4 countries. Schwarcz et al. (2021) evaluate the development of economic and social indicators in V4 countries. Four authors, Rezabek, Marek, Doucek, and Nedomova (2022), conduct a comparison of the development of selected macroeconomic indicators of the regions of the V4 Countries. Matoušková (2022) compares economic cycles of the Slovak economy and the other Visegrad Four Countries. Soreg and Bermudez-Gonzalez (2021) conduct a comparative analysis of the socioeconomic development paths of Bulgaria, Croatia, and Romania and Hungary, examining whether Hungary is converging, diverging, or stagnating relative to the three Visegrad Four economies and the Balkans.

2 Theory and Methodology

2.1 Gross Domestic Product Calculation

GDP can be calculated in three ways. Using the GDP production approach, GDP is calculated as the sum of the gross added value of individual institutional sectors or industries and net taxes on products that are not allocated to sectors and industries. This is also a balancing item of the production account for the national economy as a whole, where production is captured on the resource side and intermediate consumption on the use side. Gross added value is the difference between production and intermediate consumption. Given that production is valued in basic prices and used in purchase prices, the resource side of the national economy is supplemented by taxes less product subsidies. The calculation can be written in the form

$$GDP = Production - Intermediate\ consumption + Taxes\ on\ products - Subsidies\ on\ products.$$

The GDP expenditure approach calculates GDP as the sum of final use of goods and services by resident units (actual final consumption and gross capital formation) and the balance of exports and imports of goods and services. Real final consumption is derived through natural social transfers from final consumption expenditure by households, government, and non-profit institutions serving households. Gross capital formation is broken down into gross fixed capital formation, changes in inventories, and net acquisitions of valuables. The calculation can be written in the form

$$GDP = Final\ consumption\ expenditure + Gross\ capital\ formation + Exports\ of\ goods\ and\ services - Imports\ of\ goods\ and\ services.$$

The GDP income approach calculates GDP as the sum of primary income for the national economy as a whole: compensation of employees, taxes on production and imports less subsidies, and gross operating surplus and mixed income (net operating surplus and mixed income and consumption of fixed capital). The calculation can be written in the form

$$GDP = Compensation\ of\ employees + Taxes\ on\ production\ and\ imports - Subsidies + Net\ operating\ surplus + Net\ mixed\ income + Consumption\ of\ fixed\ capital.$$

2.2 Time Series Model Construction

Exponential is classified as an adaptive method of modelling the trend of a time series, and is used when it is not possible to fit a time series with a single trend function over the entire length of the series. However, it is possible to describe such a series with a trend function with variable parameters. This is local trend modelling, also known as the so-called gradual trend concept. Exponential smoothing is suitable for making short-term trend predictions of a time series. It is a technique that develops the idea of smoothing a time series using moving averages. This method uses all previous values of the time series, with the weight of these observations decreasing towards the past according to the exponential function $w_t = (1 - \alpha) \cdot \alpha^{n-t}$, where t is the time variable, n is the length of the time series, and α is a smoothing constant taking values from the interval (0; 1). The issue of exponential smoothing is explained in detail, for example, in Hyndman and Athanasopoulos (2021) or Hyndman et al.(2008).

Using interpolation criteria, a suitable exponential smoothing model was selected. In this approach, we measure the accuracy of time series smoothing by average residual characteristics that can be generalized to any time series model. These include Mean Error – *ME*, Mean Square Error = variance – *MSE*, Mean Absolute Error – *MAE*, Mean Absolute Percentage Error – *MAPE*, and Mean Percentage Error – *MPE*.

The suitability of the chosen type of exponential smoothing was further verified using the autocorrelation and partial autocorrelation functions and by calculating the Durbin-Watson statistic. The essence of exponential smoothing is explained in detail in Hyndman et al. (2008).

2.3 Autocorrelation and Partial Autocorrelation Functions

The autocorrelation function is a tool for assessing the dependence between individual values of a time series. It is a procedure that estimates the correlation coefficients for individual pairs of equally distant values in a time series. Figures 1 and 2 present illustrative examples of Brown's linear exponential smoothing for the GDP per capita for time series of Czechia and Hungary, including the residual autocorrelation function and the residual partial autocorrelation function, where the individual columns represent the estimated correlation coefficients.

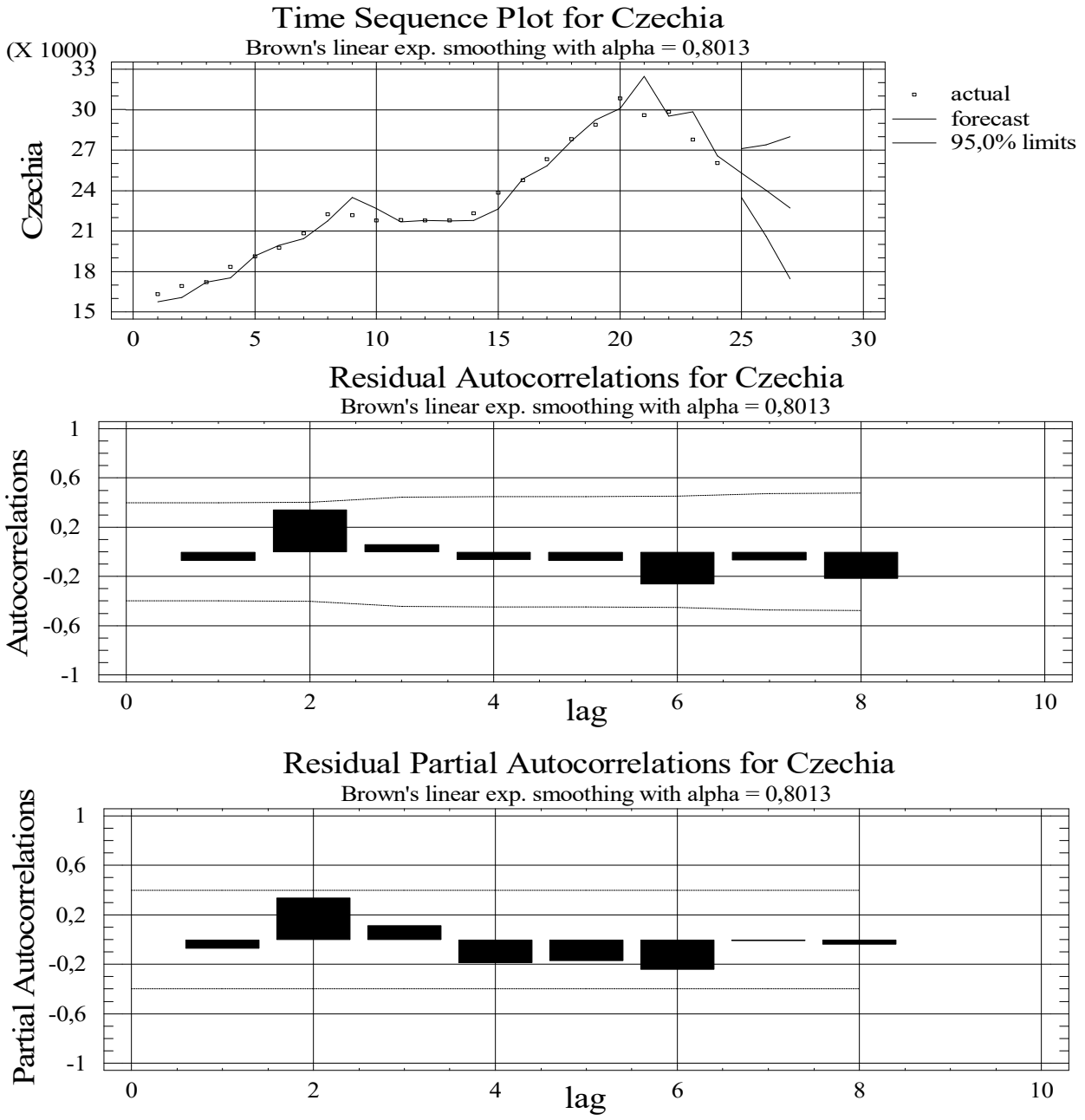


Figure 1 Construction of a Browni's linear exponential smoothing model for the time series of real annual GDP per capita for the Czechia, residual autocorrelation function and residual partial autocorrelation function

In the figures of residual autocorrelation function, a dashed line is visible at a distance of ±2times the standard error of the estimate from zero for each estimated correlation coefficient. This boundary is important for determining the significance of individual correlation coefficients. In general, the correlation coefficients represented by the bars are insignificantly different from zero if the bars do not exceed (above or below) the dashed boundaries. Conversely, bars exceeding these boundaries indicate the significance of the respective correlation coefficients.

The residual partial autocorrelation function, like the residual autocorrelation function, serves as a tool to assess the dependence between individual values of a time series, but it estimates partial correlation coefficients for individual pairs of values in the time series. Figures 1 and 2 graphically depict the course of the residual partial autocorrelation function, too. The individual columns represent the estimated partial correlation coefficients. The

dashed line in the figures again represents the distance ±2 times the standard error of the estimate from the zero for each estimated partial correlation coefficient, which is important for determining the significance of the individual partial correlation coefficients. The partial correlation coefficients represented by the columns are insignificantly different from zero if the columns do not exceed the dashed boundaries. Columns exceeding these boundaries indicate the significance of the respective partial correlation coefficients.

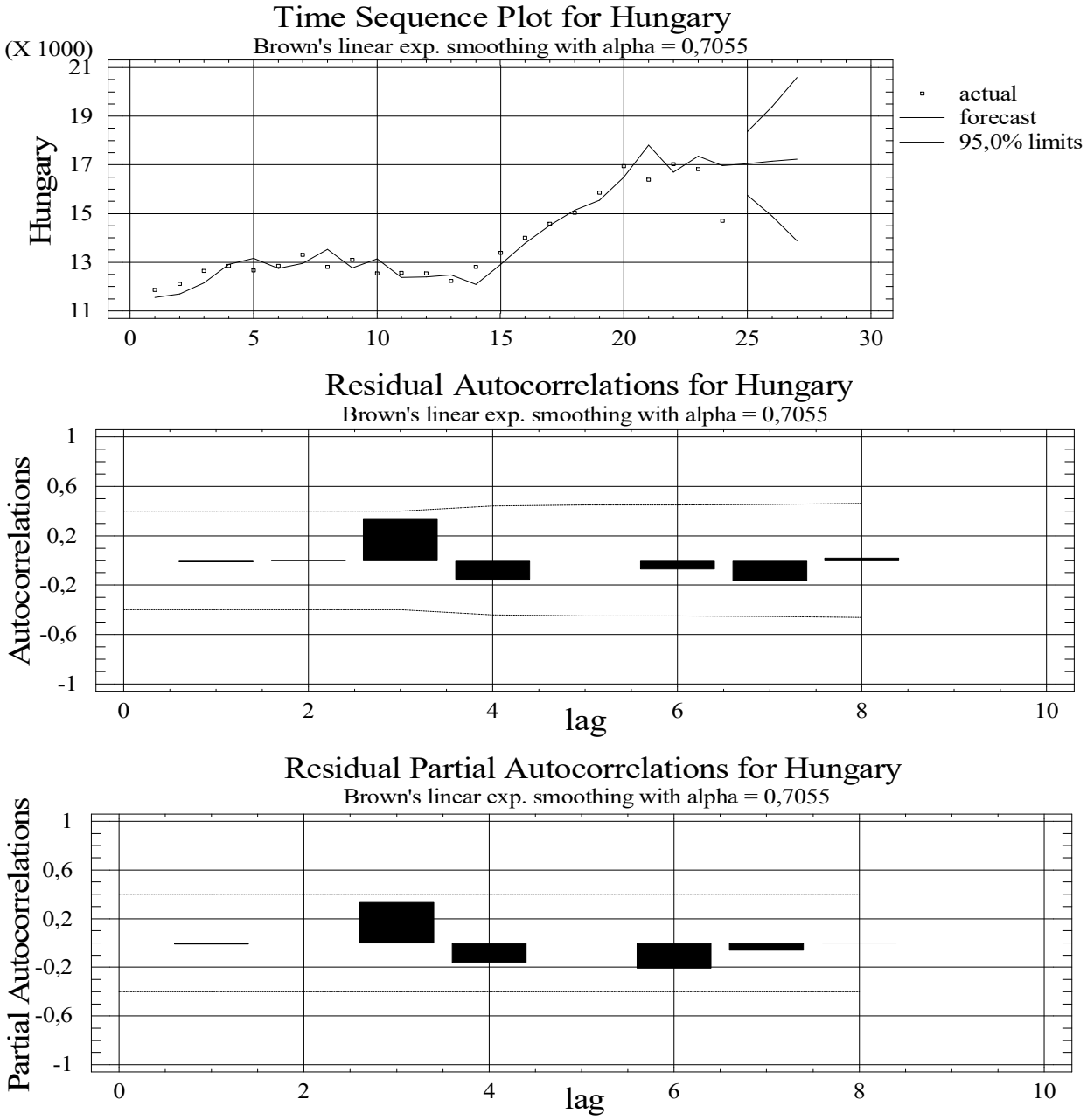


Figure 2 Construction of a Browni's linear exponential smoothing model for the time series of real annual GDP per capita for the Hungary, residual autocorrelation function and residual partial autocorrelation function

Country	Model
Czechia	Brown's linear exp. smoothing with alpha = 0,8013
Hungary	Brown's linear exp. smoothing with alpha = 0,7055
Poland	Brown's quadratic exp. smoothing with alpha = 0,5344
Slovakia	Holt's linear exp. smoothing with alpha = 0,9999 and beta = 0,043

Table 1 Exponential smoothing models of real annual GDP per capita of V4 countries

The results obtained are shown in Figures 1 and 2. It is clear from the figures that none of the estimated correlation coefficients and estimated partial correlation coefficients exceed the dashed line, so we did not demonstrate

statistical significance of any of these coefficients at the 5% significance level. The values of the Durbin-Watson statistics were also in the interval (1.4; 2.6), thus close to the value of two. The selected exponential smoothing therefore appears to be appropriately chosen. The selected types of exponential smoothing of GDP per capita time series of all Visegrad Four countries are presented in Table 1.

3 Results and Conclusion

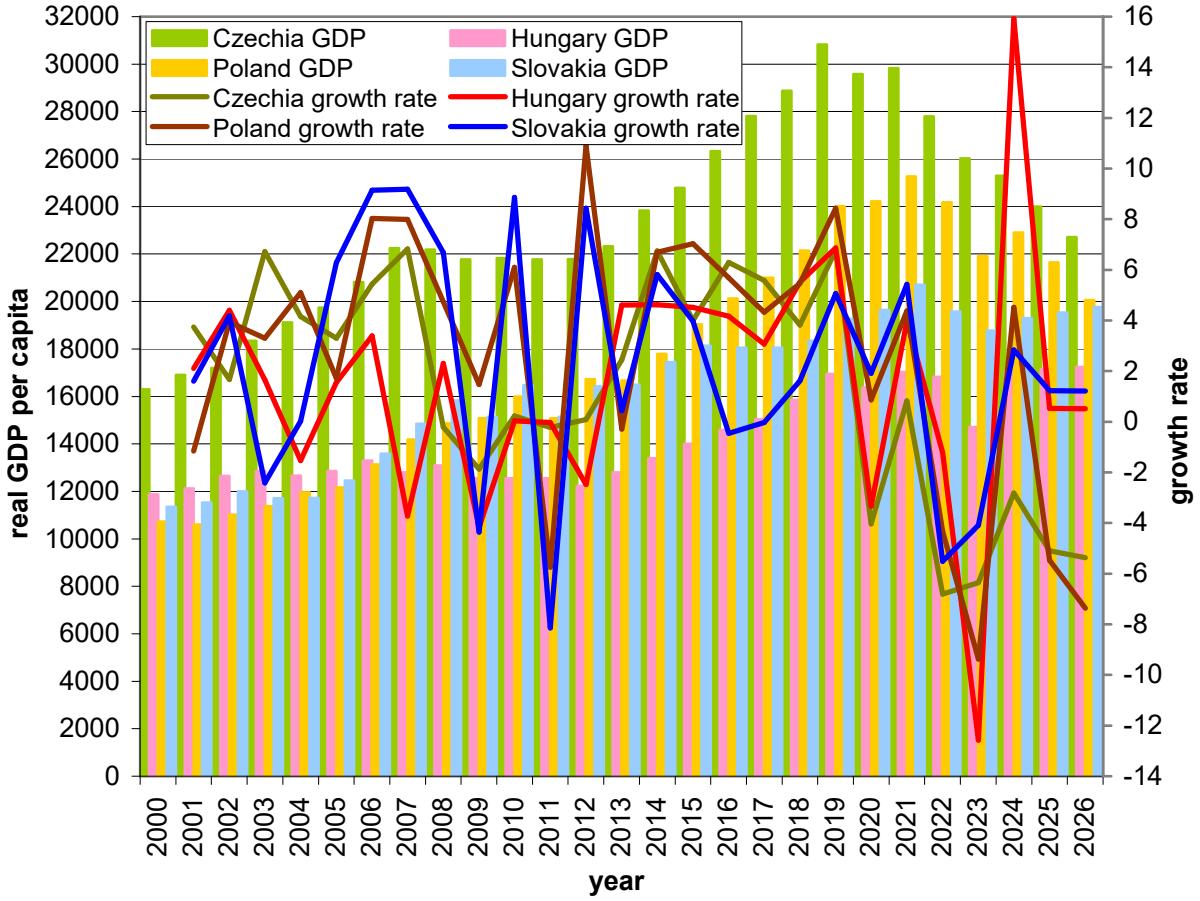


Figure 3 Development of real annual GDP per capita (USD) and growth rate (%) of real annual GDP per capita in 2000–2023 including predisctions for 2024–2026

The constructed time series models show that the V4 countries (Czechia, Hungary, Poland and Slovakia) showed stable growth in real GDP per capita in the period 2000–2019, reflecting their convergence with more advanced European Union economies. However, after 2020, under the influence of the pandemic COVID-19, inflation, energy crisis and geopolitical instability, a turning point occurred. A decline began in 2022, which continues in the forecasts for 2024–2026, see Figure 3.

Czechia

2000–2019: Development was stable, except for the crisis in 2009. GDP per capita increased from 16,301 USD to 30,826 USD in 2019. The peak of growth was recorded in 2015–2019.

2020–2023: Stagnation and subsequent decline since the COVID-19 pandemic. From 30,826 USD in 2019, the value decreased to 26,025 USD in 2023. The research Kotorov, I., Krasylnykova, Y., Demel, J. & Blaschke, P. (2023) investigates the impact of the COVID-19 pandemic on economic growth in the Czech Republic, Germany and Poland, and the research findings show that the COVID-19 pandemic has indeed affected the development of the economies under study. The impact of the pandemic was reflected in a decline in GDP in all three countries in 2020, and in 2021, GDP in all analyzed countries immediately resumed its growth.

2024–2026 (prediction): A continued decline is expected to reach 22,713 USD in 2026. This means a loss of almost 26% compared to the peak in 2019. The Czech Republic will be the V4 most affected by the long-term decline.

Hungary

2000–2019: Growth was slower and more volatile than in the Czech Republic and Poland. After a slight decline in 2007–2009, growth resumed only after 2013.

2020–2023: The pandemic decline was followed by a very sharp decline in 2023 (-12.6%), the largest among the V4. The research Zsibók (2022) uses descriptive and multivariate statistical analysis and finds that significant changes occurred especially in the highest and middle-performing regions and that territorial disparities temporarily narrowed.
2024–2026 (prediction): A sharp return to growth is predicted in 2024 (+15.9%), but the pace slows significantly thereafter. Hungary's GDP per capita remains below the 2019 level, indicating structural problems.

Poland
2000–2019: The most stable and fastest growth among the V4 countries – GDP per capita more than doubled. Poland was the only one of the V4 not to fall into recession during the 2009 crisis.
2020–2023: After weak growth during the pandemic (2020), GDP per capita fell the most since 1989 – from USD 24,004 in 2019 to USD 21,908 in 2023.
2024–2026 (prediction): Predictions show fluctuations with a slight increase in 2024, but an overall downward trend to USD 20,051 in 2026. Nevertheless, Poland remains the best performing V4 economy in terms of growth in the long term.

Slovakia
2000–2019: Rapid growth in GDP per capita, especially in 2004–2008, followed by stagnation after 2010. Slovakia successfully caught up with other V4 members.
2020–2023: Slovakia experiences a decline from 2022, but less dramatic than in the Czech Republic or Hungary. The authors Adamišin et al.(2021) try to draw attention to the problem of the country's economy in connection with government and economic restrictions related to the coronavirus issue.
2024–2026 (prediction): Unlike other countries, predictions show moderate and stable growth. GDP per capita in 2026 is expected to reach $19,773, which is a decrease compared to 2019, but the pace of decline is relatively moderate.

Acknowledgements

This paper was subsidized by the funds of institutional support of a long-term conceptual advancement of science and research number IP400040 at the Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic.

References

Adamišin P, Butoracová Šindleryová I, Čajková A (2021). Coronavirus vs. Real Cause of the European Economic Crisis – Comparing Slovak and German National Model Example. On-line Journal Modelling the New Europe, 37, 78–101.

Hyndman R, Koehler A B, Ord J K, Snyder R D (2008). Forecasting with Exponential Smoothing: The State Space Approach. Springer Science & Business Media.

Kotorov I, Krasylnykova Y, Demel J, Blaschke P (2023). The Effect of the COVID-19 Pandemic on Economic Growth and R&D Spending in Czechia, Germany, and Poland. ACC Journal, 29(2), 37–48.

Krkošková R (2021). Causality between Energy Consumption and Economic Growth in the V4 Countries. Technological and Economic Development of Economy, 27(4), 900–920.

Matoušková E (2022). A szlovák gazdaság és a többi visegrádi négy ország gazdasági ciklusainak összehasonlítása (Comparison of the Economic Cycles of the Slovak Economy and the Other Visegrad Four Countries). Public Finance Quarterly, 67(3),

Rezabek P, Marek L Doucek P, Nedomova L (2022). A Comparison of the Development of Selected Macroeconomic Indicators of the Regions of the V4 Countries. Quality Innovation Prosperity, 26(2), 21–38.

Schwarcz P, Kováčik M, Valach M (2021). The Development of Economic and Social Indicators in V4 Countries. Acta Polytechnica Hungarica, 18(2), 47–68.

Soreg K, Bermudez-Gonzalez G (2021). Measuring the Socioeconomic Development of Selected Balkan Countries and Hungary: A Comparative Analysis for Sustainable Growth. Sustainability, 13(2), 736.

Zsibók Z (2022). Drivers of Regional Economic Growth in Hungary before and during the Pandemic Crisis. Acta Universitatis Sapientiae, European and Regional Studies, 22, 73–89.

Received: 2025-03-17 Last revision: 2025-07-23 Accepted: 2025-07-23

Portfolio Selection in the Return and absolute Deviation Space Incorporating ESG Indicators

Juraj Pekár[1] , Ivan Brezina[2] , Marian Reiff [3]

Abstract: Investment strategy optimization uses portfolio models that diversify assets to minimize risk and maximize return. The paper defines investment risk as the absolute deviation from the expected return. Given the growing importance of sustainable investing, the model is extended with ESG (Environmental, Social, Governance) criteria to ensure consistency with the principles of socially responsible investing. The application of the proposed approach is demonstrated on Dow Jones Industrial Average (DJIA) stocks, providing a practical insight into integrating ESG factors into investment decision-making.

Keywords: portfolio, ESG, optimization

JEL Classification: C61, G11
AMS Classification: 90C05, 91G*

1 Introduction

The creation of a financial portfolio is based on exploring a combination of asset acquisitions so that the portfolio has desirable characteristics, with each investor seeking to create the best performing portfolio that - given his or her risk appetite - represents the most advantageous combination of different return and risk-taking activities (diversification reduces risk). Thus, the asset portfolio represents the diversified assets held by the investor, and its performance is an expression of the portfolio's risk-return profile.
Today, the demand for sustainable investments, in the context of sustainable development, is on the rise. While implementing sustainability requirements may be perceived as increasing risk, socially responsible investing (ESG Investing) is gaining popularity. This approach prioritizes environmental, social, and governance (ESG) factors, reflecting a growing investor interest in supporting socially responsible companies.
Sustainable investing (e.g., Friede et al., 2015; Renneboog et al., 2008) generally refers to the search for investment opportunities that benefit the environment, the company's behavior towards its employees, and its responsibility towards the state. Such investment aims to support companies that positively impact the above indicators, which are represented by environmental, social, and corporate governance (ESG) criteria.
The paper aims not to analyze the assets that make up the calculated efficient portfolios but to confirm or refute the general assumption that sustainability-conscious investing increases risk in the short term (sustainable investing may ignore socially "irresponsible" sectors but can be highly profitable). To test this general hypothesis, the calculations were performed on the assets of the DJIA stock index, but the approach presented can be applied to any set of assets.
The first part of the paper introduces a portfolio selection model that incorporates an investment sustainability condition, based on the absolute deviation risk measure. The following section applies this model to analyze DJIA stock index assets. This practical demonstration shows the increase in risk when the investment sustainability condition is considered, sparking interest in the potential impact on the risk-return ratio.

2 Portfolio selection model enhanced with ESG indicators

Because every rational investor solves the dilemma of maximizing possible returns while minimizing risk, Markowitz (1952) created a mathematical model that became one of the most influential tools in financial investing and remains one of the cornerstones of investment planning today. Markowitz's model seeks the optimal

[1] University of Economics in Bratislava/Department of Operations Research and Econometrics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia, juraj.pekar@euba.sk.
[2] University of Economics in Bratislava/Department of Operations Research and Econometrics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia, ivan.brezina@euba.sk.
[3] University of Economics in Bratislava/Department of Operations Research and Econometrics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia, marian.reiff@euba.sk.

combination of assets in a portfolio, balancing the relationship between return and risk (e.g., Bodie et al., 2014; Fabozzi et al., 2002).

In 1991, Konno and Yamazaki (1991) reported a model that minimizes the hazard rate. Unlike Markowitz (Markowitz, 1952), the variance is replaced by absolute variance. This model is simpler than Markowitz's because it can be transformed to a linear form. Moreover, it eliminates the need to compute the covariance matrix **C**.

Let r_{jt} be the t-th ($t=1,2,...T$, where T is the number of periods used in the construction of the model) component of the discrete random variable X_j represented by the vector $\mathbf{r}_j$ for $j = 1, 2,...n$, whose values we consider to be available, whether they are historical returns or simulated return values. We also assume that the expected value of the random variable X_j can be expressed as a geometric mean derived from these data, assuming an equal probability of occurrence for each state. Then the expected return can be expressed as follows:

$$E_j = E\left(\mathbf{r}_j\right) = \left(\prod_{t=1}^{T}\left(1+r_{jt}\right)\right)^{\frac{1}{T}} - 1 \tag{1}$$

Based on this, the portfolio's expected return function can be written as follows:

$$E(\mathbf{w}) = \sum_{j=1}^{n} E_j w_j \tag{2}$$

Of course, the goal of any investor is to make the value as great as possible. In the above model, the absolute deviation, which can be expressed as a risk function $\delta(\mathbf{w})$, acts as the risk:

$$\delta\left(\mathbf{w}\right) = E\left[\left|\sum_{j=1}^{n}\mathbf{r}_j w_j - E\left[\sum_{j=1}^{n}\mathbf{r}_j w_j\right]\right|\right] = \frac{1}{T}\sum_{t=1}^{T}\left|\sum_{j=1}^{n}\left(r_{jt} - E_j\right)w_j\right| \tag{3}$$

Consequently, based on the above risk measure, the portfolio selection problem can be formulated as follows:

$$\begin{aligned}
&\min \delta\left(\mathbf{w}\right) = \frac{1}{T}\sum_{t=1}^{T}\left|\sum_{j=1}^{n}\left(r_{jt} - E_j\right)w_j\right| \\
&\sum_{j=1}^{n} E_j w_j \geq E_P \\
&\sum_{j=1}^{n} w_j = 1 \\
&w_j \geq 0, j = 1,2,...n
\end{aligned} \tag{4}$$

The absolute value problem can be transformed into a linear problem (Pekár and Furková, 2014) as follows:

$$\begin{aligned}
&\min \delta\left(\mathbf{w},\mathbf{y}\right) = \frac{1}{T}\sum_{t=1}^{T} y_t \\
&y_t + \sum_{j=1}^{n}\left(r_{jt} - E_j\right)w_j \geq 0, \quad t = 1,2,...,T \\
&y_t - \sum_{j=1}^{n}\left(r_{jt} - E_j\right)w_j \geq 0, \quad t = 1,2,...,T \\
&\sum_{j=1}^{n} E_j w_j \geq E_P \\
&\sum_{j=1}^{n} w_j = 1 \\
&w_j \geq 0, j = 1,2,...n
\end{aligned} \tag{5}$$

where the variable y_t represents the absolute deviation in each period t ($t=1,2,...T$, where T is the total number of input periods) from the expected value of the return.

For sustainable investing, the values of ESG indicators need to be included in model (5), and the ESG indicator has been added to the model as part of the analysis. The concept of ESG first appeared in 2004 (Corporate Finance Institute). Still, it was not until 2020 that the concept of ESG evolved into a comprehensive framework that incorporates key elements in relation to a company's environmental, social and governance impacts. ESG has become an essential part of financial investors' decision-making. The ESG concept is based on three core components - environmental (climate change, greenhouse emissions, natural resource depletion, waste and pollution, energy efficiency, deforestation, biodiversity, etc.) and environmental (climate change, greenhouse gas emissions, natural resource depletion, waste and pollution, energy efficiency, deforestation, biodiversity, etc.).); social (human capital management, improving working conditions, employment opportunities, human rights,

employee diversity, health and safety, child labor and slavery, community engagement, philanthropy, etc.), governance (business ethics, fair wages, diversity and board structure, bribery and corruption, political lobbying, tax strategy, compliance, etc.)
The parameter used to build the model was: the desired ESG value (E_{ESG}). The ESG values (ESG_j, the indicator for the j-th asset, j= 1,2,...n) were used as input parameters. The model can be extended with the following conditions:

$$\sum_{j=1}^{n} ESG_j w_j \leq E_{esg} \tag{6}$$

3 Analysis of the implementation of sustainability conditions

For an exemplary application of the presented model, it is important to correctly determine the time period for analyzing historical stock prices. This time span is crucial to obtain relevant data on calculating the average return. In the presented study, the evolution of the weekly prices of 29 stocks of the DJIA index over the specified period, namely from January 1, 2020 to December 31, 2024, was applied
In addition to stock price performance, ESG metrics for selected DJIA index stocks should be obtained. The relevant values are available from the rating agency Sustainalytics at finance.yahoo.com (Sustainalytics, 2024). The agency typically publishes ESG ratings annually, and 2023 data was used in the analysis.
In the solution, the maximum weighted ESG value for the portfolio was set as the median value, which takes the value of 20 for the above stocks.
Portfolio optimization accepting the investment sustainability condition was performed using Python, and the results (efficient portfolios) are shown in Table 1 in the ESG section. Calculations were also performed to compare the impact of the sustainability conditions without considering the ESG condition (Table 1, non-ESG section).

	With ESG		Without ESG	
Portfolio	AD	Ep	AD	Ep
P1	1,460%	0,198%	1,430%	0,208%
P2	1,470%	0,240%	1,440%	0,250%
P3	1,500%	0,280%	1,470%	0,290%
P4	1,540%	0,320%	1,530%	0,330%
P5	1,600%	0,360%	1,590%	0,360%
P6	1,680%	0,400%	1,680%	0,400%
P7	1,810%	0,440%	1,810%	0,440%
P8	2,040%	0,480%	2,020%	0,480%
P9	2,370%	0,520%	2,330%	0,520%
P10	2,960%	0,560%	2,960%	0,560%

Table 1 Efficient portfolios with and without ESG

Source: Author's own calculations

The number of computed efficient portfolios for the two cases considered is 10, with the values in column AD being the computed minimum values of the portfolio's absolute deviation at the specified rates of return given in column Ep (sorted in ascending order). The values analyzed clearly show that the differences in the portfolios' absolute deviations are only more pronounced at lower portfolio return rates Ep.

Figure 1 shows a graphical interpretation of the results. It also shows that the Ep to AD ratios for the ESG and non-ESG considered for efficient portfolios take on different values only for lower values of the portfolio return Ep. This confirms the general assumption that investing in sustainable assets (with ESG) carries a higher risk for adequate considered returns but that this difference disappears at higher returns.

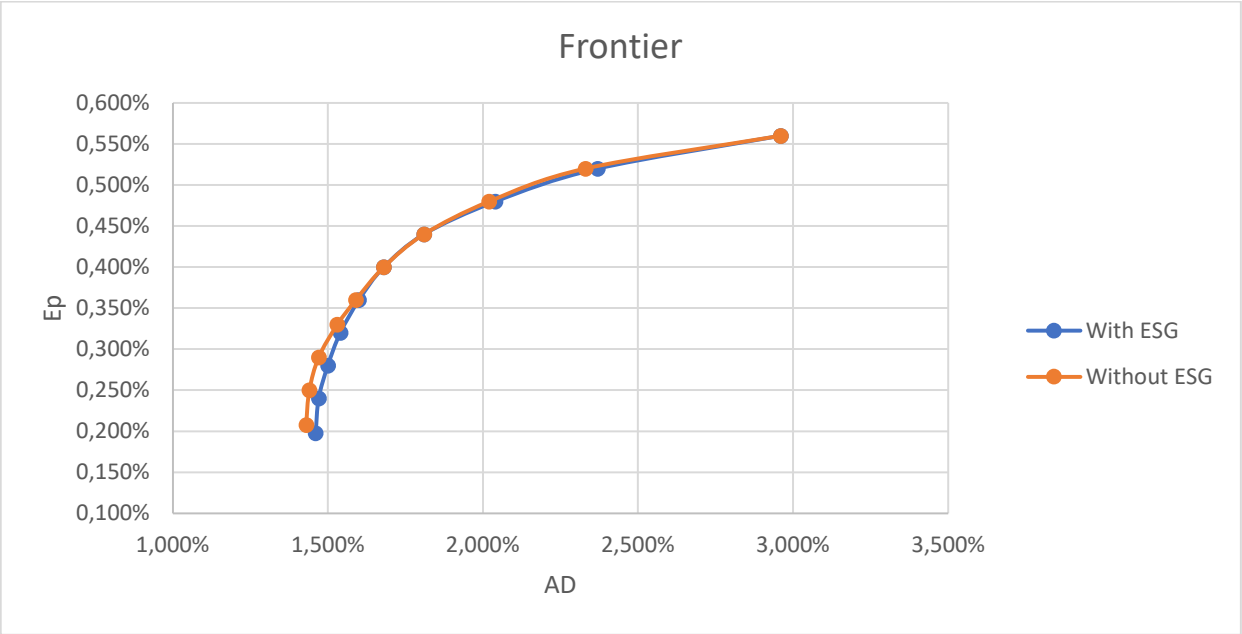


Figure 1 Effective boundary in case of ESG boundary and without ESG

Source: Author's own calculations

4 Conclusion

The paper analyzes the impact of considering the sustainability condition in the portfolio selection model. Sustainability has been implemented by adding a structural constraint to the model that limits the maximum value of the portfolio's weighted average ESG score to the level of the median frontier. When applying the model to DJIA index stocks, the results of portfolio optimization with and without the ESG constraint were compared.
The results showed that the inclusion of ESG criteria leads to a slight increase in risk (absolute deviation) compared to the traditional model without ESG constraints. This increase in risk is due to the restriction of the selection of investment assets to those that meet the required sustainable standards. However, despite this increased risk, ESG portfolios are becoming increasingly popular due to the growing demand for socially responsible investments.
Based on the analysis conducted, it can be concluded that while sustainable investing may reduce diversification and slightly increase investment risk, it can bring long-term benefits in the form of lower volatility, higher reputation of firms in the portfolio and potentially better performance in times of economic uncertainty. Therefore, investors must consider risk and return and the broader impact of their investments on society and the environment.

Acknowledgements

This work was supported by the Grant Agency of Slovak Republic - VEGA grant no. 1/0120/23 "Environmental models as a tool for ecological and economic decision making ".

References

Bodie Z, Kane A, Marcus A. J (2014*)*.Investments (10th ed.). McGraw-Hill Education.

Fabozzi F J, Gupta F, Markowitz H M (2002) The Legacy of Modern Portfolio Theory. The Journal of Investing, 11(3), 7–22.

Friede G, Busch T, Bassen A (2015) ESG and Financial Performance: Aggregated Evidence from More than *2000* Empirical Studies. Journal of Sustainable Finance & Investment, 5(4), 210–233.

Konno H, Yamazaki H (1991) Mean-Absolute Deviation Portfolio Optimization Model and Its Applications to Tokyo Stock Market. Management Science, 37(5), 519-531.

Markowitz H (1952) Portfolio Selection. The Journal of Finance, 7(1), 77-91.

Renneboog, L., Horst, J. T., & Zhang, C. (2008). *Socially Responsible Investments: Institutional Aspects, Performance, and Investor Behavior*. Journal of Banking & Finance, 32(9), 1723-1742.

Sustainalytics. (2024). *ESG Risk Ratings Methodology - Version 3.1*. Retrieved from https://www.sustainalytics.com

Received: 2025-06-20 Last revision: 2025-07-06 Accepted: 2025-07-14

Revisiting the Almost Ideal Demand System with Respect to Substitutability and Complementarity: Suitability for Simulating Consumer Behavior

Petr Krautwurm[1]

Abstract. This paper investigates the suitability of the Almost Ideal Demand System as a foundation for simulating consumption behavior conditional on known relationships between products. It derives the system's exact Marshallian demand functions and corresponding cross-price elasticities, assessing how these align with established theoretical definitions of complementarity and substitutability; namely the "q", "p", "gross", and "net" perspectives. While the Almost Ideal Demand System satisfies core axioms of consumer theory and offers interpretable parameter structure, its lack of an explicit utility function limits its connection to preference-based definitions. The paper concludes that although the system is analytically elegant, alternative formulations, such as quadratic utility functions with interaction terms, may offer better operational grounding for simulations involving context-sensitive product relationships.

Keywords: Almost Ideal Demand System, Consumer, Cross-price elasticity, Substitutes, Complements

JEL Classification: D01, D11
AMS Classification: 91B16

1 Introduction

Relationships between products are one of the most important aspects of any economic analysis. The central question is to what extent these products can be used together for a common purpose, or conversely, to what extent they are indistinguishable. In essence, the study of product relationships reduces to identifying whether goods are complements or substitutes.

Over the course of economic thought, a substantial body of literature has addressed the nature of substitutes and complements (see Hicks and Allen 1934; Russell et al. 1993; Manchanda 1999; Shocker et al. 2003; Berry et al. 2014; Kort et al. 2018; Ruiz et al. 2020; Weinstein 2022), with many influential frameworks proposed to infer such relationships from empirical data (Shocker et al. 2004; Kim et al. 2012; Netzer et al. 2012; Ruiz et al. 2020). Yet, what remains missing is a unified simulational framework that enables researchers to ex ante define product relationships, simulate the consumer behavior, and then observe the resulting patterns of such behavior.

There are two main reasons why such a framework has not yet been established. First, there exist multiple competing and mutually inconsistent definitions of what constitutes substitutes and complements, often with nontrivial implications. These include the so-called "p" and "q" relationships (Hicks and Allen 1934b; Samuelson 1974; Seidman 1993), as well as "net" and "gross" relationships (Samuelson 1974; Weinstein 2022), along with additional contextual qualifications (Berry et al. 2014; Kort et al. 2018). Second, Uzawa (1962) showed that it is impossible to construct a single function that could model different relationships between distinct product pairs via elasticity of substitution and keep these elasticities constant across all quantity configurations. In practical terms, this means that substitution relationships between products are inherently context-dependent. Together, these issues complicate the development of a generalizable framework for exploring how interactions between distinct products shape emergent purchasing patterns.

This paper evaluates the potential of the Almost Ideal Demand System, introduced by Deaton and Muellbauer (1980), as a benchmark simulational framework for modeling such relationships. As such, cross-price elasticity (Hicks and Allen 1934; Deaton 1987) corresponding to this framework will be the focus of this research.

The purpose of current paper is to open the door for future exploration. The paper's main contribution is to assess whether the AIDS framework can serve as a foundation for modeling known relationships between goods and analyzing the resulting patterns of market behavior.

[1] Prague University of Economics and Business, Department of Econometrics, W. Churchill sq. 1938/4, Prague, 130 67, The Czech Republic, petr.krautwurm@vse.cz

2 Defining Relationships between Products

Broadly, relationships between products hinge on whether goods are used together or are interchangeable. However, such definitions are highly context-dependent (see Berry et al. 2014 for a detailed discussion on how contextual factors influence complementarity). For example, cucumbers and tomatoes may be complements in the context of making a salad, but substitutes when assembling a sandwich. As such, it is implausible to expect a stable relationship between two products that holds across all consumption contexts.

Still, it may be useful to adopt a working definition that captures the average tendency across the average context. One approach is to define product relationships with so-called "q" and "p" perspectives (Hicks and Allen 1934; Samuelson 1974; Seidman 1993). The "q" perspective corresponds to how one product affects the marginal utility of another product, while "p" perspective corresponds to how the demand for product is affected by another product's price. As such, the "q" perspective describes the problem from the perspective of preferences, while "p" does it from the perspective of choice within market setup.

Another, related distinction is between "gross" and "net" relationships (Samuelson 1974; Weinstein 2022). The "gross" perspective captures real-world behavioral responses to price changes without compensating for other factors like income. In contrast, the "net" perspective isolates preference-driven effects by considering what the agent would do if compensated for such changes (particularly, changes in real income). That is, we examine how behavior shifts when the agent is made indifferent (in utility terms) between the original and new situations.

Conceptually, the "p" and "gross" perspectives are closely aligned; both reflect realized market behavior. On the contrary, the "q" and "net" perspectives, attempt to isolate underlying preferences; yet, each in their own manner. The distinction becomes clearer in the formal exposition below.

2.1 Perspective in Detail

As noted, the "q" and "p" perspectives define product relationships through preferences and choices, respectively.

Assuming an agent with rational preferences over n products, we can describe these preferences over quantities of said products $x_1, \dots x_n$ using a utility function $U(x_1, \dots, x_n)$. The marginal utility of product i is given by $MU_i = \partial U / \partial x_i$, which quantifies the contribution of x_i to overall utility. The "q" relationships are then defined as:

$$\text{"q" relationships: } \begin{cases} \frac{\partial MU_i}{\partial x_j} > 0 & \text{[Complements]} \\ \\ \frac{\partial MU_i}{\partial x_j} < 0 & \text{[Substitutes]} \end{cases}$$

According to these definitions, "q" complements enhance the marginal utility of another product, while "q" substitutes diminish it. In intuitive terms, a complement makes another product "bring more to the table", whereas a substitute reduces its marginal contribution; not because it tastes worse, but because the presence of one product reduces the need for the other. This makes the "q" perspective an appealing and conceptually natural foundation.

In contrast, the "p", "gross", and "net" perspectives are based on observable choice behavior in response to price changes. All three perspectives relate to demand functions, but differ in the specific variant considered.

Given a linear budget constraint with linear prices $\boldsymbol{p}$ and budget M, two demands can be defined: Marshallian and Hicksian. They are similar in both of them being a function of prices $\boldsymbol{p}$, but Marshallian demand is also function of budget M, while Hicksian demand is a function of utility needed to be reached u.

Given a linear budget constraint with prices $\boldsymbol{p}$ and total budget M, we define two types of demand:

Marshallian demand: $x_i^* = (\mathbf{x}^*)_i$ is the solution to primary problem of maximizing the utility:

$$\mathbf{x}^*(\boldsymbol{p}, M) \in \arg\max_{\mathbf{x}} \left\{ U(\boldsymbol{x}) \;\middle|\; \boldsymbol{p}\boldsymbol{x} = M \right\} \tag{1}$$

Hicksian demand: $\widetilde{x}_i = (\widetilde{\mathbf{x}})_i$ is the solution to the dual problem of minimizing the expenditures such that the utility reaches a certain level u:

$$\widetilde{\mathbf{x}}(\boldsymbol{p}, u) \in \arg\min_{\mathbf{x}} \left\{ \boldsymbol{p}\boldsymbol{x} \;\middle|\; U(\boldsymbol{x}) = u \right\} \tag{2}$$

The "p" and "gross" relationships are derived from Marshallian demand, such as:

$$\text{“p” and “gross” relationships: } \begin{cases} \frac{\partial x_i^*}{\partial P_j} < 0 & \text{[Gross Complements]} \\ \\ \frac{\partial x_i^*}{\partial P_j} > 0 & \text{[Gross Substitutes]} \end{cases}$$

Meanwhile, the “net” relationships stem from Hicksian demand:

$$\text{“net” relationships: } \begin{cases} \frac{\partial \widetilde{x}_i}{\partial P_j} < 0 & \text{[Net Complements]} \\ \\ \frac{\partial \widetilde{x}_i}{\partial P_j} > 0 & \text{[Net Substitutes]} \end{cases}$$

As such, “net” relationships closely resemble the “q” definition; both aim to isolate preference-driven effects. However, they remain distinct, as the “net” relationships rely on a compensation mechanism that adjusts income to offset the utility effects of price changes.

Thus, while “q” relationships offer a foundational and definitional understanding of complementarity and substitutability, they are difficult to operationalize due to the inaccessibility of the underlying utility function. In contrast, “gross” and “net” relationships, though less conceptually pure, are empirically tractable. Both Marshallian and Hicksian demands can be estimated. In particular, Hicksian demand can be inferred from Marshallian demand using the Slutsky equation (Hicks and Allen 1934).

For these reasons, a desirable simulational framework would ideally relate all three definitions, starting from the “q” perspective and systematically deducing the “gross” and “net” relationships.

3 Almost Ideal Demand System

The Almost Ideal Demand System (AIDS) is a framework that approximates what the expenditure function would look like if it were derived from arbitrary, well-behaved utility function.

The expenditure function $E(\mathbf{p}, u)$ reflects the minimum expenditure required to achieve utility level u at price vector $\mathbf{p}$ and is constructed from the expenditures $\mathbf{px}$ corresponding precisely to Hicksian demand $\mathbf{p}\widetilde{\boldsymbol{x}}(\mathbf{p}, u)$. For the AIDS, this function has the following form:

$$\ln(E(\mathbf{p}, u)) = \alpha_0 + \sum_i \alpha_i \ln(p_i) + \frac{1}{2} \sum_i \sum_j \gamma_{ij} \ln(p_i) \ln(p_j) + \beta_0 \prod_i p_i^{\beta_i} \cdot u \tag{3}$$

Here, all prices $p_i > 0$, and symmetry is assumed: $\gamma_{ij} = \gamma_{ji}$. For tractability, it is useful to decompose the expenditure function into two parts:

$$A(\mathbf{p}) = \alpha_0 + \sum_i \alpha_i \ln(p_i) + \frac{1}{2} \sum_i \sum_j \gamma_{ij} \ln(p_i) \ln(p_j) \tag{4}$$

and

$$B(\mathbf{p}) = \beta_0 \prod_i p_i^{\beta_i} \tag{5}$$

The next step is to derive the explicit demand functions implied by this setup, which allows us to compute cross-price elasticities and assess how the AIDS reflects relationships between products.

3.1 Deriving demands

While the AIDS is often specified in terms of budget shares, it is analytically more useful to derive the exact Marshallian demands. To do this, we must first transform the expenditure function into its corresponding indirect utility function $V(\mathbf{p}, M)$ and then apply Roy’s identity (Varian 1993):

$$x_k^*(\mathbf{p}, M) = -\frac{\frac{\partial V(\mathbf{p},M)}{\partial p_k}}{\frac{\partial V(\mathbf{p},M)}{\partial M}} \tag{6}$$

To apply Roy's identity, we use the fact that $u = V(\mathbf{p}, M)$, which implies $E(\mathbf{p}, u) = E(\mathbf{p}, V(\mathbf{p}, M)) = M$ (Varian 1993). Substituting this into the expenditure function yields:

$$V(\mathbf{p}, M) = \frac{\ln(M) - A(\boldsymbol{p})}{B(\boldsymbol{p})} = \frac{\ln(M) - \alpha_0 - \sum_i \alpha_i \ln(p_i) - \frac{1}{2}\sum_i \sum_j \gamma_{ij} \ln(p_i) \ln(p_j)}{\beta_0 \prod_i p_i^{\beta_i}} \tag{7}$$

Additionally, it is useful to split the formula into two parts: First, define the nominator as $A_0(\mathbf{p}, M) = \ln(M) - A(\mathbf{p})$; then, adjust the denominator as $B_0(\mathbf{p}) = 1/B(\mathbf{p})$, leading to $V(\mathbf{p}, M) = A_0(\mathbf{p}, M) \cdot B_0(\mathbf{p})$.

The differentiation in the denominator of Roy's identity is straightforward, as there is only one term involving M in the original expenditure function:

$$\frac{\partial V(\mathbf{p}, M)}{\partial M} = \frac{1}{M}\frac{1}{B(\mathbf{p})} \tag{8}$$

The more challenging part is the differentiation in the numerator. Applying the chain rule gives:

$$\frac{\partial V(\mathbf{p}, M)}{\partial p_k} = \frac{\partial A_0(\mathbf{p}, M)}{\partial p_k} \cdot B_0(\mathbf{p}) + A_0(\mathbf{p}, M)\frac{\partial B_0(\mathbf{p})}{\partial p_k} \tag{9}$$

It is trivial to verify that $\partial A_0/\partial p_k = -\partial A/\partial p_k$, which simplifies the expression to:

$$\frac{\partial V(\mathbf{p}, M)}{\partial p_k} = -\frac{\partial A(\mathbf{p})}{\partial p_k} \cdot B_0(\mathbf{p}) + \Big(\ln(M) - A(\mathbf{p})\Big)\frac{\partial B_0(\mathbf{p})}{\partial p_k} \tag{10}$$

We then compute the derivatives of the key components:

$$\frac{\partial A(\mathbf{p})}{\partial p_k} = \alpha_k \frac{1}{p_k} + \frac{1}{2}\left(2 \cdot \gamma_{kk}\frac{1}{p_k}\ln(p_k) + 2\sum_{i \neq k}\gamma_{lk}\frac{1}{p_k}\ln(p_i)\right) = \frac{1}{p_k}\left(\alpha_k + \sum_i \gamma_{ik}\ln(p_i)\right) \tag{11}$$

and

$$\frac{\partial B_0(\mathbf{p})}{\partial p_k} = -\frac{\beta_k}{p_k}\frac{1}{\beta_0 \prod_i p_i^{\beta_i}} = -\frac{\beta_k}{p_k}\frac{1}{B(\mathbf{p})} \tag{12}$$

Substituting these into the expression above yields:

$$\frac{\partial V(\mathbf{p}, M)}{\partial p_k} = -\frac{1}{p_k}\left(\alpha_k + \sum_i \gamma_{ik}\ln(p_i)\right)\frac{1}{B(\mathbf{p})} - \frac{1}{B(\mathbf{p})}\Big(\ln(M) - A(\mathbf{p})\Big)\left(\frac{\beta_k}{p_k}\right) \tag{13}$$

Multiplying by $B(\boldsymbol{p})$, (-1), and M gives us the final form of the Marshallian demand derived via Roy's identity:

$$x_k^*(\mathbf{p}, M) = -\frac{\frac{\partial V(\mathbf{p},M)}{\partial p_k}}{\frac{\partial V(\mathbf{p},M)}{\partial M}} = \frac{M}{p_k}\left(\alpha_k + \sum_i \gamma_{ik}\ln(p_i) + \beta_k\Big(\ln(M) - A(\mathbf{p})\Big)\right) \tag{14}$$

This expression (though relatively rare in the literature) provides a complete characterization of consumption behavior across n products, for a consumer with a utility function consistent with the original expenditure function.

The interpretation is intuitive and relatively nice. The term M/p_k represents the maximum possible quantity of good k that could be purchased given budget M. Thus, the remaining factor is a scaling term between 0 and 1 that

adjusts this maximum based on other parameters. For economic goods, this term is strictly positive, and summing over all n goods yields 1. This leads to parameter restrictions: $\sum_i \alpha_i = 1$, $\sum_i \gamma_{ij} = \sum_j \gamma_{ji} = 0$, $\sum_i \beta_i = 0$.

These conditions provide additional interpretive value: $\alpha_k \in (0, 1)$ can be seen as an unconditional taste for good k, $\gamma_{kj} \in (-\infty, \infty)$ reflects interaction with other goods, and $\beta_k \in (-\infty, \infty)$ captures sensitivity to real income changes via price effects. The framework thus offers a reasonably interpretable model of consumption behavior.

The key question is how these parameters relate to the concepts of complementarity and substitutability introduced in Section 2. This can be addressed, at least partially, by analyzing cross-price elasticities or by retrieving the utility function corresponding to this system.

3.2 Cross-price Elasticity

To assess how the AIDS aligns with the "gross" and "p" perspectives, we compute cross-price elasticities of demand, defined as:

$$\zeta_{ij} = \frac{\frac{\partial x_i}{x_i}}{\frac{\partial p_j}{p_j}} = \frac{\partial \ln(x_i)}{\partial \ln(p_j)} \tag{15}$$

These elasticities serve the same conceptual role as the "gross" and "p" measures, indicating whether products are complements or substitutes, but being scaled by current prices and quantities for logarithmic perspective. Nevertheless, like their counterparts, they are centered around zero, facilitating the interpretation.

To compute them, we require the derivative of x_k^* with respect to price p_j of another good:

$$\frac{\partial x_k^*(\mathbf{p}, M)}{\partial p_j} = \frac{M}{p_k}\left(\gamma_{jk}\frac{1}{p_j} - \beta_k \frac{\partial A(\mathbf{p})}{\partial p_j}\right) \tag{16}$$

Using equation (11), we find:

$$\frac{\partial x_k^*(\mathbf{p}, M)}{\partial p_j} = \frac{M}{p_k}\left(\gamma_{jk}\frac{1}{p_j} - \beta_k \frac{1}{p_j}\left(\alpha_j + \sum_i \gamma_{ij}\ln(p_i)\right)\right) \tag{17}$$

Multiplying this by $p_j / x_k^*(\mathbf{p}, M)$, where the denominator is the exact demand from equation (14), and simplifying gives the exact formula for cross-price elasticity:

$$\zeta_{ij} = \frac{\frac{\partial x_i}{x_i}}{\frac{\partial p_j}{p_j}} = \frac{\gamma_{jk} - \beta_k\left(\alpha_j + \sum_i \gamma_{ij}\ln(p_i)\right)}{\alpha_k + \sum_i \gamma_{ik}\ln(p_i) + \beta_k\left(\ln(M) - A(\mathbf{p})\right)} \tag{18}$$

As shown earlier, the denominator lies strictly between 0 and 1, so the sign of the elasticity depends entirely on the numerator. As such, the interpretation of the parameters is more clear. The term γ_{jk} reflects the direct price effect on demand, modulated by the broader decision context via β_k. If $\beta_k = 0$, then $\gamma_{jk} > 0$ implies substitutability, and $\gamma_{jk} < 0$ implies complementarity. As such, if all indirect effects were neutralized (i.e., compensated) then γ_{kj} would closely approximate the "net" relationship between goods. A more definitive answer, however, would require deriving the exact Hicksian demand functions and assessing their structure, echoing the earlier discussion of context dependence in Section 2.

The main limitation of this framework is structural. The AIDS allows only pairwise interactions via γ_{kj} or generalized market effects via β_k. Yet, in reality, relationships among goods often occur in sets: for instance, two sets of products may be substitutable together but not individually. This type of higher-order complementarity or substitutability cannot be captured by the model.

4 Conclusion

This paper examined the AIDS by deriving its exact Marshallian demand functions and the corresponding formula for cross-price elasticity. It aimed to evaluate how the system aligns with different theoretical perspectives on

product relationships (particularly the “q”, “p”, “gross”, and “net” definitions) in order to assess its potential as a benchmark framework for simulating consumption behavior conditional on inter-product relationships.

The AIDS offers a compelling combination of analytical tractability and theoretical richness. As its name implies, it comes remarkably close to satisfying the ideal requirements of consumer demand theory. Nevertheless, despite fulfilling the standard axioms of microeconomic theory, its practical applicability to the present research question is limited by a crucial issue: the absence of a closed-form utility function consistent with its expenditure function. This limitation constrains its alignment with the “q” perspective and complicates interpretability in preference-based terms.

In light of this, it may be worthwhile to explore alternative functional forms, such as quadratic utility functions with interaction terms, that relax some traditional assumptions in favor of greater operational simplicity and interpretability.

Acknowledgements

This work was supported by The Internal Grant Agency of Prague University of Economics and Business [VŠE IGS F4/39/2025].

References

Berry S, Khwaja A, Kumar V, Musalem A, Wilbur K, Allenby G, Anand B, Chintagunta P, Hanemann W, Jeziorski P, Mele A (2014) Structural models of complementary choices. Marketing Letters 25(3):245–256. https://doi.org/10.1007/s11002-014-9309-y

Deaton A (1987) Estimation of own- and cross-price elasticities from household survey data. Journal of Econometrics 36(1–2):7–30.

Hicks J, Allen R (1934) A reconsideration of the theory of value. Part II. A mathematical theory of individual demand functions. Economica 1(2):196. https://doi.org/10.2307/2548749

Kim H, Kim J, Chen Q (2012) A product network analysis for extending the market basket analysis. Expert Systems with Applications 39(8):7403–7410. https://doi.org/10.1016/j.eswa.2012.01.066

Kort P, Taboubi S, Zaccour G (2018) Pricing decisions in marketing channels in the presence of optional contingent products. Central European Journal of Operations Research 28(1):167–192. https://doi.org/10.1007/s10100-018-0527-x

Manchanda P, Ansari A, Gupta S (1999) The “shopping basket”: A model for multicategory purchase incidence decisions. Marketing Science 18(2):95–114. https://doi.org/10.1287/mksc.18.2.95

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. Marketing Science 31(3):521–543. https://doi.org/10.1287/mksc.1120.0713

Russell G, Bucklin R, Srinivasan V (1993) Identifying multiple preference segments from own- and cross-price elasticities. Marketing Letters 4(1):5–18. https://doi.org/10.1007/bf00994183

Ruiz F, Athey S, Blei D (2020) SHOPPER: A probabilistic model of consumer choice with substitutes and complements. Annals of Applied Statistics 14(1):1–27. https://doi.org/10.1214/19-aoas1265

Samuelson P (1974) Complementarity: An essay on the 40th anniversary of the Hicks-Allen revolution in demand theory. Journal of Economic Literature 12(4):1255–1289.

Seidman L (1989) Complements and substitutes: The importance of minding p’s and q’s. Southern Economic Journal 56(1):183. https://doi.org/10.2307/1059065

Shocker A, Bayus B, Kim N (2003) Product complements and substitutes in the real world: The relevance of “Other products.” Journal of Marketing 68(1):28–40. https://doi.org/10.1509/jmkg.68.1.28.24032

Uzawa H (1962) Production functions with constant elasticities of substitution. The Review of Economic Studies 29(4):291–299.

Varian H (1993) Microeconomic Analysis, 3rd ed. W. W. Norton and Company, New York.

Weinstein J (2022) Direct complementarity. Working paper, Washington University in St. Louis.

Received: 2025-05-14 Last revision: 2025-07-07 Accepted: 2025-07-29

Efficiency of Regional Transport Systems

Eva Štichhauerová[1], Miroslav Žižka[2]

Abstract. The paper compares the efficiency of 13 regional transportation systems in the Czech Republic. The comparison is based on a two-stage Data Envelopment Analysis using slack-based models. In the first stage, technical efficiency is evaluated, i.e. how efficiently regions are able to arrange transport services from bus and rail operators. The second stage examines service effectiveness, i.e. how efficiently the transport services are used by passengers. At the same time, the influence of external factors such as population density, the density of the road and rail network and the level of motorization of the population on both scores is examined. The research is based on data from 2023. The results show considerable variation between regions. In terms of technical efficiency, five regions (Karlovy Vary, Liberec, Olomouc, Pardubice, Usti and Labem) scored highest. In the second phase, only two systems (South Moravian and Moravian-Silesian regions) achieved the maximum score. Further research showed that the density of transport infrastructure influences the level of technical efficiency. Service effectiveness is positively influenced by population density and inversely proportionally by the density of transport networks and the level of motorization of the population.

Keywords: Data Envelopment Analysis, integrated transport systems, public transport, service effectiveness, slack-based measure model, technical efficiency.

JEL Classification: C61, H54, R42
AMS Classification: 90C08

1 Introduction

Regions in the Czech Republic play an important role in the procurement, organization and financing of regional public transportation. They decide on the scope of public transport, the choice of operators for both regular bus services and regional rail transport, and the price of fares on their territory. To manage public transport, the regions establish specialized companies - regional public transport organizers, which manage regional integrated transport systems. In matters of public transport organization, the regions must work together with the Ministry of Transport, which commissions long-distance interregional rail transport, and with the cities, which operate the urban public transport systems.

The level of interconnection of the different public transport systems (urban, regional, inter- and supra-regional) varies from region to region. Among the oldest systems are the Prague Integrated Transport System, which has been in operation since 1993, and the Integrated Transport System of the Moravian-Silesian Region, which has been in operation since 1995 (Drdla ,2008). On the contrary, in some regions such as South Bohemian or Karlovy Vary, integrated transport systems are still less developed.

The impetus for the research was an article (Sůra, 2024) which sought to compare the efficiency of public funds spent on rail transport in individual regions. In the cited article, efficiency was measured using various simple ratios (e.g. subsidy/train-km, cost/train-km, number of trips per capita, etc.). The aim of this paper is to assess the efficiency of regional transport systems in a multi-criteria manner using the Data Envelopment Analysis (DEA) method. DEA is a suitable method for this purpose because it is used in the evaluation of the efficiency of public services. In addition to rail transport, we also included regular bus services in our research.

2 Theoretical background

The design of efficient transport systems is one of the topical issues in literature. The main objective of integrated transport systems is to improve their performance, both in terms of efficiency and effectiveness. Efficiency measures the correct functioning of a given process, i.e. the efficiency of the allocated resources. Effectiveness

[1] Technical University of Liberec, Faculty of Economics, Department of Business Administration and Management, Studentská 2, Liberec 1, Czechia, eva.stichhauerova@tul.cz.
[2] Technical University of Liberec, Faculty of Economics, Department of Business Administration and Management, Studentská 2, Liberec 1, Czechia, miroslav.zizka@tul.cz.

expresses whether the resources were spent purposefully (Ji et al. 2024). Meta-study (Daraio et al., 2016) provides an overview of the inputs and outputs used in analyses of the technical efficiency of transport systems. Typical inputs include natural indicators such as the number of employees (segmented into drivers and other service staff), number of vehicles, fuel consumption, and vehicle capacity; among the financial indicators used are e.g. capital expenditure, labour costs, fuel and energy costs, maintenance, operating expenses, etc. Outputs can be, e.g. transport performance (vehicle kilometres, vehicle hours), transport performance or services (number of passengers, passenger kilometres, number of journeys, vehicle utilization) and revenues (fare revenues, operating revenues). Furthermore, Daraio et al. (2016) identified factors affecting technical efficiency. These include service characteristics and quality (network length, average speed, vehicle age, number of stops, lines), socio-demographic factors such as population density, individual motorization rate, GDP per capita, unemployment rate, income inequality, etc., organizational factors (contract type, company size, etc.), subsidies and externalities. In terms of methods and approaches, SFA and DEA models are mainly used for the assessment.

The second basic component of performance is effectiveness. This can be measured as either cost (also operational) or service effectiveness. Cost or operational effectiveness relates the services provided (e.g. number of passengers) to the inputs (e.g. number of vehicles, length of network). Service effectiveness measures services provided (e.g. number of passengers, revenue) and outputs (e.g. passenger kilometres). Žižka (2017) used a two-stage DEA to assess the efficiency and effectiveness of public transport operators. In the first stage, the technical efficiency of the system was evaluated based on three inputs (number of vehicles, drivers, fleet age) and one output (passenger kilometres). The second phase measured service effectiveness with passenger-kilometers as the only input and two outputs (passenger numbers and revenue).

DEA can be used to assess the efficiency of public services, particularly where we are faced with the absence of market prices. Public transport is an appropriate area of application because prices (fares) in regional transport (as opposed to commercial long-distance transport) are regulated and cover only part of the costs. The difference between costs and revenues is covered by the contracting authority's subsidy from the regional budget.

DEA is a widely used performance measurement technique that evaluates the efficiency of decision-making units (DMUs) by comparing the ratio of inputs to outputs. Among the various DEA models, the traditional models, namely the Charnes, Cooper, and Rhodes (CCR) model and the Banker, Charnes, and Cooper (BCC) model, are foundational. The CCR model assumes constant returns to scale (CRS), while the BCC model allows for variable returns to scale (VRS), thus providing a more flexible framework for assessing efficiency under different operational conditions (Zhao 2015; Li et al. 2019). Both models are characterized as radial, meaning they focus on proportional reductions in inputs or proportional increases in outputs, which can lead to limitations in accurately capturing inefficiencies, particularly when inputs and outputs do not change in a uniform manner (Nourani et al. 2019). In contrast, slack-based DEA models, introduced by Tone in 2001, address some of the shortcomings of traditional DEA models by incorporating slacks – excess inputs or shortfalls in outputs – into the efficiency measurement process (Tone 2001). The Slack-Based Measure (SBM) model allows for non-proportional adjustments, meaning it can account for situations where inputs and outputs change at different rates, thus providing a more nuanced view of efficiency (Nourani et al. 2019). This approach is particularly beneficial in contexts where DMUs may have varying degrees of inefficiency that are not adequately captured by the radial models, as it directly measures the inefficiencies associated with input excesses and output shortfalls (Maltseva et al. 2020; Rashid and Mohd Zawawi 2024). Moreover, while traditional DEA models provide a scalar measure of efficiency based on the radial approach, the SBM model offers a more comprehensive assessment by allowing for the simultaneous consideration of both input and output slacks (Seth et al. 2024). This results in a more accurate representation of a DMU's performance, as it reflects the real operational inefficiencies rather than just the proportional changes in inputs and outputs (Ajirlo et al., 2019; Vittal et al., 2021). For these reasons, we used the SBM DEA model for our analysis.

3 Data and methodology

For our research, we used data from the 2023 Transport Yearbook (Slabá and Houšť, 2023). For the analysis, we merged the Central Bohemian region and the capital city of Prague into one territorial unit, which is closely linked economically and by transport. From the yearbook we obtained data on the number of passengers carried by rail and public bus transport in each region. From the same source we obtained data on the length of roads, railway lines and the number of registered vehicles in the regions. Data on electrified lines were obtained from the Railway Administration. Another important source was the region's reports on public service obligations for 2023, which contained data on the performance of bus (kilometres) and passenger rail transport (train kilometres) and the amount of compensation paid in both modes of transport (in CZK million). Compensation reflects the difference

between the costs incurred by operators, including a reasonable profit, and the revenue received from passengers. Data on population and area of the regions were obtained from a public database (CZSO, 2025).

The design of the model can be seen in Figure 1. We used a two-stage DEA analysis. The first phase assesses how cost-effectively regions purchase transportation services from bus and rail operators. The second phase measures how efficiently the purchased services are used by passengers. Overall efficiency is then given by the product of the two scores. Finally, the relationship between the individual scores and selected external factors such as population density, degree of motorization, density of transport networks and electrification of rail lines were examined.

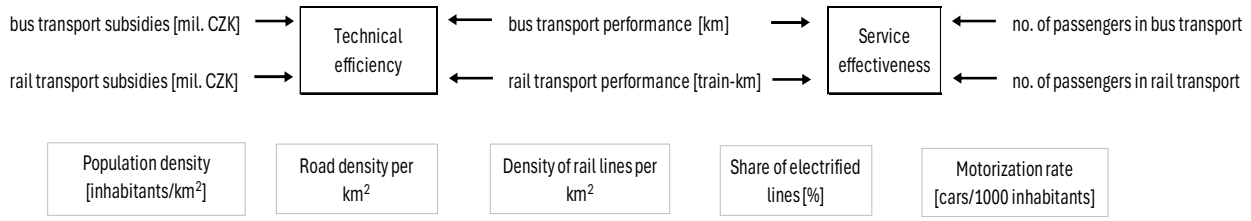


Figure 1 Inputs and outputs of the model phases

The SBM DEA model (1) by Zhu (2014) with CRS assumption was used to analyze the effectiveness and efficiency.

$$\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to

$$\sum_{j=1}^{n} \lambda_j x_{ij} + s_i^- = x_{i0} \quad i = 1, 2, \dots, m; \tag{1}$$

$$\sum_{j=1}^{n} \lambda_j y_{rj} - s_r^+ = y_{r0} \quad r = 1, 2, \dots, s;$$

$$\lambda_j, s_i^-, s_r^+ \geq 0$$

where x_{ij} are inputs, y_{ij} are outputs, λ_j is the weight of the DMU, s^+ a s^- are slack variables.

The non-parametric Spearman rank-correlation coefficient r_s (2) taking values between -1 and +1 was used to assess the relationships between the individual scores calculated by the SBM DEA model and external factors.

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} (x_{ri} - y_{ri})^2}{n(n^2 - 1)} \tag{2}$$

Where x_{ri} a y_{ri} are the order within the ascending order of the observed values of the variables x and y respectively and n is the number of subjects observed.

4 Results

Table 1 shows the results of the SBM DEA models for both phases. Five regions (Karlovy Vary, Liberec, Olomouc, Pardubice, Usti and Labem) were the best ranked in terms of their ability to provide transport performance from individual operators in bus and rail transport for the financial resources spent. The average efficiency rate of the first phase is quite high (0.8943). The largest reserves in this aspect exist in the transport system of the South Bohemian Region. Using the DEA model, peer efficient systems can be designed for the inefficient units, see Table 2, including the relevant weights. The average effectiveness score in the second phase was significantly lower, only 0.5811. This shows that the supply of transport services does not always meet adequate demand on the part of passengers. This may be a signal that it is either oversized or inappropriately matched to passengers' needs. Only two systems - in the South Moravian and Moravian-Silesian regions - achieve the maximum score of 1.0. In the second phase, integrated transport systems in the South Bohemian and Usti and Labem regions proved to be

the least efficient. Analogous to the first phase, suitable benchmarks including their weights can be established for the inefficient units.

Region	Technical efficiency	Service effectiveness	Overall efficiency
Moravian-Silesian	0.8263	1.0000	0.8263
South Moravian	0.6350	1.0000	0.6350
Pilsen	0.7400	0.7373	0.5456
Olomouc	1.0000	0.5355	0.5355
Prague + Central Bohemian	0.8779	0.5731	0.5031
Karlovy Vary	1.0000	0.4982	0.4982
Vysocina	1.0000	0.4978	0.4978
Zlin	0.8010	0.6188	0.4960
Pardubice	1.0000	0.4762	0.4762
Liberec	1.0000	0.4534	0.4534
Hradec Kralove	0.9346	0.4610	0.4308
Usti and Labem	1.0000	0.3600	0.3600
South Bohemian	0.8110	0.3428	0.2780
Mean	0.8943	0.5811	0.5027

Table 1 Efficiency and effectiveness scores of integrated transport systems

Table 1 shows that the most efficient in terms of technical efficiency and service effectiveness is the integrated transport system of the Moravian-Silesian region, and with some distance, the system in the South Moravian region. On the contrary, the least efficient system is the one in the South Bohemian region, followed by the transport system in the Usti and Labem region.

Region	Technical efficiency - Benchmark systems	Service effectiveness – Benchmark systems
South Moravian	Olomouc (0.50), Pardubice (1.23)	x
Moravian-Silesian	Olomouc (0.49), Pardubice (0.82)	x
Prague + Central Boh.	Olomouc (1.10), Pardubice (1.81)	South Moravian (1.02)
Hradec Kralove	Pardubice (1.14)	South Moravian (0.16), Moravian-Silesian (0.08)
Pardubice	x	South Moravian (0.16), Moravian-Silesian (0.07)
Liberec	x	South Moravian (0.20)
Olomouc	x	South Moravian (0.36), Moravian-Silesian (0.10)
Pilsen	Pardubice (1.14)	Moravian-Silesian (0.46)
Zlin	Olomouc (0.39), Pardubice (0.40)	South Moravian (0.32)
Karlovy Vary	x	South Moravian (1.00)
Usti and Labem	x	South Moravian (0.04), Moravian-Silesian (0.35)
South Bohemian	Olomouc (0.08), Pardubice (1.11)	South Moravian (0.20)
Vysocina	x	South Moravian (0.24)

Table 2 Peer systems for inefficient transport systems (lambda values in brackets)

In the next stage, the correlation coefficients between the individual technical efficiency and service effectiveness scores and selected indicators characterizing external factors in the region were determined, see Table 3.

Score	Population density	Road density	Density of rail lines	Electrified lines	Motorization rate
Technical efficiency	-0.1590	0.5115	0.4450	0.0954	-0.0838
Service effectiveness	0.4480	-0.3714	-0.3631	0.0275	-0.3109

Table 3 Spearman rank-correlation coefficients

Table 3 shows that there is a moderately strong correlation between levels of technical efficiency and service effectiveness and transport infrastructure in terms of the density of the road and rail network. This suggests that it is technically and economically easier to organize transport services in areas where there is a dense transport

network. However, in the case of service effectiveness, the relationship is inversely proportional. Thus, better transport infrastructure reduces the rate of public transport use. Similarly, the correlation between service effectiveness and the degree of motorization is also inversely proportional. A lower degree of motorization leads to a higher use of public transport and vice versa. Population density also has a moderate effect on service effectiveness. This means that in a densely populated area, the use of public transport increases. Furthermore, it was found that the proportion of electrified rail lines is essentially unrelated to either technical efficiency or service effectiveness.

5 Conclusions

The results of the research show that there are shortcomings in the effective functioning of regional integrated transport systems in the Czech Republic, both in transport performance provision and in their use by passengers. Regions, or the organizations they set up to operate integrated transport systems, can improve technical efficiency by better selecting operators based on open public tenders for individual territorial units. Service effectiveness can be improved by matching the supply of transport services to demand. In some cases, the current supply of transport services may appear too high and should be reduced. However, lower use of public transport systems may also be because the quality of existing services does not match the expectations of potential passengers. This means, for example, that services run at inconvenient times, connections between bus and rail do not wait for each other, the fare system is too complicated, etc. In this case, the degree of effectiveness can be increased by developing the quality of the integrated transport system, which should induce new demand. Integrated transport systems in the Czech Republic have a different history and are developed in different ways in terms of the interconnection of modes and tariff integration. For example, in the South Bohemian region, which emerges from this analysis as the lowest ranked, full integration of bus, urban and rail transport has not yet occurred and is only planned. On the other hand, the systems in the Moravian-Silesian, South Moravian region and Prague Integrated Transport (which has gradually expanded to cover the whole territory of the Central Bohemian region) have been in operation for many years and the degree of integration of the different modes of transport is at a high level.

Among the external factors, technical efficiency was found to be moderately and directly proportional to the density of the road and rail network. This implies that it is easier to organize transport performance in the regions with a dense network. Vehicles have shorter turnaround times and can carry out a higher volume of operations in a given time. Examples are Usti and Labem, Pardubice and Central Bohemian regions. In contrast, in the South Bohemian and Zlin regions, the density of transport infrastructure is lower and the technical efficiency scores come out lower. A larger number of factors influence service effectiveness. It is confirmed that population density has a positive impact on the use of transport services. Higher population density generates stronger transport flows, which has a positive impact on the service effectiveness measure. This is the case for South Moravian or Moravian-Silesian regions, whereas regions with a scarce population (South Bohemian, Karlovy Vary, Vysocina) have worse service effectiveness rates. A certain exception is Prague and the Central Bohemian region, which show only an average level of use of the transport services offered. However, this result may also be influenced by social and political factors. In Prague, for example, children up to 15 years of age travel free of charge (unlike in other regions), which has an impact on the statistics for paying passengers. The number of free passengers is not reflected in the statistics. Furthermore, the density of the transport network and the level of motorization have an indirect proportional effect on service effectiveness. For example, the above three regions with high density of transport infrastructure (Usti and Labem, Pardubice and Central Bohemian regions) have lower service effectiveness scores. The Moravian-Silesian region has the lowest motorization rate and a service effectiveness score of one. On the other hand, Prague and the Central Bohemian, Pilsen and South Moravian regions have the highest motorization rates while having lower service effectiveness scores.

Finally, it should be noted that only 13 units entered the evaluation, which can be seen as a limit of the research. However, it should be remembered that these are all territorial units and from this perspective the researchers worked with a population. It must also be considered that public regional transport does not operate as a purely commercial service but is influenced by various social and political influences (typically the level of fares, the extent of free transport) which make it particularly difficult to compare with each other. It is in such an environment that the DEA method can be a useful tool for assessing the efficiency of public spending.

Acknowledgements

Supported by institutional funds for long-term conceptual development of the research organization of the Technical University of Liberec. Project: Multi-criteria approaches to measuring business performance.

References

Ajirlo SF, Amirteimoori A, Kordrostami S (2019) Two-stage additive integer-valued data envelopment analysis models: A case of Iranian power industry. J Model Manag 14:199–213. https://doi.org/10.1108/JM2-12-2017-0132

CZSO (2025) Statistics VDB. https://vdb.czso.cz/vdbvo2/faces/en/index.jsf?page=statistiky. Accessed 6 Feb 2025

Daraio C, Diana M, Di Costa F, et al (2016) Efficiency and effectiveness in the urban public transport sector: A critical review with directions for future research. Eur J Oper Res 248:1–20. https://doi.org/10.1016/j.ejor.2015.05.059

Drdla P (2008) IDS v České republice - Srovnání a zvláštnosti. Perner´s Contacts 3:69–74

Ji W, Huang Z, Gao G, Zheng P (2024) Evaluation of integrated transport efficiency and equity at the county level——taking the counties in ningbo city as an example. Transp Policy 148:257–272. https://doi.org/10.1016/j.tranpol.2024.01.011

Li Y, Chiu Y, Lu LC (2019) New Energy Development and Pollution Emissions in China. Int J Environ Res Public Health 16:1764. https://doi.org/10.3390/ijerph16101764

Maltseva V, Na J, Kim G, Ha H-K (2020) Efficiency Analysis of Russian rail freight transportation companies with super slack-based measurement Data Envelopment Analysis. J Int Logist Trade 18:77–89. https://doi.org/10.24006/jilt.2020.18.2.077

Nourani M, Ting IWK, Lu W-M, Kweh QL (2019) Capital structure and dynamic performance: Evidence from ASEAN-5 banks. Singap Econ Rev 64:495–516. https://doi.org/10.1142/S021759081650034X

Rashid NRM, Mohd Zawawi IS (2024) Measuring the Efficiency of water supply service: Application of Charnes, Cooper Rhodes and slack-based measure models. ASM Sci J 19:1–6. https://doi.org/10.32802/asmscj.2023.1036

Seth H, Deepak D, Ruparel N, et al (2024) Assessment of working capital management efficiency – a two-stage slack-based measure of data envelopment analysis. Manag Finance 50:1344–1365. https://doi.org/10.1108/MF-08-2020-0432

Slabá R, Houšť R (2023) Transport Yearbook Czech Republic 2023

Sůra J (2024) Až čtyřnásobný rozdíl. Unikátní srovnání ukazuje, kolik kraje stojí cestující ve vlacích. In: Zdopravy.cz. https://zdopravy.cz/nezvykle-porovnani-kolik-stoji-jednotlive-kraje-zeleznicni-doprava-na-kilometr-a-ktery-je-vlakovy-premiant-225448/. Accessed 20 Jan 2025

Tone K (2001) A slacks-based measure of efficiency in data envelopment analysis. Eur J Oper Res 130:498–509. https://doi.org/10.1016/S0377-2217(99)00407-5

Vittal B, Nellutla R, Reddy MK (2021) Selection and Analysis of input-output variables using Data envelopment analysis of decision making units - Indian private sector banks. Int J Eng Adv Technol 10:119–127. https://doi.org/10.35940/ijeat.E2674.0610521

Zhao H (2015) The Empirical Study on R&D Efficiency of High-Tech Industry in China. Int J Trade Econ Finance 6:49–52. https://doi.org/10.7763/IJTEF.2015.V6.441

Zhu J (2014) Quantitative Models for Performance Evaluation and Benchmarking, 3rd edn. Springer International Publishing, Cham

Žižka M (2017) An Assessment of the efficiency and effectiveness of the services of urban transport operators in the Czech Republic. Transform Bus Econ 42:134–152

Received: 2025-05-23 Last revision: 2025-06-24 Accepted: 2025-06-25

Assignment mutation in Evolution Algorithm to improve solutions in Flexible Job Shop Scheduling

František Koblasa[1], Miroslav Vavroušek[2]

Abstract. The Flexible Job Shop Scheduling Problem (FJSP) is widely recognised for its ability to model complex manufacturing environments. A key challenge in FJSP is the assignment of operations to available machines, as machine selection directly influences production efficiency, resource utilisation, and overall scheduling performance. Besides the crossover operator, which influences machine assignments, mutation plays an important role in the Evolution algorithm. Random mutation is usually used as a tool that prevents optimisation from getting stuck in local optima and enriches the population of solutions so that they do not converge to one solution.
This article focuses on enhancing the machine selection process within evolutionary algorithms by introducing mutations that are not strictly random. It uses information about a particular scheduling instance to suggest changing machine assignments.
The research evaluates the impact of assignment mutation on convergence and computational efficiency in comparison with a simple evolution algorithm without mutation and with a standard random mutation technique.
To validate the approach, known FJSP benchmark instances are used. The experimental framework assesses how different deterministic-based mutation strategies influence scheduling outcomes, providing insights into their practical applicability in real-world flexible job shop environments.

Keywords: flexible job shop scheduling problem, evolution algorithm, mutation, machine assignment.

JEL Classification: C63, L23
AMS Classification: 90C59

1 Introduction

This paper is inspired by a specific real-world scheduling problem in the moulding industry, where very few operations with a high volume of machines can process particular operations (moulding, drying, assembly, packing, etc.). This model can be described as a Flexible Job Shop Scheduling Problem (FJSP). FJSP is a well-established model for representing complex manufacturing environments. It extends the classic Job Shop Scheduling Problem, which is already NP-hard, by allowing operations to be assigned to one of several alternative machines. This flexibility more accurately reflects real-world production settings, where multiple machines may be capable of performing the same task. As such, it was selected to deal with the above-mentioned manufacturing system.

Besides classical dispatching rules, metaheuristic algorithms, especially population-based ones, such as Evolutionary (genetic) algorithms, are used to solve FJSP. Mutation operators play a key role in such algorithms by introducing variability and preventing premature convergence. In FJSP, mutations typically target either operation sequencing or machine selection. This paper focuses on how machine assignment mutations influence the optimisation timespan and single criterion of makespan. There are four research questions (RQ) defined to study the influence of the mutation parameter and develop a new one:

- RQ1 – Is a type of mutation influencing the final objective function?
- RQ2 – Is the mutation type influencing the generation in which the population converges to a local extreme?
- RQ3 – Is the mutation type influencing the optimisation timespan?
- RQ4 – What are the chances that a mutation will improve or worsen the solution?

In order to answer the defined RQ, this paper is organised as follows. The second chapter focuses on a mathematical model of FJSP and assignment decision-making problems. The third chapter focuses on known mutation operators to find research gaps and to prepare experiments. The fourth chapter proposes mutations to be tested and

[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

evaluated. Theoretical benchmarks are used in the following chapter to evaluate the quality of the schedule (makespan), convergence of the population and optimisation timespan in order to answer before defined RQ.

2 Definition of the FJSP model

A flexible Job Shop (FJSP) scheduling problem is an NP-hard problem which consists of proper sequencing and job assignment operation to achieve the optimal value of the objective function, which is most typically, among others, possible to select, makespan i.e. C_{max} (1), maximal completion time c_{ini}. There are, in general, 5 major mathematical models (Demir and Kürşat İşleyen, 2013) available to describe FJSP and for the purpose of this paper simplified M3 model is used (Gao et al., 2006; Sun et al., 2015). Equation (2) ensures that the job follows specific operation sequence using completion time c_{ik} of operation O_{ik} and processing time t_{ikj} of kth operation of job i. Disjunctive constraints (3) represent that operation O_{gh} should not start before O_{ik} or that the operation O_{hg} must be completed before the starting of operation O_{ij} if they are assigned on the same machine j by decision variables x_{ikj} resp. x_{hgj}. Constraint (4) ensures that the operation will be processed by one machine from the set of machines A_{ik} at one time. Constraint (5) defines the non-negativity of completion time, and constraint (6) defines the decision variable to be boolean. Job indexes are noted as i and h where $i, h = 1,2.,\ldots,n$; j is machine index, $j = 1,2,\ldots,m$; k, g are operation indexes; $k, g = 1,2,\ldots, n_i$,where n_i is the total number of operations of job i.

$$C_{max} = max\{c_{in_i}\} \tag{1}$$

$$\text{s.t. } c_{ik} - c_{i(k-1)} \geq t_{ikj}x_{ikj}, k = 2, \ldots, n_i, \forall i, j \tag{2}$$

$$\left[(c_{hg} - c_{ik} - t_{hgj})x_{hgj}\right] \geq 0 \ \vee \left[(c_{ik} - c_{hg} - t_{ikj})x_{ikj} \geq 0\right] \forall i, j, g, h \tag{3}$$

$$\sum\nolimits_{x_{ikj} \in A_{ik}} x_{ikj} = 1 \, \forall i, k \tag{4}$$

$$c_{ik} \geq 0, \forall i, k \tag{5}$$

$$x_{ikj} \in \{0,1\}, \forall i, k, j \tag{6}$$

This paper focuses on selecting proper machine j from the list of available machines A_{ik} in genetic algorithm operator of mutation that are described in the following chapter.

3 Mutation operators in GA for FJSP – types and probabilities

Two key parameters need to be defined when applying mutation in Genetic Algorithms (GA). First, what type of mutation should be used? Second, how likely is this mutation to be applied to let solutions escape local optima and to improve the current solution? There are, in general, two types of mutations in FJSP, as it contains two decision-making problems (job sequencing and machine selection). This paper focuses mainly on machine assignment. However, it has to be pointed out that sequencing mutation is, in general, mostly applied as a random change in the sequence of a given solution or as hybridisation by knowledge-based local search.

Similar to sequencing mutation, the most commonly used mutation in machine assignment is a random change in the chromosome gene. In machine selection, it means selecting an operation and switching machines to another random machine (Yao et al., 2024). An alternative to a random change in machine genes is a two-point swap mutation. (Chen et al., 2025) introduces a two-point swap mutation in which two operations are randomly selected, and their machines are swapped. That requires finding operations with the same list of machines (Mei et al., 2024). In theoretical problems with maximal machine flexibility (all operations can be done on all machines), it is not a computationally complex operation; however, in real-life problems, it can increase computation complexity significantly with rising differences in the vectors of available machines.

Modification of above mentioned can be inheriting a machine from another randomly selected gene (Li et al. 2024b). A pair of machine genes is selected with a given probability, and based on another random selection, one gene inherits from another if a random threshold is passed. Inversion mutation randomly selects a chromosome sequence and reverses the sequence of particular genes (Abdel-Basset et al., 2023). All the above-mentioned mutations require, as in other cases, that the change to the machine is feasible. If not, the change has to be discarded or the chromosome repaired.

Completely random mutation is usually an efficient way to suppress the convergence of the population too fast to the local optimum. However, there are improved approaches which use knowledge about a particular problem at increased computational cost.

(Ma et al., 2025) presented four types of knowledge-based mutations. First, it selects a random operation, and it changes the machine to the one with the shortest processing time (Hao et al., 2025). The second one is that it first selects machines with the highest utilisation, then picks a random operation from the list of operations assigned to this machine and redirects that to the machine with lower utilisation (Yang et al., 2025). The third type relocates operations from machines with higher than average utilisation to machines with utilisation lower than average. The last one is first analysing the critical path, and then reassigning every operation on the critical path to machines with lower utilisation load. The critical path is also presented in (Deng et al., 2025; Zhao et al., 2024), where a random critical path operation is selected and put on a random machine available from the list.

Completion time (makespan) is used in (Gong et al., 2024) where a random operation is selected on a machine with the highest makespan, and an alternative machine with a lower makespan is selected.

Besides the type of mutation, the probability of applying it is a key mutation parameter. There are two major systems used while determining if the mutation is used or not – gene and individual mutation percentage. In individual gene percentage, there is usually a defined probability that an individual gene will be changed, and the algorithm runs through the whole set of genes of all individuals. When a random number is lower than predefined percentage P_m=<0;1> (e.g. P_m=0.1), a predefined mutation is performed. Setting an appropriate level of P_m is usually based on a rule of thumb (Chen et al., 2025) as or a matter of parameter tuning (Deng et al., 2025) usually done by an orthogonal experiment. Generally, in the case of gene-defined mutation, it is state-of-art knowledge that P_m>0.1 makes GA closer to random search as parent inheritance is broken (Rivera et al., 2022).

In an individual mutation, P_m represents the probability that the mutation will be performed not on a gene, but on an individual. P_m higher than in single gene mutation, reaching values up to P_m=0.3-0.6. In those cases, an individual is selected to perform a usually complex mutation on the whole gene.

4 Mutation operation design

Based on the review in the previous chapter, there are two major parameters to set when designing a mutation type in GA/EA for FJSSP machine assignment. The first is a way of selecting a gene to be changed, followed by placing the operation on another machine. Figure 1 provides a classification overview of the defined parameters along with known and designed parameters.

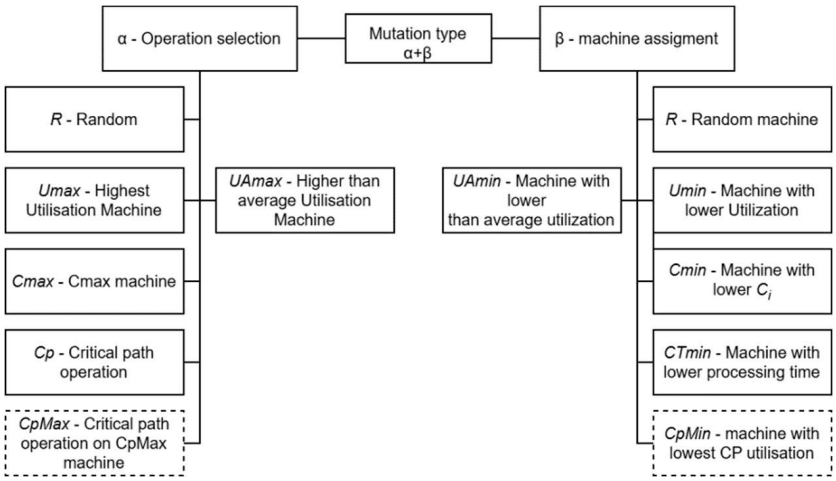


Figure 1 Mutation types categories – Operation selection and machine assignment.

In the first parameter (α - operation selection), the machine gene random selection (*R*) is the most commonly used one, as it is simple to implement and requires no additional knowledge about the scheduling problem. Selection of random operation from the machine with the highest Utilisation (*Umax*), or random operation from the random machine with higher than average utilisation (*UA*) is a typical knowledge-based variant. Single criterion (minimisation of total competition of all jobs) mutation selection is represented by selecting a random operation from a machine that defines the makespan (*Cmax*). The most sophisticated operation selection is based on critical path theory, and it involves selecting a random operation from the list of critical operations (*Cp*).

The second parameter is machine assignment (β). In operation selection, assignment to the random machine is the most common (*R*). It is mostly combined with a random selection of operations (*R*+*R*); however, random swap, where the machine assignment of two randomly selected operations is also used. Machine assignment based on utilisation consists of selecting a random operation from the machine with a utilisation lower than the current

machine utilisation (*Umin*) or the machine with the lowest average utilisation (*UAmin*). They are usually combined in two reasonable pairs (*Umax+Umin* and *UAmax+UAmin*). Assignment to a machine with an earlier competition time (*Cmin*) is usually paired with *Cmax*. Selection of a shorter processing time machine (*CTmin*) is generally used with all the above-mentioned systems, most frequently with random selection (*R+CTmin*).

Based on the review and based on observed properties of real-world manufacturing systems (moulding industry), two new parameters were defined. A list of critical path operations is usually large, so the chance of finding the proper critical operation to be changed can be low. In order to minimise the list of available operations, *α=CpMax* is defined. *CpMax* represents the selection of an operation which lies on the Critical path and on the machine which contains the largest amount (from the duration of the cycle time perspective) of such operations. *β= CpMin* represents a machine which has lower utilisation by CP operations. Both of operators are inspired by combining *α=Cp* and *α=Umax* with *β=Umin.*

5 Tests and result discussion

The above-defined mutations are tested on the Evolution algorithm during the reproduction process, which follows the generally known process of selecting parents for reproduction-crossover-mutation-elimination. Tested EA is using a double (sequencing and assigning) Random key chromosome representation of *N*=200 individuals in population, roulette wheel selection of parent solutions, uniform crossover with crossover probability $P_c = 0.25$, set of seven defined mutations defined in previous chapter (see Figure 1 and Table 1) with mutation probability $P_m = 0.3$ to mutate individual (not a gene) and elitist elimination strategy where parents and childes are sorted by objective function of makespan (1) and better *N* individuals survives in to new generation. The algorithm has three stopping conditions. It stops after a maximal number of generations*, G=200,* after the population converges to a local optimum, or if there is no improvement after 20 generations.

To evaluate the above-defined RQ, EA is tested on known MK02 and MK07 FJSP problems (Brandimarte, 1993), which are the closest in terms of number of operations and machine assignment flexibility to the real-world example mentioned before. MK02, with a theoretical lower bound LB=24, consists of 10 jobs with 5-7 operations each, which can be processed on all 6 available machines. MK07 (LB=133) contains 20 jobs with 5 operations each, which can be processed on all of 5 machines.

Table 1 presents the best, average, and standard deviation of makespan results after 20 repetition runs, along with the optimisation timespan of one generation "t" and the generation in which optimisation has stopped (Conv.). Runs with mutation are compared with EA, which does not have a mutation operator (*No mutation*).

Mutation	Cmax MK02 LB = 24					Cmax MK07 LB= 133				
α+β	Best	Avg	SD	t [s]	Conv.	Best	Avg	SD	t [s]	Conv.
No mutation	32	34.1	1.5	10.5	74.0	152	169.0	10.1	12.7	90.7
R+R	31	33.2	1.3	10.8	84.1	150	159.1	4.1	14.2	116.6
R+CTmin	30	31.9	1.3	10.6	73.9	153	159.3	5.3	18.2	100.2
Umax+Umin	31	32.6	1.2	14.0	89.0	150	160.9	6.9	20.3	108.7
CP+R	29	32.9	3.8	13.9	90.0	150	156.6	4.9	22.7	115.6
CP+Ctmin	29	30.6	1.2	14.0	81.3	150	156.5	5.1	22.1	107.9
CP+Umin	29	32.1	1.9	13.6	85.8	149	157.9	5.6	28.9	113.9
CPmax+CPmin	30	32.4	1.8	13.7	80.1	152	159.2	5.0	32.2	112.0

Table 1 MK02 and MK07 makespan and optimisation process statistics

Results show that the mutation has a significant impact on both the objective function and convergence. However, the type of mutation had no statistically significant (data did not have a normal distribution, Kruskal-Wallis was used) impact on end results – both makepan (RQ1) and generation in which generation (RQ2) stops while testing on MK02 and MK07. Mutation type does influence the optimisation timespan significantly (RQ3), especially in the case of *α= Umax, CP* and *CPmax* and *β=Umin, CPmin*, as there is a necessity to calculate utilisation (*Umax*) and critical path (CP). It is questionable to use more sophisticated mutations, such as *Umax+Umin* and *CP*;*CPmax+β*, with a higher timespan, but no significant influence on the objective function. Differences are higher in MK07 than in MK02, as there is a significantly different number of operations (100 vs 58 operations), so the critical path (search space for mutations with CP) state space is expected to be bigger.

Further, mutations were evaluated from the point of improving the solution during evolution (RQ4). Results show that there is a significant difference in changing an individual for both better and worse (see Table 2) from a point of makespan. There is also a significant difference based on which state of optimisation the change is made.

Probability of making better individual is much higher in early stages of optimisation (Q1 – first quarter of generations), than in the last (Q4).

MK02 [%] Chance of Better / Worse individual														
Q	*R+R*		*R+CTMin*		*Umax+Umin*		*Cp+R*		*Cp+CTMin*		*Cp+Umin*		*CpMax+CpMin*	
	Better	Worse	Better	Worse	Better	Worse	Better	Worse	Better	Worse	Better	Worse	Better	Worse
Q1	17.97	29.89	15.97	12.68	26.45	17.26	48.29	27.78	41.60	9.54	43.73	18.41	51.18	20.92
Q2	8.86	41.62	7.85	13.45	18.90	25.55	29.45	42.97	25.53	11.76	27.29	30.65	37.96	31.27
Q3	4.37	56.85	3.56	16.06	10.61	41.20	14.10	58.27	11.47	16.51	15.07	42.16	20.25	44.36
Q4	2.60	66.60	2.25	21.27	5.08	58.31	5.68	70.81	4.73	23.09	6.86	53.03	10.29	53.02
MK07 [%] Chance of Better / Worse individual														
Q	*R+R*		*R+CTMin*		*Umax+Umin*		*Cp+R*		*Cp+CTMin*		*Cp+Umin*		*CpMax+CpMin*	
	Better	Worse	Better	Worse	Better	Worse	Better	Worse	Better	Worse	Better	Worse	Better	Worse
Q1	18.56	30.70	13.00	11.28	27.31	19.87	55.55	24.39	39.49	7.12	45.08	14.07	57.95	20.33
Q2	9.46	46.34	5.45	13.02	21.18	25.97	37.40	39.61	24.92	7.66	35.72	22.28	44.66	30.56
Q3	4.15	65.02	2.38	16.67	13.94	38.67	16.16	58.10	11.71	11.97	13.99	45.14	21.71	50.88
Q4	2.35	75.28	1.24	18.38	6.62	54.09	5.61	73.46	3.64	12.32	4.82	60.97	6.32	70.80

Table 2 Percentage of improvement and worsening individuals by mutation

As the main purpose of using mutation is to keep the population diverse and make some large shifts, the least desirable quality is to have the same solution from the perspective of the objective function. There are several mutations which have a significantly higher chance of generating the same solution, which can be calculated as 100%-Better-Worse. *R+CTMin* (71-82%) and *CP+CTMin* (40-80%) were mutations which has the highest percentage of the same solutions, and such are not fitting the defined role of the mutation. Preferred are mutations (*Cp+R* and *CpMax+CpMin*) that make individuals different in the majority of cases (are the same in only 20-40%). While comparing mutations based on the improvement, significant differences are identified. *CpMax+CpMin* makes better solutions in 6-10% cases, even in the last moments of optimisations (Q4). Generally, the best from points of diversity, quality of mutation and stability of results measured by makespan is *Umax+Umin*, which has about 5% of improved and 50% of worsened solutions with just 45% of the same solutions. However, this assumption is limited by results from just two FJSSP cases (MK02;MK07) and as a result, *CpMax+CpMin* indicates that results can be very different in different models.

6 Conclusion

The presented research shows that while the mutation operators in Evolutionary Algorithms for FJSSP significantly influence convergence and maintain population diversity, the specific type of mutation used does not consistently lead to statistically significant improvements in final solution quality (RQ1) or reaching high number generations in which the population is diverse and able to improve (RQ2).

Simpler mutations, such as random operation and machine selection, offer lower computational cost, while more sophisticated, knowledge-based mutations, as those involving utilisation and critical path analysis, require more computational resources (RQ3) without always yielding superior makespan outcomes. Nevertheless, certain combinations like *CpMax+CpMin* and *Umax+Umin* show potential in improving diversity (RQ4), mutation quality, and performance stability, especially in larger problem instances (MK07).

That leads to the assumption that random-based mutations are more suitable for frequent applications (higher P_m), while mechanisms such as *CpMax+CpMin* and *Umax+Umin* are better suited to hybridising EA in the form of a local search to find a better solution. It also seems that selection, elimination and crossover operators have a higher influence on the optimisation process than mutation. In order to prove that, experiments using mutation as the only reproduction parameter and studying the influence of the above-mentioned by applying different kinds of parameters have to be performed. This can be studied in further research, along with testing it on more benchmark instances, in order to have more generalised evaluations.

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 2025. The research reported in this paper was supported by institutional support for nonspecific university research.

References

Abdel-Basset M, Mohamed R, El-Shahat D, Sallam KM (2023) An efficient hybrid optimization method for Fuzzy Flexible Job-Shop Scheduling Problem: Steady-state performance and analysis. Eng. Appl. Artif. Intell. 123, 106249. https://doi.org/10.1016/j.engappai.2023.106249

Brandimarte P (1993)Routing and scheduling in a flexible job shop by tabu search. Ann. Oper. Res. 41, 157–183. https://doi.org/10.1007/BF02023073

Chen J, Wang C, Xu B, Liu S (2025) Discrete Multi-Objective Grey Wolf Algorithm Applied to Dynamic Distributed Flexible Job Shop Scheduling Problem with Variable Processing Times. Appl. Sci. 15, 2281. https://doi.org/10.3390/app15052281

Demir Y, Kürşat İşleyen, S (2013) Evaluation of mathematical models for flexible job-shop scheduling problems. Appl. Math. Model. 37, 977–988. https://doi.org/10.1016/j.apm.2012.03.020

Deng L, Qiu Y, Di Y, Zhang L (2025) A knowledge-driven memetic algorithm for distributed green flexible job shop scheduling considering the endurance of machines. Appl. Soft Comput. 170, 112697. https://doi.org/10.1016/j.asoc.2025.112697

Gao J, Gen M, Sun L (2006) Scheduling jobs and maintenances in flexible job shop with a hybrid genetic algorithm. J. Intell. Manuf. 17, 493–507. https://doi.org/10.1007/s10845-005-0021-x

Gong G, Tang J, Huang D, Luo Q, Zhu K, Peng N, (2024) Energy-efficient flexible job shop scheduling problem considering discrete operation sequence flexibility. Swarm Evol. Comput. 84, 101421. https://doi.org/10.1016/j.swevo.2023.101421

Hao L, Zou Z, Liang X (2025) Solving multi-objective energy-saving flexible job shop scheduling problem by hybrid search genetic algorithm. Comput. Ind. Eng. 200, 110829. https://doi.org/10.1016/j.cie.2024.110829

Ma J, Gao W, Tong W (2025) A deep reinforcement learning assisted adaptive genetic algorithm for flexible job shop scheduling. Eng. Appl. Artif. Intell. 149, 110447. https://doi.org/10.1016/j.engappai.2025.110447

Mei Z, LuY, Lv L (2024) Research on Multi-Objective Low-Carbon Flexible Job Shop Scheduling Based on Improved NSGA-II. Machines 12, 590. https://doi.org/10.3390/machines12090590

Rivera M.M., Ochoa-Zezzatti A, Serna SP, (2022) Chapter 3 - Embedded system for model characterization developing intelligent controllers in industry 4.0, in: Hassanien, A.E., Chatterjee, J.M., Jain, V. (Eds.), Artificial Intelligence and Industry 4.0, Intelligent Data-Centric Systems. Academic Press, pp. 57–91. https://doi.org/10.1016/B978-0-323-88468-6.00004-8

Sun L, Lin L, Wang Y, Gen M, Kawakami H (2015) A Bayesian Optimization-based Evolutionary Algorithm for Flexible Job Shop Scheduling. Procedia Comput. Sci., Complex Adaptive Systems San Jose, CA November 2-4, 2015 61, 521–526. https://doi.org/10.1016/j.procs.2015.09.207

Yang Z, Hu X, Li Y, Liang M, Wang K, Wang L, Tang H, Guo S(2025) A Q-learning-based improved multi-objective genetic algorithm for solving distributed heterogeneous assembly flexible job shop scheduling problems with transfers. J. Manuf. Syst. 79, 398–418. https://doi.org/10.1016/j.jmsy.2025.02.002

Yao Y, Wang Q, Wang C, Li X, Gao L, Xia K (2024) Knowledge-based multi-objective evolutionary algorithm for energy-efficient flexible job shop scheduling with mobile robot transportation. Adv. Eng. Inform. 62, 102647. https://doi.org/10.1016/j.aei.2024.102647

Zhao S, Zhou H, Zhao Y, Wang D (2024) DQL-assisted competitive evolutionary algorithm for energy-aware robust flexible job shop scheduling under unexpected disruptions. Swarm Evol. Comput. 91, 101750. https://doi.org/10.1016/j.swevo.2024.101750

Received: 2025-06-05 Last revision: 2025-07-16 Accepted: 2025-07-29

On the efficient design of a network under multiple criteria

Robert Hlavatý [1], Helena Brožová[2]

Abstract. The problem of searching for a maximum flow between a source and a terminal in a network with limited edge capacities is a classic problem of graph theory. We assume that the edge capacities are not explicitly given but subject to a predefined lower and upper bound. We aim to design a network that would enable maximum flow between two vertices, while additional criteria may be considered, such as the cost of edge capacity building or flow risk mitigation. We propose a methodology for constructing the optimal edge capacities with respect to a given budget. The optimal edge capacities represent an efficient design in a multiobjective sense. Our approach is inspired by the DeNovo optimisation framework with specific adjustments. We demonstrate our methodology using an artificial example. A comparison with traditional multiobjective approaches is shown, and it turns out that none of them can be considered superior to the others, as is usual in multiobjective optimisation. However, the average performance of our approach appeared to outperform the other methods, which implies that it does matter which method is applied for a particular case.

Keywords: DeNovo optimisation, efficient design, metaoptimisation, multiobjective optimisation, network flow

JEL Classification: C61
AMS Classification: 05C21, 90C17, 90C35

1 Introduction

Network flows in the operations research context have been extensively studied since the original paper of Ford and Fulkerson, 1956. Comprehensive books (Ahuja et al, 1993, Finke, 2008) exist on the topic as well as surveys (e.g., Aronson, 1989, Cruz-Mejía and Letchford, 2023). Our contribution deals with the case of network flows wherein the uncertainty is involved. The issue of uncertainty in network flows is justified by the practical requirements for solving these types of problems. Numerous papers have discussed different methodologies for capturing uncertainty. In the network flow context, the uncertainty is usually imposed on edge capacities. One can mostly encounter a stochastic or robust approach. For the earlier, Thapalia et al, 2012 investigate the problem optimality with stochastic edge capacities and compare it to its deterministic variant. Such an approach has practical uses, as shown by Mozafari and Zabihi, 2020, who use the stochastic approach to solve the water supply problem. Their methodology is combined with a robust approach, which is another suitable way to deal with uncertainty. One can discover up-to-date papers or preprints dedicated to robust network flows such as Matthews et al, 2019, Dokka et al, 2020 or Garuba et al, 2021. All these papers, whether stochastic-based or robust, are focused on finding solutions that are, in a given sense, resilient to changes and secure the realisation of the solution for the given degree of uncertainty.

The approach we discuss in our contribution has a different background. We initially assume that the edge capacities are not explicitly given at all, but have their predefined lower and upper bounds. Such an initial setting is strongly related to interval analysis, which has been applied to network flows by Baykasoğlu and Özbel, 2021 who use the interval linear programming and game theory for assessing a network with multiple owners. The interval-based framework is often used for commodity flows such as gas (Qiao et al, 2017) or electricity (Zhang et al, 2018). Unlike in these interval-based papers, we aim not to search for particular solutions under uncertainty. Instead, we propose a way of setting the efficient edge capacity for every edge within its predefined interval, such that the resulting graph has properties desired by its owner. At the same time, it is considered that building an edge capacity is bounded by unit costs (possibly different for every edge) and the total building costs are constrained by an available budget. The overall objective is designing a network with such edge capacities that would yield the best objectives when we consequently optimise over this network with multiple objectives, and while we desire to minimise the budget needed for its design.

[1] CZU Prague, Department of Systems Engineering, Kamýcká 129, 16500 Praha - Suchdol, hlavaty@pef.czu.cz
[2] CZU Prague, Department of Systems Engineering, Kamýcká 129, 16500 Praha - Suchdol, brozova@pef.czu.cz

2 Methods

Prior to numerical examples in the next section, we present the mathematical framework used for designing the "optimal[1] network" for a minimum budget.

2.1 Network flow with multiple objectives

Let us consider the generalised network flow problem with r objectives formulated as a linear programming problem (1). Given a directed graph $G = (V, E)$, unknown edge capacities $y_{ij} \geq 0$ for each edge $(i, j) \in E$, source vertex $s \in V$ and terminal vertex $t \in V$, the following formulation maximizes the flow from s to t and the flow on the edge $(i, j) \in E$ is x_{ij}. Additional $r - 1$ generic objectives are considered and are maximised (or minimised) over all edges $(i, j) \in E$ for given edge weights w_{ij}^k for every k-th objective.

$$\begin{array}{ll} \max \sum_{j:(s,j)\in E} x_{sj} & \\ \max \sum_{(i,j)\in E} w_{ij}^k x_{ij} & k = 2, \ldots, r \end{array} \quad \text{subject to} \quad \begin{array}{c} \sum_{j\in V} x_{ji} - \sum_{j\in V} x_{ij} = 0 \quad \forall i \in V \setminus \{s, t\} \\ 0 \leq x_{ij} \leq y_{ij} \quad \forall (i, j) \in E \\ l_{ij} \leq y_{ij} \leq u_{ij} \\ u_{ij} \geq l_{ij} \geq 0, \forall (i, j) \in E \end{array} \tag{1}$$

The first objective is separated from the others since the maximum flow is a typical objective present in the most applications, the rest of the r objectives is optional. However, to make our methodology work, we require that $|r| \geq 2$. The first constraint set represents the flow conservation on all vertices except s and t. The second constraint set makes each flow x_{ij} fit the edge capacity y_{ij}. The remaining constraints restrict the edge capacity to the interval $[l_{ij}, u_{ij}]$.

2.2 DeNovo optimization basics

The methodology we propose is inspired by the DeNovo optimisation framework of Zelený, 1986 that was primarily designed for linear programs in production planning. The specific structure of the linear program for network flows does not allow the use of the DeNovo methodology in its original form, but we get inspired by its principles. The DeNovo optimisation problem is formally defined as follows:

$$\max \sum_j c_{jk} x_j, k = 1, \ldots, r \quad \text{subject to} \quad \begin{array}{c} \sum_j a_{ij} x_j \leq y_i, \forall i \\ \sum_i q_i y_i \leq B \\ y_i \geq 0, \forall i \\ x_j \geq 0, \forall j \end{array} \tag{2}$$

The problem (2) has r objectives, the right-hand sides y_i of the first constraint set are variables. The unit cost of every right-hand side y_i is q_i. The interpretation of the problem (2) is to find an efficient setting of all y_i with respect to all r criteria and the given limited budget B. First, the problem (2) needs to be solved as a single-criterion problem for every objective function k. This leads to obtaining r ideal solutions, let these solutions be denoted as $z_k^*, \forall k$. Knowing the ideal solutions, we further construct the following optimisation model:

$$\min \sum_i q_i y_i \quad \text{subject to} \quad \begin{array}{c} \sum_j^n a_{ij} x_j \leq y_i, \forall i \\ \sum_j c_{jk} x_j = z_k^*, k = 1, \ldots, r \\ y_i \geq 0, \forall i \\ x_j \geq 0, \forall j \end{array} \tag{3}$$

If there is an optimal solution to the problem (3), then it is also an efficient solution to the problem (2) that achieves ideal values for all k objectives for a minimal budget. At this point, two issues may occur:

1. the problem (3) does not have a feasible solution.
2. the problem (3) has a feasible solution, but the objective exceeds the available budget B.

If the first issue occurs, then it is necessary to relax the constraint $\sum_j c_{jk} x_j = z_k^*, \forall k$ and instead of trying to achieve ideal values for all k objectives, the constraint is changed to $\sum_j c_{jk} x_j = \hat{z}_k, \forall k$. Let us call the $\hat{z_k}$ achievable values instead of ideal values. The achievable values lie on the Pareto front of the problem (2). To obtain the achievable values, we particularly take advantage of the STEP method of Benayoun et al, 1971. Using the STEP

[1] efficient in the multiobjective sense

method principles, we obtain an efficient solution $\hat{\boldsymbol{z}}(\hat{z}_k)$ that has the minimum weighted distance from the ideal point $\boldsymbol{z}^*(z_k^*)$. A detailed description of adopting this method to the DeNovo framework is shown, e.g., by Hlavatý and Brožová, 2023. If the second issue is encountered, achieving the ideal values of all k objectives is impossible under the current budget. In this case, a concept of *optimum path ratios* introduced by Shi, 1995 should be applied under the classic DeNovo framework to find a solution that respects the budget but is achievable at the same time. Unfortunately, in the case of network flows, due to specific dependence among variables y_i, it is impossible to apply this concept because the edge capacities are not directly proportional to the objectives and the budget, as in the classic DeNovo concept. The linear transformation of the results fails in our case, but our computation experiments showed that shifts in the budget have approximate linear dependence on the edge capacities.

3 Results

In this section, we solve the multiobjective network flow problem with interval-based capacities under the DeNovo framework. We demonstrate our cause on an artificial example and show our computational experiments afterwards.

3.1 Example assignment

Let us consider the following network topology as in the Figure 1. The network is a graph $G(V, E)$ with $|V| = 8, |E| = 15$ a source vertex S, terminal vertex T and all edges have the same (without loss of generality) predefined lower and upper bounds $l_{ij} = 10, \forall(i, j) \in E$ and $u_{ij} = 50, \forall(i, j) \in E$.

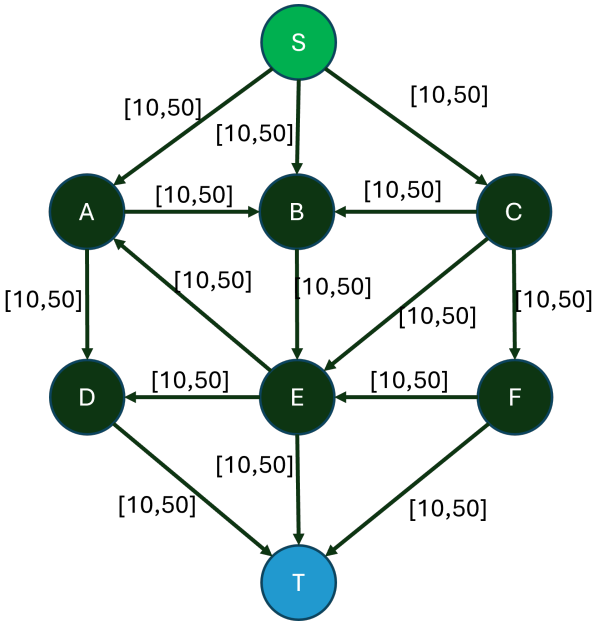


Figure 1 Network topology

Further, we assume two objectives for simplicity, while the first one is maximising the flow from S to T, which is represented by the following objective function:

$$z_1 := \max\ (x_{sa} + x_{sb} + x_{sc}) \tag{4}$$

and the second objective minimising the risk score over all edges:

$$\begin{gathered} z_2 := \min \boldsymbol{w}^\top \boldsymbol{x} \\ \boldsymbol{w}^\top = (2, 6, 10, 8, 5, 10, 1, 10, 6, 4, 7, 10, 1, 1, 5) \\ \boldsymbol{x}^\top = (x_{sa}, x_{sb}, x_{sc}, x_{ab}, x_{ad}, x_{be}, x_{cb}, x_{ce}, x_{cd}, x_{dt}, x_{ea}, x_{ed}, x_{et}, x_{fe}, x_{ft}) \end{gathered} \tag{5}$$

The components of $\boldsymbol{w}(w_{ij})$ were randomly generated over uniform distribution on $[1, 10] \cap \mathbb{Z}$. The goal is to compute appropriate edge capacities within the predefined intervals with respect to both objectives and the constrained budget B:

$$\begin{gathered} \boldsymbol{q}^\top = (64, 1, 16, 4, 67, 18, 67, 37, 61, 17, 65, 14, 13, 67, 63) \\ \boldsymbol{q}^\top \boldsymbol{y} \leq B \end{gathered} \tag{6}$$

The components of $\boldsymbol{q}(q_{ij})$ were randomly generated over uniform distribution on $[1, 100] \cap \mathbb{Z}$.

3.2 Solution procedure

Let us now follow with the step-by-step procedure for finding an efficient network design. Recall the problem (1). First, it is necessary to solve two separate single-criterion problems with the objectives (4) and (5) as follows:

$$\begin{gathered}\max z_1 \\ \text{subject to} \\ \sum_{j\in V} x_{ji} - \sum_{j\in V} x_{ij} = 0 \quad \forall i \in V \setminus \{s,t\} \\ 0 \le x_{ij} \le y_{ij} \quad \forall (i,j) \in E \\ 10 \le y_{ij} \le 50 \\ \sum_{(i,j)\in E} q_{ij} y_{ij} \le B\end{gathered} \tag{7}$$

$$\begin{gathered}\min z_2 \\ \text{subject to} \\ \sum_{j\in V} x_{ji} - \sum_{j\in V} x_{ij} = 0 \quad \forall i \in V \setminus \{s,t\} \\ 0 \le x_{ij} \le y_{ij} \quad \forall (i,j) \in E \\ 10 \le y_{ij} \le 50 \\ \sum_{(i,j)\in E} q_{ij} y_{ij} \le B\end{gathered} \tag{8}$$

The optimal solutions of (7), (8) are z_1^* and z_2^* respectively. The value z_1^* represents the maximum flow from S to T with the edge capacities designed within the available budget B. The z_2^* interprets as the minimum overall network risk associated with the edge capacities designed within the available budget B. In both cases, the flow constraints are secured to make any flow within these capacities feasible. Next, we follow the second stage of the DeNovo framework as in (3):

$$\begin{gathered}B^* = \min \sum_{(i,j)\in E} q_{ij} y_{ij} \\ \text{subject to} \\ z_1 = z_1^* \\ z_2 = z_2^* \\ \sum_{j\in V} x_{ji} - \sum_{j\in V} x_{ij} = 0 \quad \forall i \in V \setminus \{s,t\} \\ 0 \le x_{ij} \le y_{ij} \quad \forall (i,j) \in E \\ 10 \le y_{ij} \le 50\end{gathered} \tag{9}$$

$$\begin{gathered}\hat{B} = \min \sum_{(i,j)\in E} q_{ij} y_{ij} \\ \text{subject to} \\ z_1 = \hat{z}_1 \\ z_2 = \hat{z}_2 \\ \sum_{j\in V} x_{ji} - \sum_{j\in V} x_{ij} = 0 \quad \forall i \in V \setminus \{s,t\} \\ 0 \le x_{ij} \le y_{ij} \quad \forall (i,j) \in E \\ 10 \le y_{ij} \le 50\end{gathered} \tag{10}$$

Problem (9) minimises the budget for building edge capacities while achieving the ideal objective values for both objectives and preserving the logic of the network flow. Alternatively, if the solution to the problem (9) fails to exist, then the problem (10) is solved with a similar interpretation where the ideal objective values are replaced by achievable objective values. By solving either (9) or (10), we obtain an efficient network capacity design respecting both objectives and available budget.

3.3 Experiments and comparisons

Note that in the previous section, specific data were on the input, except for no specific budget was defined. Given the randomly generated edge unit cost vector $\boldsymbol{q}$ in section 3.1, one can determine the minimum and maximum theoretical budget needed to solve the problem, considering the interval bounds for the edge capacities. The theoretical budget ranges in $[5740, 28700] = [10\|\boldsymbol{q}\|_1, 50\|\boldsymbol{q}\|_1]$ if all edges are set to their lower or upper bound, respectively. However, the realistic budget range, for which the results make sense[2] is $[7620, 18540]$. The tighter interval was yielded by the computational experiments and has no trivial computation formula, as it depends on the network topology. We have repeated the procedure described in section 3.2 for different budget levels ranging within the latter interval. To emphasise the difference between our approaches and the more straightforward ways, a comparison has been made for different budget levels. Since the problem in question is a multiobjective optimisation problem, traditional methods were also applied to solve our example: Goal programming (11)and Weighted sum of objectives (12).

[2] for a budget too low, the problem is not feasible for the second objective. For budget too high, useless capacities are being built

$$
\begin{aligned}
&\min \left(\frac{\delta_1}{3} + \frac{\delta_2}{\|\boldsymbol{w}\|_1} \right) \\
&\text{subject to} \\
&x_{sa} + x_{sb} + x_{sc} + \delta_1 = z_1^* \\
&\sum_{(i,j)\in E} w_{ij} x_{ij} - \delta_2 = z_2^* \\
&\sum_{j\in V} x_{ji} - \sum_{j\in V} x_{ij} = 0 \quad \forall i \in V \setminus \{s,t\} \\
&0 \le x_{ij} \le y_{ij} \quad \forall (i,j) \in E \\
&10 \le y_{ij} \le 50 \\
&\sum_{(i,j)\in E} q_{ij} y_{ij} \le B
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
&\max \left(\frac{x_{sa} + x_{sb} + x_{sc}}{3} - \sum_{(i,j)\in E} \frac{w_{ij} x_{ij}}{\|\boldsymbol{w}\|_1} \right) \\
&\text{subject to} \\
&\sum_{j\in V} x_{ji} - \sum_{j\in V} x_{ij} = 0 \quad \forall i \in V \setminus \{s,t\} \\
&0 \le x_{ij} \le y_{ij} \quad \forall (i,j) \in E \\
&10 \le y_{ij} \le 50 \\
&\sum_{(i,j)\in E} q_{ij} y_{ij} \le B
\end{aligned}
\tag{12}
$$

Applying all three approaches (DeNovo, Goal programming and Weighted sum) yields the results in table 1. The approaches in the table are mutually non-dominated, and none is superior to the others in the multiobjective

De Novo			**Goal programming**			**Weighted sum**		
Flow	Risk	Budget	Flow	Risk	Budget	Flow	Risk	Budget
88	1480	9071	82	1480	8000	77	1243	8000
100	1480	11637	87	1480	9000	83	1317	9000
88	1782	8022	92	1480	10000	90	1391	10000
93	1920	8527	97	1480	11000	97	1466	11000
98	1962	9304	104	1540	12000	104	1540	12000
103	2004	10082	110	1619	13000	110	1619	13000
108	2045	10861	118	1769	14000	118	1769	14000
113	2087	11702	125	1919	15000	125	1919	15000
118	2129	12582	132	2069	16000	132	2069	16000
123	2177	13507	139	2219	17000	139	2219	17000
130	2251	14666	146	2369	18000	146	2369	18000
137	2326	15949	150	2450	19000	150	2450	19000

Table 1 Methods comparison

optimisation sense. However, we attempted to compare the results by calculating values of ratio *flow/budget* and product *risk*budget*. These *flow/budget* and *risk*budget* measures were consequently normalised separately to the interval $[0, 1]$ and the normalised values summed for each budget level. The sum of normalised ratios can be considered an index of solution quality that includes the performance of both *flow value* and *risk value* to *budget*. The index takes values between $[0, 2]$, the higher is better. Once again, the indices are mutually non-dominated, but the average index value $\bar{\alpha}$ was calculated for each method as in Table 2.

DeNovo	**Goal programming**	**Weighted sum**
$\bar{\alpha}$=0.93	$\bar{\alpha}$=0.74	$\bar{\alpha}$=0.76

Table 2 Average performance index

4 Conclusion

We have proposed a new method for setting up an efficient network capacity. Different approaches yield generally incomparable values, but our method performed better on average. It needs to be further investigated if the overall performance of the DeNovo approach can yield better average results than other multiobjective techniques.

Acknowledgements

This paper was supported by the Czech Science Foundation (GAČR) project No. 25-15714S.

References

Ahuja RK, Magnanti TL, Orlin JB (1993) Network Flows: Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs, NJ

Aronson JE (1989) A survey of dynamic network flows. Annals of Operations Research 20:1–66. https://doi.org/10.1007/BF02216922

Baykasoğlu A, Özbel BK (2021) Explicit flow-risk allocation for cooperative maximum flow problems under interval uncertainty. Operational Research 21(4):2149–2179. https://doi.org/10.1007/s12351-019-00500-5

Benayoun R, Montgolfier J, Tergny J, et al (1971) Linear programming with multiple objective functions: Step method (stem). Mathematical Programming 1(1):366–375

Cruz-Mejía O, Letchford AN (2023) A survey on exact algorithms for the maximum flow and minimum-cost flow problems. Technical report, Lancaster University Management School

Dokka T, Garuba F, Goerigk M, et al (2020) An efficient approach to distributionally robust network capacity planning. arXiv preprint arXiv:2004.04493 [math.OC]

Finke G (ed) (2008) Operations Research and Networks. John Wiley & Sons, Hoboken, NJ

Ford LR, Fulkerson DR (1956) Maximal flow through a network. Canadian Journal of Mathematics 8:399–404

Garuba F, Goerigk M, Jacko P (2021) A comparison of discrete and polyhedral uncertainty sets for robust network design. arXiv preprint arXiv:2003.10507 [math.OC]

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), Prague, Czech Republic, pp 105–110

Matthews LR, Gounaris CE, Kevrekidis IG (2019) Designing networks with resiliency to edge failures using two-stage robust optimization. European Journal of Operational Research 279(3):704–720. https://doi.org/10.1016/j.ejor.2019.06.021

Mozafari M, Zabihi A (2020) Robust water supply chain network design under uncertainty in capacity. Water Resources Management 34:4093–4112. https://doi.org/10.1007/s11269-020-02658-6

Qiao Z, Guo Q, Sun H, et al (2017) An interval gas flow analysis in natural gas and electricity coupled networks considering the uncertainty of wind power. Applied Energy 201:343–353. https://doi.org/10.1016/j.apenergy.2016.12.020

Shi Y (1995) Studies on optimum-path ratios in multicriteria de novo programming problems. Computers & Mathematics with Applications 29:43–50

Thapalia B, Wallace WS, Kaut M, et al (2012) Single source single-commodity stochastic network design. Computational Management Science 9(1):139–160. https://doi.org/10.1007/s10287-010-0129-0

Zelený M (1986) Optimal system design with multiple criteria: De novo programming approach. Engineering Costs and Production Economics 10:89–94

Zhang C, Chen H, Guo M, et al (2018) Dc power flow analysis incorporating interval input data and network parameters through the optimizing-scenarios method. International Journal of Electrical Power & Energy Systems 96:380–389. https://doi.org/10.1016/j.ijepes.2017.10.005

Received: 2025-06-06 Last revision: 2025-06-06 Accepted: 2025-07-07

Mean-CVaR portfolio choice problem with endogenous randomness

Erik Kočandrle[1]

Abstract. Portfolio optimization problems include a random element corresponding to the losses or returns of various assets. The usual assumption is that the randomness is exogenous, meaning that the distribution of losses can not be influenced by the investor's decisions. However, there may be cases where this assumption is not completely sound, for example when we deal with a combination of a large investor and a company with a small market capitalization. We present a portfolio optimization problem with endogenous randomness, minimizing the CVaR risk measure, while assuming that a large enough investment in an asset can influence its loss distribution and improve performance. After discretization of the investment effects, various partitions of the feasibility set can be constructed based on the investor's budget and asset's market capitalization. We study the possible changes in portfolio composition under different budgets.

Keywords: Portfolio optimization, endogenous randomness, conditional value at risk

JEL Classification: C44, G11
AMS Classification: 90C15

1 Introduction

In order to define problems in the mean-risk portfolio optimization framework, we need to be able to quantify risk. The functions that serve this purpose are called risk measures. One of the most important risk measures is the Conditional Value at Risk, whose properties were studied in Pflug, 2000 for example. Since then it became a very popular choice for its coherency and handling of extreme losses. It was used and studied in the mean-risk portfolio choice context in important papers Rockafellar and Uryasev, 2000 and Rockafellar and Uryasev, 2002. Portfolio optimization problems naturally include a random element in the form of asset losses/returns, which makes them fall into the category of stochastic optimization. In the standard case, the distributions of these losses are assumed to be fixed, since it is unlikely that a small investor could somehow influence the performance of a large company. However, there may be cases where this assumption is not completely valid. Let us imagine, for example, a large investor who chooses to invest in very small companies in hopes of elevating their performance. The effect of such a large investment can have many forms, but in our model, we will assume that a significant enough investment can enhance the expected performance of an asset. We will measure the size of a company by its market capitalization, and an investment will be deemed large, if it covers a significant share of the company stocks. Generally, the type of randomness which depends on decisions, is called endogenous or decision dependent. It is still a relatively new topic, but of rising interest. One of the first major works dedicated to endogenous randomness is Jonsbråten et al, 1998. In the more financial context, we can list notable works in this field, like the pricing problem with subsequent asset-liability bank management model in Kopa and Rusý, 2021, or the paper Kopa and Rusý, 2023, studying the robustness via contamination of stochastic optimization problems with endogenous randomness. This paper is structured as follows. In section 2 we formulate the general shape of a CVaR portfolio choice problem with endogenous randomness, while subsequently defining the specific type of decision dependence. In section 3 we present a numerical example for 8 small company stocks and several investor budgets. Lastly, in section 4 we conclude the results of the paper and provide ideas for further research.

2 Model definition

We will generally measure a stock's performance by its daily losses, but the returns can be simply obtained by changing the signs. In this paper, we will use the well-known risk measure Conditional Value at Risk on the significance level of $\alpha = 0.95$, shortly denoted as CVaR_α. Intuitively, the risk of a stock will be measured by the average of 5% worst possible losses. A portfolio will be considered optimal, if it minimizes the risk while enforcing some minimal requirement on the expected return. A portfolio of $d \in \mathbb{N}$ assets will be understood as

[1] Charles University, Faculty of mathematics and physics, Department of probability and mathematical statistics, Sokolovská 83, 186 75, Prague 8, Czech Republic, kocandrle@karlin.mff.cuni.cz

a vector of weights $\boldsymbol{\lambda} \in \Lambda = \left\{ \boldsymbol{\lambda} \in \mathbb{R}_+^d : \sum_{i=1}^d \lambda_i = 1 \right\}$, where λ_i denotes the fraction of wealth invested into the i-th stock. These weights will serve as decision variables in our model and since we assume the existence of endogenous randomness, their values will have an effect on the loss distribution.

2.1 General formulation of decision dependent randomness

For each feasible investment choice $\boldsymbol{\lambda} \in \Lambda$, the decision dependent asset losses will be stored in the random vector $\boldsymbol{\xi}(\boldsymbol{\lambda}) = (\xi_1(\boldsymbol{\lambda}), ..., \xi_d(\boldsymbol{\lambda}))^T : (\Omega, \mathcal{A}) \to (\mathbb{R}^d, \mathcal{B}^d)$ with a distribution $\mathcal{P}(\boldsymbol{\lambda})$, where $\xi_i(\boldsymbol{\lambda})$ denotes the random daily loss of the i-th asset under portfolio choice $\boldsymbol{\lambda}$. We assume all present variables to have a finite expectation and all present optimization models to have an optimal solution. The daily loss of the whole portfolio can then be computed as $\boldsymbol{\lambda}^T \boldsymbol{\xi}(\boldsymbol{\lambda}) = \sum_{i=1}^d \lambda_i \cdot \xi_i(\boldsymbol{\lambda})$. The general CVaR_α minimization problem with endogenous randomness is then formulated as

$$\begin{aligned} &\min_{\boldsymbol{\lambda} \in \Lambda} \ \text{CVaR}_\alpha\left(\boldsymbol{\lambda}^T \boldsymbol{\xi}(\boldsymbol{\lambda})\right) \\ &\text{s.t.} \ -\mathbb{E}[\boldsymbol{\lambda}^T \boldsymbol{\xi}(\boldsymbol{\lambda})] \geq r_{min}, \end{aligned}$$

or equivalently (as seen in Rockafellar and Uryasev, 2000)

$$\begin{aligned} &\min_{\boldsymbol{\lambda} \in \Lambda, a \in \mathbb{R}} \left(a + \frac{1}{1-\alpha} \cdot \mathbb{E}[\boldsymbol{\lambda}^T \boldsymbol{\xi}(\boldsymbol{\lambda}) - a]^+ \right) \\ &\text{s.t.} \ -\mathbb{E}[\boldsymbol{\lambda}^T \boldsymbol{\xi}(\boldsymbol{\lambda})] \geq r_{min}. \end{aligned}$$

The problem is, we do not know the exact shape of the loss distribution, so we need to estimate it. We will collect $S \in \mathbb{N}$ daily loss observations from the recent past and store them inside a matrix $\mathbb{L} \in \mathbb{R}^{S \times d}$, where the i-th row $\boldsymbol{\xi}^i \in \mathbb{R}^d$ is the i-th observation of the loss vector. This scenario matrix will serve as the basis of operation, and will be used in the case, where no endogenous randomness is present. The goal is now to define the functional dependence $\boldsymbol{\lambda} \mapsto \boldsymbol{\xi}(\boldsymbol{\lambda})$. We assume that the support of each decision dependent loss distribution $\mathcal{P}(\boldsymbol{\lambda})$ consists of transformed scenarios from the matrix $\mathbb{L}$, but under the same empirical probabilities $1/S$. The scenario matrix for a specific portfolio $\boldsymbol{\lambda} \in \Lambda$ will be denoted $\mathbb{L}(\boldsymbol{\lambda}) \in \mathbb{R}^{S \times d}$ and the i-th loss vector scenario $\boldsymbol{\xi}^i(\boldsymbol{\lambda})$. Note that unlike the daily losses, which we use to compute the risk measure, the minimal return parameter r_{min} will represent the expected yearly return, in order to exist on a more intuitive scale. Considering the standard 252 trading days in a year, we can compute the decision dependent vector of annualized expected returns of our stocks as $\boldsymbol{r}(\boldsymbol{\lambda}) = -252 \cdot \frac{1}{S} \sum_{i=1}^S \boldsymbol{\xi}^i(\boldsymbol{\lambda})$. Let us formulate the general model of interest (1) in its linear reformulation.

$$\begin{aligned} &\min_{\boldsymbol{\lambda} \in \Lambda, a \in \mathbb{R}, z_1, ..., z_m \in \mathbb{R}} \left(a + \frac{1}{1-\alpha} \cdot \sum_{i=1}^S \frac{1}{S} \cdot z_i \right) \\ &\text{s.t.} \ \boldsymbol{\lambda}^T \boldsymbol{r}(\boldsymbol{\lambda}) \geq r_{min} \\ &\quad z_i \geq \boldsymbol{\lambda}^T \boldsymbol{\xi}^i(\boldsymbol{\lambda}) - a, \ i = 1, ..., S \\ &\quad z_i \geq 0, \ i = 1, ..., S. \end{aligned} \tag{1}$$

2.2 Definition of decision dependence

Let $M_1, ..., M_d > 0$ be the market capitalizations of the chosen companies at the time of investment and $B > 0$ the investor's budget. It is now time to assess what we consider a significant investment, capable of influencing the loss distribution. We work with six different ownership levels for each asset. For a stock $i \in \{1, ..., d\}$, the ownership level $k_i = 0$ is achieved if the investor obtains less than 10% of the market capitalization. Level $k_i = 1$ is achieved if he obtains between 10% and 20%. Similarly, level $k_i = 2$ is between 20% and 30%, level $k_i = 3$ between 30% and 40%, level $k_i = 4$ between 40% and 50% and lastly, level $k_i = 5$ if the investor obtains more than 50% of the stock's market capitalization.

Let us now look at an arbitrary asset $i \in \{1, ..., d\}$ and its portfolio weight $\lambda_i \in [0, 1]$. We know the present market capitalization M_i and the budget B. We need to compute the thresholds for these ownership levels in terms of portfolio weights. The amount of funds invested into this asset is equal to $\lambda_i \cdot B$. The base threshold will always be considered as $T_i^0 = 0$, meaning that nothing was purchased. The weight λ_i^* needed to hit the 10% threshold solves the equation

$$\lambda_i^* \cdot B = 0.1 \cdot M_i,$$

so we can compute the threshold as $T_i^1 = 0.1 \cdot \frac{M_i}{B}$.

Analogically, for the 20% threshold we have $T_i^2 = 0.2 \cdot \frac{M_i}{B}$, then $T_i^3 = 0.3 \cdot \frac{M_i}{B}$, $T_i^4 = 0.4 \cdot \frac{M_i}{B}$, $T_i^5 = 0.5 \cdot \frac{M_i}{B}$. Lastly, $T_i^6 = \frac{M_i}{B}$ is the weight needed to buy all the stocks. For simplicity, we will assume that $T_i^6 > 1 \; \forall i = 1, ..., d$, i.e. none of the market capitalizations can be completely covered by the investor's budget. The reason for this assumption will be explained a little later on.

To summarize, each of the d assets generates six possible ownership levels, which can be indexed by elements of the set $\{0, 1, 2, 3, 4, 5\}$. The total number of ownership combinations for the whole portfolio is 6^d, and we can represent them using a set $K = \{(k_1, ..., k_d)^T : k_i \in \{0, 1, 2, 3, 4, 5\} \; \forall i\}$. Each element of this set is a vector of ownership levels $\boldsymbol{k} = (k_1, ..., k_d)^T$. We assume that the effect of investment on asset performances can be discretized in this manner. For each asset, passing over a non-zero ownership threshold will trigger a change in the loss scenarios. This implies that we do not have to define the decision dependent loss distribution for each $\boldsymbol{\lambda} \in \Lambda$ separately, but we can divide the set Λ to several subsets, where all portfolios in a subset share the same loss distribution. For each ownership combination $\boldsymbol{k} \in K$, we denote the set of portfolios satisfying this combination as 2.

$$\Lambda(\boldsymbol{k}) = \Lambda(k_1, ..., k_d) = \left\{ \boldsymbol{\lambda} \in \Lambda : \lambda_i \in [T_i^{k_i}, T_i^{k_i+1}) \; \forall i \in \{1, ..., d\} \right\}. \tag{2}$$

Under the previously mentioned assumption that $T_i^6 > 1 \; \forall i = 1, ..., d$, we can see that these sets form a disjoint partition of the set of all portfolios Λ. In other words, $\cup_{\boldsymbol{k} \in K} \Lambda(\boldsymbol{k}) = \Lambda$ and $\Lambda(\boldsymbol{k}_1) \cap \Lambda(\boldsymbol{k}_2) = \emptyset$ for each $\boldsymbol{k}_1, \boldsymbol{k}_2 \in K$. On each of these sets, we define a common random loss vector $\boldsymbol{\xi}(\boldsymbol{k})$ with an empirical probability distribution $\mathcal{P}(\boldsymbol{k})$ on a scenario set $\mathbb{L}(\boldsymbol{k})$. Big advantage of this approach is, that the whole optimization problem with endogenous randomness can be split into smaller subproblems with exogenous randomness. To unify the notation with the general problem definition 1, we conclude that for each ownership combination $\boldsymbol{k} \in K$ it holds that $\boldsymbol{\xi}(\boldsymbol{\lambda}) \overset{\text{a.s.}}{=} \boldsymbol{\xi}(\boldsymbol{k})$ and $\mathbb{L}(\boldsymbol{\lambda}) = \mathbb{L}(\boldsymbol{k})$ for all $\boldsymbol{\lambda} \in \Lambda(\boldsymbol{k})$.

It remains to define the exact form of the decision dependent scenario sets $\mathbb{L}(\boldsymbol{k})$. Remember that $\mathbb{L} \in \mathbb{R}^{S \times d}$ is the scenario matrix observed from data, which is used in the case where no change in scenarios occurs. This corresponds to the situation where the investor owns less than 10% of each market capitalization, i.e. $k_i = 0 \; \forall i \in \{1, ..., d\}$. The elements of the base matrix $\mathbb{L}$ are denoted $\mathbb{L}_{j,i} = \xi_i^j$, i.e. the j-th loss scenario of i-th stock. To define how the scenarios change with the parameter $\boldsymbol{k} \in K$, we need to define some quantities. For each asset $i \in \{1, ..., d\}$ we define the *translation factor* as the absolute value of the mean daily loss, i.e.

$$\tau_i = \left| \frac{1}{S} \cdot \sum_{j=1}^{S} \mathbb{L}_{j,i} \right| = \left| \frac{1}{S} \cdot \sum_{j=1}^{S} \xi_i^j \right|.$$

The second quantity of interest will be the *influence factor* δ, which specifies the strength of the effect of large investments on loss scenarios. A significant investment into an asset $i \in \{1, ..., d\}$ will lower all the loss scenarios by subtracting some percentage of the translation factor, while the percentage depends on the ownership level k_i and the influence factor δ. In other words, for each ownership combination $\boldsymbol{k} = (k_1, ..., k_d)^T \in K$, the matrix of loss scenarios is defined as

$$\mathbb{L}(\boldsymbol{k}) = \begin{bmatrix} \mathbb{L}_{1,1} - \tau_1 \cdot (\delta \cdot k_1) & \cdots & \mathbb{L}_{1,d} - \tau_d \cdot (\delta \cdot k_d) \\ \mathbb{L}_{2,1} - \tau_1 \cdot (\delta \cdot k_1) & \cdots & \mathbb{L}_{2,d} - \tau_d \cdot (\delta \cdot k_d) \\ \vdots & \ddots & \vdots \\ \mathbb{L}_{S-1,1} - \tau_1 \cdot (\delta \cdot k_1) & \cdots & \mathbb{L}_{S-1,d} - \tau_d \cdot (\delta \cdot k_d) \\ \mathbb{L}_{S,1} - \tau_1 \cdot (\delta \cdot k_1) & \cdots & \mathbb{L}_{S,d} - \tau_d \cdot (\delta \cdot k_d) \end{bmatrix}.$$

This essentially means that by reaching an ownership level k_i for an asset i, its expected return is improved by a multiplicative factor of $(1 + \delta \cdot k_i)$. The translation factors are defined using absolute values on purpose, so that the investment effect is always positive. For example, if we consider an asset i and an influence factor $\delta = 0.05$, owning less than 10% of the market capitalization does not change its performance, obtaining 10% improves the return by 5%, obtaining 20% improves the return by $2 \cdot 5\% = 10\%$ and so on.

2.3 Final model formulation

As we mentioned earlier, the full problem with endogenous randomness can be split into $|K| = 6^d$ subproblems with exogenous randomness. After solving all of them, we simply choose the one with the lowest optimal value. The number of subproblems can be very large but we have to keep in mind that a majority of them will likely

be infeasible. Let us denote the set of combinations, which generate a non-empty set of possible portfolios, by $K^* = \{\boldsymbol{k} \in K : \Lambda(\boldsymbol{k}) \neq \emptyset\}$. This means that if $\boldsymbol{k} \in K^*$, the funds can be distributed in a way that satisfies the constraints 2. We can provide a simple rule to cut out combinations which are sure to be infeasible, without the need to solve their corresponding subproblems. For each combination $\boldsymbol{k} = (k_1, ..., k_d)^T \in K$, there are two possible situations which make satisfying the constraints (2) impossible.

- The investor does not have enough funds to satisfy the constraints (2). This happens when $\sum_{i=1}^{d} T_i^{k_i} > 1$, i.e. even if he buys the minimal allowed amount of each asset, the budget is still insufficient.
- The investor has too much funds to satisfy the constraints (2). This happens when $\sum_{i=1}^{d} T_i^{k_i+1} \leq 1$, i.e. even if he buys the maximal allowed amount of each asset, it still does not cover the whole budget. The $\leq$ sign is there because the feasibility intervals from 2 are open from the right, since obtaining the exact upper threshold $T_i^{k_i+1}$ would trigger another scenario change.

To summarize, if $\boldsymbol{k} \in K^*$, both the conditions (3) must hold. Note that K^* only ensures feasibility with respect to the weight constraints, the corresponding subproblem can still be infeasible because of the minimal return constraint.

$$\begin{aligned} \sum_{i=1}^{d} T_i^{k_i} &\leq 1, \\ \sum_{i=1}^{d} T_i^{k_i+1} &> 1. \end{aligned} \tag{3}$$

We are now prepared to define the final model, which is solved by iterating through all the possibly feasible combinations $\boldsymbol{k} \in K^*$, solving the subproblems with fixed scenario vectors $\boldsymbol{\xi}^i(\boldsymbol{k})$, $i = 1, ..., S$ and choosing the best one overall. The full model (4) can then be formulated with explicitly written constraints as

$$\begin{aligned} \min_{\boldsymbol{k} \in K^*} \Bigg[\min_{\boldsymbol{\lambda} \in \mathbb{R}^d, a, z_1, ..., z_S \in \mathbb{R}} \Bigg(a + \frac{1}{1-\alpha} \cdot \sum_{i=1}^{S} \frac{1}{S} \cdot z_i \Bigg) \Bigg] \\ \text{s.t. } \boldsymbol{\lambda}^T \boldsymbol{r}(\boldsymbol{k}) \geq r_{min} \\ \sum_{i=1}^{d} \lambda_i = 1 \\ \lambda_i \geq 0, \ i = 1, ..., d \\ \lambda_i \in [T_i^{k_i}, T_i^{k_i+1}), \ i = 1, ..., d \\ z_i \geq \boldsymbol{\lambda}^T \boldsymbol{\xi}^i(\boldsymbol{k}) - a, \ i = 1, ..., S \\ z_i \geq 0, \ i = 1, ..., S. \end{aligned} \tag{4}$$

3 Numerical example

We worked with $d = 8$ stocks of small cap companies chosen from the Yahoo stock screener. We gathered daily loss data from the period between January 1st 2022 and December 31st 2024, yielding 751 observations in total, which should be enough for $\text{CVaR}_{0.95}$ estimation. The underlying information are summarized in Table 1 (from now on the assets will be ordered from left to right according to their expected return). A highly important parameter in our

Metric	EARN	LARK	SPE	SFBC	SBFG	ISSC	GOTRF	CAAS
Exp. yearly return	0.0115	0.0406	0.0461	0.1015	0.1170	0.1962	0.2143	0.3957
$\text{CVaR}_{0.95}$ of daily losses	0.0448	0.0520	0.0337	0.0311	0.0439	0.0607	0.0879	0.0870
Market cap. [mil. USD]	191.9820	142.9630	165.2690	126.0820	140.0640	188.2530	160.6890	123.4600

Table 1 Selected stocks information

model is the investor's budget B, and while it is not directly written in the model formulation 4, it is hidden inside the threshold values T_i^j. The larger the budget, the more ownership combinations become feasible. We worked with a grid of five different budgets, 1, 30, 60, 90 and 120 mil. USD. The first budget is chosen rather small on purpose, in order to not be able to purchase a significant share of any stock (so $\boldsymbol{k} = \boldsymbol{0}$ is the only feasible combination). Consequently, the generated problem is then a classical $\text{CVaR}_{0.95}$ portfolio optimization problem with exogenous randomness. The last budget is chosen in a way, that it is possible to obtain the largest ownership level for some asset. We computed tables of optimal portfolios on a grid of feasible yearly returns and we presented them for the lowest and the highest budget. Note that the problem with higher budget has more feasible returns, since the asset

performance can be improved by activations of endogenous randomness. We will use color differentiation inside the tables, where each colorful weight signifies that some threshold has been crossed, yielding a higher ownership level. The color scheme can be assessed from the legends above the tables.

We will operate with the influence factor of $\delta = 0.05$, meaning that passing to a higher ownership level improves the performance by additional 5% of the base performance. The case with the lowest budget can be seen in Table 2. See that there are no colors, since it is a problem with exogenous randomness. In Table 3 we can see the optimal portfolios for the highest budget. The general trend to note is, that since we can improve the performance of our assets by substantial investments, we can invest a little less into the risky and highly profitable assets, and substitute them by less volatile and less profitable ones, while covering the same minimal return requirement. The fourth asset SFBC has the lowest base risk, and we can see that when we are able to improve its return, we invest into it a lot more. Naturally, the weight of the most profitable asset CAAS has to grow with increasing return requirement, but it grows slower when we have a sufficient budget to improve the return of less risky assets. Table 3 also includes a strange phenomenon, where some weights occur several times, which would be very unusual in the classical framework. However, we noticed that these repeating values are actually the exact ownership thresholds, meaning that the investor sometimes buys the least possible amount of an asset which still allows him to trigger the performance enhancement.

0 - 10%	10 - 20%	20 - 30%	30 - 40%	40 - 50%	50 - 100%

r_{min} [%]	EARN	LARK	SPE	SFBC	SBFG	ISSC	GOTRF	CAAS	$CVaR_{0.95}$ [%]
10.0	0.0712	0.1236	0.2045	0.2963	0.1365	0.0849	0.0308	0.0523	1.6898
12.5	0.0203	0.0837	0.1972	0.3267	0.1080	0.1359	0.0602	0.0681	1.7602
15.0	0	0.0763	0.1131	0.3625	0.0854	0.1688	0.0822	0.1117	1.9930
17.5	0	0.0580	0.0553	0.3705	0.0649	0.1798	0.0982	0.1734	2.3285
20.0	0	0.0114	0	0.3609	0.0678	0.2265	0.1179	0.2155	2.7064
22.5	0	0	0	0.2998	0.0218	0.2422	0.1550	0.2813	3.1548
25.0	0	0	0	0.1463	0.0491	0.2997	0.1609	0.3440	3.6701
27.5	0	0	0	0.1253	0	0.2957	0.1370	0.4420	4.2371
30.0	0	0	0	0	0	0.3195	0.1762	0.5043	4.8519
32.5	0	0	0	0	0	0.2098	0.1590	0.6312	5.6348
35.0	0	0	0	0	0	0.1117	0.1291	0.7592	6.5931
37.5	0	0	0	0	0	0.0337	0.0771	0.8893	7.6601

Table 2 Optimal portfolios with a budget of 1 mil. $ and an influence factor 0.05

0 - 10%	10 - 20%	20 - 30%	30 - 40%	40 - 50%	50 - 100%

r_{min} [%]	EARN	LARK	SPE	SFBC	SBFG	ISSC	GOTRF	CAAS	$CVaR_{0.95}$ [%]
10.0	0.0712	0.1236	0.2045	0.2963	0.1365	0.0849	0.0308	0.0523	1.6880
12.5	0.0377	0.0902	0.1933	0.3243	0.1167	0.1253	0.0533	0.0591	1.7288
15.0	0	0.0805	0.1694	0.3564	0.0699	0.1569	0.0640	0.1029	1.8907
17.5	0	0.0395	0.0784	0.4203	0.0692	0.1636	0.0912	0.1378	2.1585
20.0	0	0.0447	0.0631	0.4203	0.0107	0.1736	0.0818	0.2058	2.4492
22.5	0	0	0	0.4203	0	0.2197	0.1339	0.2262	2.7800
25.0	0	0.0044	0.0275	0.3392	0	0.1863	0.1339	0.3086	3.1800
27.5	0	0	0	0.2160	0	0.3138	0.1339	0.3363	3.5664
30.0	0	0	0	0.2133	0	0.2377	0.1339	0.4150	3.9160
32.5	0	0	0	0.0941	0	0.3138	0.1339	0.4583	4.3717
35.0	0	0	0	0.0796	0	0.3138	0.0922	0.5144	4.6908
37.5	0	0	0	0	0	0.3301	0.0985	0.5715	5.1904
40.0	0	0	0	0	0	0.1980	0.1390	0.6630	5.8214
42.5	0	0	0	0	0	0.1681	0.0753	0.7566	6.5148
45.0	0	0	0	0	0	0.0787	0.0754	0.8459	7.2385
47.5	0	0	0	0	0	0.0143	0.0548	0.9309	7.9813

Table 3 Optimal portfolios with a budget of 120 mil. $ and an influence factor 0.05

Another thing of interest is the efficient frontier, which is the curve of all efficient portfolios, signifying different optimal trade-offs between risk and return. The lowest budget frontier should have the standard non-decreasing concave shape, since it corresponds to a classical mean-CVaR problem with exogenous randomness. However, in Figure 1 we can see that the higher budget frontiers introduce a sort of inflection points, corresponding to the weights dynamically crossing ownership thresholds. For the largest budget, we can see that the jumps in the frontier correspond to color changes inside Table 3.

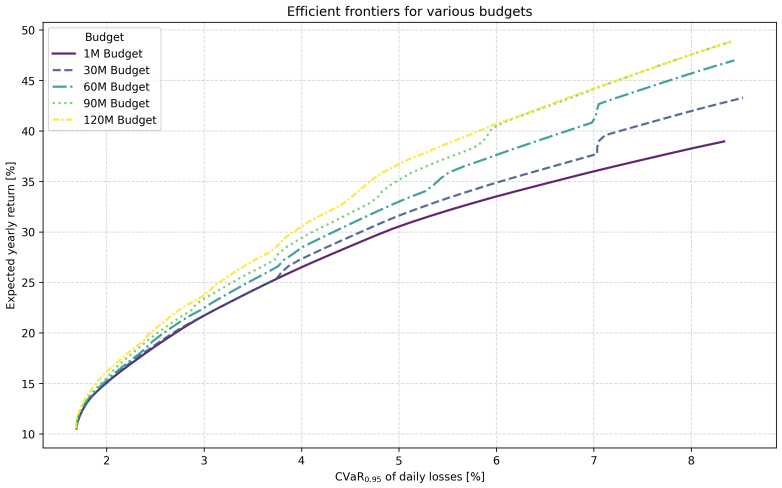


Figure 1 Efficient frontiers for an influence factor of 0.05

4 Conclusion

We presented a mean-CVaR portfolio choice problem, where an investor with a large enough budget can influence the distribution of stock losses/returns, based on the obtained share of market capitalization. From the presented tables and figure, we can safely conclude that introducing this type of endogenous randomness can notably change the portfolio structure. The presented model is adaptive to budget and for a small investor it generates a standard problem with fixed distributions. Gradually increasing the budget allows for more possible ownership combinations, therefore more and larger changes to the loss distributions. The general effect is, that since we are able to get a higher expected return from a stock by investing a significant share of funds into it, we can give larger weights to stocks which are less risky and less profitable, while still covering the same minimal return requirement. The discrete jump structure of the decision dependence we used causes the efficient frontier to lose its concavity property and the inflection points correspond to jumps in the loss scenarios after crossing an ownership threshold. After testing a few other influence effects and different segmentations of the market capitalization, we have seen that the principles demonstrated in this paper do not change. The segmentation choice determines the frequency of the frontier jumps, while the influence factor controls their magnitude. Possible continuation of this work could be to use endogenous randomness in a multistage setting, like the asset-liability management problem in Vitali et al, 2017, or to utilize a different risk measure for the mean-risk setting.

Acknowledgements

This work has been supported by the Grant Agency of Charles University (GA UK) under a project number 362225.

References

Jonsbråten TW, Wets RJB, Woodruff DL (1998) A class of stochastic programs with decision dependent random elements. Annals of Operations Research 82(0):83–106

Kopa M, Rusý T (2021) A decision-dependent randomness stochastic program for asset-liability management model with a pricing decision. Annals of Operations Research 299(1):241–271

Kopa M, Rusý T (2023) Robustness of stochastic programs with endogenous randomness via contamination. European Journal of Operational Research 305(3):1259–1272

Pflug G (2000) Some remarks on the value-at-risk and the conditional value-at-risk. Probabilistic constrained optimization: Methodology and applications pp 272–281

Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. Journal of risk 2:21–42

Rockafellar RT, Uryasev S (2002) Conditional value-at-risk for general loss distributions. Journal of banking & finance 26(7):1443–1471

Vitali S, Moriggia V, Kopa M (2017) Optimal pension fund composition for an italian private pension plan sponsor. Computational Management Science 14:135–160

Received: 2025-05-20 Last revision: 2025-07-01 Accepted: 2025-07-03

Cluster Patterns and Coopetition Potential Among EU Countries: An Analysis Based on the Travel and Tourism Development Index

Petra Vašaničová[1], Martina Košíková[2]

Abstract. This study examines the potential for cross-border coopetition—simultaneous cooperation and competition—among European Union (EU) countries in the travel and tourism sector, using cluster analysis based on data from the Travel and Tourism Development Index (TTDI). Specifically, it focuses on 14 indicators within the Travel and Tourism (T&T) Resources dimension for 27 EU member states across three years: 2019, 2021, and 2024. The primary aim is to determine whether countries consistently belong to the same clusters over time or shift between them, which would have implications for formulating coordinated tourism policies. Cluster analysis was conducted using Ward's method, which produces clusters of similar size and shape by maximizing intra-cluster homogeneity. The number of clusters was determined using the Duda-Hart Je(2)/Je(1) index and the pseudo-T^2 stopping rule. A dendrogram was employed to visually represent the hierarchical structure of clusters. In all three years, four distinct clusters were identified. Notably, only four countries changed their cluster affiliation over the analyzed period: the Netherlands in 2019, Cyprus in 2021, and Belgium and Croatia in 2024. The overall stability in cluster membership suggests a promising foundation for regional cooperation in travel and tourism development across the EU, even among countries that traditionally compete for similar tourist markets.

Keywords: travel and tourism development index, cluster analysis, european union, tourism resources, coopetition, competition, cooperation

JEL Classification: L83, C38
AMS Classification: 62H30

1 Introduction

Coopetition, defined as the simultaneous interplay of cooperation and competition among actors, forms a dynamic network in which entities collaborate while also pursuing their own competitive goals (Chim-Miki and Batista-Canino, 2017b). This duality operates within a structure where participants leverage complementary resources, enhancing the mutual value of their interactions (Lorgnier and Su, 2014). In the context of tourism, coopetition can be particularly beneficial, as it helps mitigate competitive risks by capitalizing on the synergies derived from collaborative strategies (Chim-Miki and Batista-Canino, 2017a). Tourism destinations, due to their complex structures and reliance on shared resources, are frequently cited as environments well-suited to the development of such coopetitive relationships (Della Corte and Sciarelli, 2012). Coopetition is a natural characteristic of tourism destinations which, when strategically managed, can serve as a source of competitive advantage and contribute to the creation of social value (Chim-Miki et al., 2024).

The aim of this paper is to explore the potential for cross-border coopetition among European Union (EU) member states in the travel and tourism sector by identifying patterns of similarity through cluster analysis. By analyzing 14 indicators from the T&T Resources dimension of the Travel and Tourism Development Index (TTDI) across the years 2019, 2021, and 2024, the study seeks to determine the consistency of cluster membership over time and its implications for formulating coordinated and regionally integrated tourism policies.

[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, martina.kosikova@unipo.sk.

2 Material and Methods

2.1 Data

The research sample consists of 27 EU countries. Data used for the analysis is obtained from the TTDI 2024 (World Economic Forum, 2024) and cover three periods: 2019, 2021, and 2024. The TTDI is a versatile tool that enables policymakers and industry professionals to assess the effectiveness of their travel and tourism strategies, management practices, and long-term planning. It takes into account a range of factors that influence the resilience and sustainability of travel destinations and markets. By leveraging the insights provided by the index, public and private sector leaders can inform the creation of innovative policies, practices, and solutions. This helps ensure that the benefits of travel and tourism are harnessed responsibly, preserving the diverse assets that underpin the sector. Ultimately, the TTDI can play a crucial role in promoting inclusive global economic growth, enhancing cultural exchange, and creating employment opportunities. The TTDI comprises 17 pillars and 102 indicators, which are distributed across these pillars.

In the present analysis, coopetition network structures were proposed based on a cluster analysis of 14 indicators drawn from pillars 12 (Natural Resources), 13 (Cultural Resources), and 14 (Non-Leisure Resources), collectively referred to as the T&T Resources dimension of the TTDI. Each of the three pillars within the T&T Resources dimension contributes to cross-border coopetition in distinct ways, as outlined below along with the specific indicators used to assess them:

- Natural Resources (pillar 12): A country's natural resources can act as catalysts for cross-border coopetition by encouraging collaborative resource management, coordinated environmental protection across borders, joint infrastructure projects, and the promotion of transnational tourism. The indicators used to assess this pillar include the Number of World Heritage natural sites (P12.1), Total known species (P12.2), Total protected areas (P12.3), Natural tourism Digital Demand (P12.4), and Number of terrestrial and freshwater ecoregions (P12.5).
- Cultural Resources (pillar 13): A country's cultural resources are essential in fostering cross-border coopetition by encouraging cultural exchange and mutual understanding, promoting heritage tourism, facilitating international education and research, and supporting collaboration in the creative industries. The indicators used to assess this pillar include the Number of World Heritage cultural sites (P13.1), Oral and intangible cultural heritage (P13.2), Number of large sports stadiums (P13.3), Cultural and entertainment tourism Digital Demand (P13.4), and Number of UNESCO Creative Cities (P13.5).
- Non-Leisure Resources (pillar 14): Non-leisure resources contribute to cross-border coopetition by attracting business and professional travel through the presence of global cities, multinational corporations, and world-class academic institutions. The indicators used to assess this pillar include the Presence of Forbes Global 2000 HQ locations (P14.1), Presence of global cities (P14.2), Number of top universities (P14.3), and Non-leisure tourism Digital Demand (P14.4).

Each indicator is scored on a scale ranging from 1 (lowest) to 7 (highest), representing the performance level of a given country. These scores are derived through normalization processes—typically using min-max scaling—to ensure comparability across countries and indicators with different original units and ranges.

2.2 Cluster Analysis

From a methodological perspective, this study employs cluster analysis to reduce the dimensional complexity of the data by combining multiple variables into a single representation, indicating each object's membership within a specific group. A hierarchical clustering approach is applied, which organizes the data into a nested structure of groups and subgroups, where each subgroup can belong to a higher-order group (Vašaničová 2025). Specifically, Ward's method is utilized, which tends to produce clusters of similar size and shape by maximizing homogeneity within each cluster. The results are visually represented using a dendrogram (hierarchical tree diagram).

To determine the optimal number of clusters, the analysis applies a stopping rule based on the Duda-Hart $Je(2)/Je(1)$ index in conjunction with the pseudo-T^2 statistics. These statistical measures offer a formal basis for determining the point at which further merging of clusters results in minimal improvement. $Je(1)$ represents the within-cluster sum of squares for the initial cluster, while $Je(2)$ reflects the combined within-cluster sum of squares after it is divided into two sub-clusters. The $Je(2)/Je(1)$ ratio indicates the relative increase in within-cluster variability between clustering steps, and the pseudo-T^2 statistic assesses the distinctiveness of the clusters being merged. The ideal cluster solution corresponds to the highest Duda-Hart index value combined with the lowest pseudo-T^2 value, indicating the most distinct cluster separation (Duda et al., 2000; Vašaničová 2025). All analyses are conducted using Stata software.

3 Results

The correlation analysis among the selected indicators revealed no problematic multicollinearity. While some moderate to strong correlations exist (e.g., values between 0.6 and 0.8), none of the variables exhibit extremely high correlation (above 0.9) that would suggest redundancy or distort the clustering process. The highest observed correlation is 0.94 between Number of large sports stadiums (P13.3) and Cultural and entertainment tourism Digital Demand (P13.4), both under the Cultural Resources pillar. Although strongly correlated, these indicators capture complementary but non-redundant aspects of cultural tourism infrastructure and visitor interest. This level of correlation is generally considered acceptable in cluster analysis, as it reflects shared but non-identical contributions to the overall data structure. Therefore, all selected variables were retained for clustering, as they offer complementary information and contribute to a more nuanced segmentation of EU member states.

Due to the formation of several smaller clusters through Ward's method, the final number of clusters was determined using the Duda-Hart *Je*(2)/*Je*(1) index. The analysis evaluated between one and fifteen potential groupings, with the optimal solution corresponding to the highest value of the index combined with the lowest pseudo-T^2 statistic. Table 1 displays the outcomes of the stopping rule, indicating that a four-cluster solution was selected.

	Number of clusters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2019	***Je*(2)/*Je*(1)**	0.478	0.679	0.635	**0.578**	0.665	0.441	0.557	0.654	0.603	0.242	0.000	0.094	0.375	0.541	0.000
	pseudo-T^2	27.3	9.94	5.75	**2.92**	4.55	2.54	3.18	1.58	3.3	3.14	.	19.4	1.66	2.55	.
2021	***Je*(2)/*Je*(1)**	0.480	0.696	0.660	**0.630**	0.563	0.303	0.415	0.518	0.316	0.181	0.417	0.398	0.231	0.597	0.000
	pseudo-T^2	27.1	9.16	5.66	**2.93**	6.21	4.6	2.82	3.73	2.17	9.07	2.8	1.52	3.33	2.7	.
2024	***Je*(2)/*Je*(1)**	0.486	0.662	0.650	**0.706**	0.597	0.537	0.277	0.644	0.497	0.189	0.432	0.000	0.434	0.000	0.000
	pseudo-T^2	26.5	10.7	4.31	**4.59**	4.73	2.59	5.22	3.32	3.03	8.58	2.63	.	1.3	.	.

Table 1 Duda-Hart Index and Pseudo-T^2
Note: Values in bold indicate the most significant breakpoint.

Figure 1 presents the dendrograms generated for the years 2019, 2021, and 2024. The dendrograms have been enhanced with labeled boxes to improve the clarity of country groupings within each cluster. Notably, only four countries changed their cluster affiliation over the analyzed period: the Netherlands in 2019, Cyprus in 2021, and Belgium and Croatia in 2024.

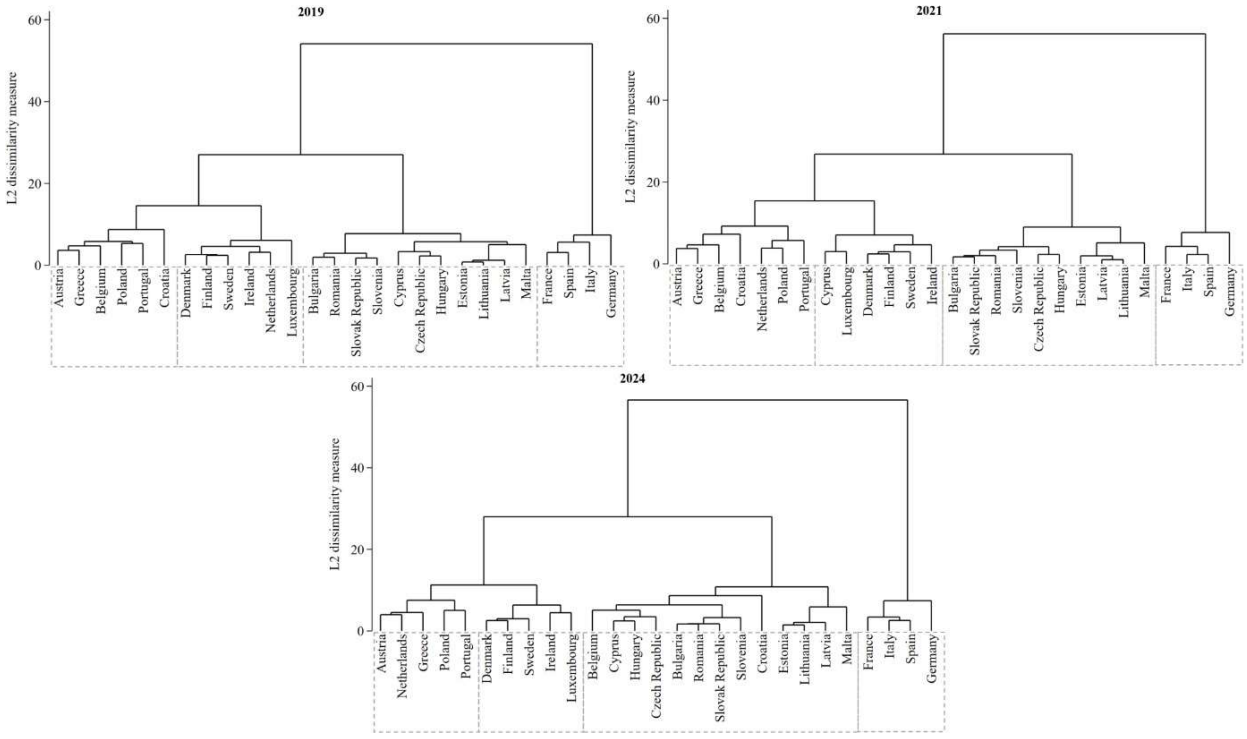


Figure 1 Dendrograms for Analyzed Years

The identified clusters can be interpreted based on the cluster centroids, which indicate the average value of the corresponding pillar within each group (see Table 2).

Cluster 1: Austria, Greece, Poland, Portugal, Netherlands (exclude 2019), Belgium (exclude 2024), Croatia (exclude 2024). Cluster 1 consists of countries with moderate development in both cultural and business tourism, but with relatively limited natural resource appeal. These countries are balanced performers but don't stand out in natural tourism capacity. Cluster 1 demonstrates its strongest performance in the domain of cultural and non-leisure resources, particularly in indicators such as oral and intangible cultural heritage (P13.2), cultural tourism digital demand (P13.4), and business-related infrastructure (P14.2 and P14.3). Conversely, the weakest performance is observed in the natural resources dimension, with particularly low scores in total known species (P12.2) and the number of ecoregions (P12.5).

Cluster 2: Denmark, Finland, Ireland, Luxembourg, Sweden, Cyprus (in 2021), Netherlands (in 2019). This cluster includes innovation-driven economies that excel in business infrastructure but have limited appeal in cultural and natural tourism. These are likely to attract business travelers more than leisure tourists. Cluster 2 exhibits its strongest results in the non-leisure travel dimension, especially in the presence of Forbes Global 2000 headquarters (P14.1), highlighting its attractiveness for business and professional travel. The weakest indicators in this cluster are associated with cultural and natural assets, including low values for total known species (P12.2) and oral and intangible cultural heritage (P13.2), indicating limited appeal in these areas.

Cluster 3: Bulgaria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Romania, Slovak Republic, Slovenia, Cyprus (exclude 2021), Belgium (in 2024), Croatia (in 2024). This cluster includes developing or emerging tourism markets in Central and Eastern Europe. They score consistently lower across all three dimensions, indicating room for growth in tourism infrastructure and branding. Cluster 3 shows relatively stronger performance in select cultural indicators, such as intangible heritage (P13.2), though the overall scores remain modest across all dimensions. This cluster records its lowest scores across a wide range of indicators, particularly in sports and entertainment infrastructure (P13.3), creative city designation (P13.5), and natural tourism digital demand (P12.4), suggesting underdeveloped tourism assets.

Cluster 4: France, Germany, Italy, Spain. These leading tourism economies have rich natural and cultural resources and perform strongly across all dimensions, especially in cultural heritage. They are the most mature and diversified tourism destinations in the group. Cluster 4 stands out with the highest performance across the cultural and natural resources dimensions, with particularly strong results in World Heritage cultural sites (P13.1), intangible cultural heritage (P13.2), and known species (P12.2), reflecting the rich heritage and biodiversity of these countries. While still relatively strong, the weakest area for this cluster lies in non-leisure tourism, especially in the presence of global headquarters (P14.1), which shows comparatively lower values than other indicators.

							2019							
Cluster	**P12.1**	**P12.2**	**P12.3**	**P12.4**	**P12.5**	**P13.1**	**P13.2**	**P13.3**	**P13.4**	**P13.5**	**P14.1**	**P14.2**	**P14.3**	**P14.4**
1	2.167	1.862	4.398	3.792	2.075	3.467	4.200	2.100	3.022	2.222	3.511	3.554	3.487	2.064
2	2.083	1.157	3.631	2.301	1.650	2.283	1.333	1.900	2.302	1.778	6.584	3.087	3.214	1.932
3	1.909	1.242	3.361	1.542	1.668	1.964	2.782	1.218	1.486	1.545	1.456	2.285	2.397	1.565
4	5.375	2.744	5.543	4.765	3.400	7.000	5.300	6.400	6.655	5.167	3.556	6.199	5.364	4.892
							2021							
1	2.286	1.771	4.270	4.217	2.007	3.629	4.257	2.200	2.931	2.524	3.636	3.705	3.598	1.950
2	1.917	1.132	3.535	2.568	1.600	2.083	1.800	1.633	1.910	1.889	6.542	3.104	2.835	1.935
3	2.100	1.258	3.479	1.740	1.705	2.120	2.920	1.220	1.452	1.733	1.266	2.171	2.439	1.631
4	5.750	2.744	5.609	5.826	3.400	7.000	5.900	6.400	6.416	6.167	3.546	6.236	5.364	4.070
							2024							
1	2.200	1.938	4.698	3.899	2.080	3.880	4.040	2.480	3.002	3.533	4.503	3.872	3.917	2.292
2	2.100	1.208	3.637	2.724	1.660	2.220	2.040	1.720	1.946	2.200	6.774	3.116	2.983	2.002
3	2.077	1.353	3.414	2.149	1.692	2.308	4.015	1.262	1.630	1.821	1.563	2.410	2.446	1.509
4	6.000	2.820	5.620	5.697	3.400	7.000	6.200	6.400	6.397	6.500	3.632	6.236	5.342	4.159

Table 2 Cluster Centroids

Note: Higher values are shown in greener shades. Cluster numbers correspond to those shown in the dendrograms (Figure 1). 'P' indicates the pillar number from the TTDI (Pillars 12, 13, and 14), and the number following the dot refers to the specific indicator's position within that pillar.

4 Discussion

Building on prior research that emphasized the role of shared tourism assets and regional integration in promoting cross-border coopetition (e.g., Kožić et al., 2019; Kropinova, 2020), the current analysis identifies four distinct clusters of European Union countries based on natural, cultural, and non-leisure tourism resources. The composition and performance of these clusters suggest differentiated opportunities for coopetition, grounded in both geographic proximity and thematic alignment.

Cluster 1 demonstrates balanced development across cultural and non-leisure tourism indicators, with moderate natural resource assets. These countries—many of which have coastlines or are recognized for their historical cities—share significant tourism flows, especially in the form of multi-country European tours. Potential for coopetition lies in the development of thematic tourism packages (e.g., heritage cities, cruise destinations, culinary and wine tourism), leveraging regional proximity and existing transportation links. For example, joint itineraries between Austria and Portugal—reflecting growing bilateral economic ties—could be expanded into the tourism sector. Greece, Portugal, and Croatia may further cooperate on "sun and sea" tourism branding and sustainability initiatives.

Cluster 2 is characterized by strong performance in non-leisure (business) tourism, particularly due to the presence of global cities and major corporate hubs. Previous research highlighted Luxembourg, Belgium, and Ireland as business tourism centres tied to finance, technology, and EU institutions (Durand, 2022). Coopetition opportunities in this cluster revolve around transnational business travel, conferences, and innovation hubs. Cross-border strategies may include coordinated promotion of conference infrastructure, start-up ecosystems, and international education linkages. Notably, the inclusion of Cyprus (2021) offers an avenue for linking leisure and business travel across seasons.

Countries in Cluster 3 show weaker scores across all indicators, yet share characteristics of emerging or developing tourism markets with rich natural and cultural heritage. These countries often lack a strong standalone global tourism image and are sometimes perceived only through the lens of their capitals or regions (Cottrell & Cottrell, 2015). Cross-border coopetition in this group can focus on joint branding of lesser-known destinations in Central and Eastern Europe; development of regional circuits (e.g., Visegrad + Baltics cultural tours); health and wellness tourism, particularly in the Baltics (Spiriajevas, 2013; Kropinova, 2021); public-private partnerships to improve service quality and visibility. Such cooperative networks would help address the shared challenge of under-recognition and provide scale economies in marketing and infrastructure development.

Cluster 4 includes Europe's tourism powerhouses, leading in natural and cultural resources. These countries already engage in significant cross-border tourism exchanges and rank among the top ten globally in TTDI performance (World Economic Forum, 2024; Kožić et al., 2019). Coopetition potential is both extensive and mature, focusing on integrated Mediterranean and wine tourism circuits (Tafel and Szolnoki, 2021); multi-city heritage tours (e.g., Paris–Madrid–Rome itineraries); joint efforts in sustainable tourism development and destination management; cultural route collaborations supported by the EU (e.g., European Capitals of Culture). The existing interconnection of tourism services, infrastructure, and marketing platforms in this cluster enables high-level, value-added coopetition based on cultural richness and shared branding initiatives.

The current cluster configuration suggests varying levels and forms of cross-border coopetition. While Cluster 4 exemplifies advanced integration and tourism leadership, Clusters 1 and 2 present strong thematic and infrastructural complementarities that can be further developed through joint programs. Cluster 3 highlights the greatest potential for cooperative uplift through shared branding, capacity building, and regional alliances, especially in Central, Eastern, and Baltic Europe. Overall, the findings align with existing literature emphasizing regionalization and collaboration as key strategies for enhancing the competitiveness and resilience of tourism destinations (Timothy and Saarinen, 2013).

5 Conclusion

This study explores the potential for cross-border coopetition among European Union member states in the travel and tourism sector by applying cluster analysis to data from the TTDI. Focusing on 14 indicators within the Travel and Tourism Resources dimension, the analysis spans three key years—2019, 2021, and 2024—and encompasses all 27 EU countries. The results highlight patterns of consistency and change in cluster membership over time, offering meaningful insights into the dynamics of regional alignment in tourism development.

The findings highlight how cross-border coopetition can drive tourism growth, resilience, and innovation. By grouping countries with similar strengths, the study shows how shared assets—like natural or cultural resources—

can support joint strategies. These insights help policymakers and industry leaders develop coordinated policies, marketing, and thematic routes. Destinations can compete and cooperate at once, boosting global visibility, efficiency, and visitor experience.

Future research could expand on these findings by applying multi-criteria decision-making (MCDM) methods such as TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) or WSA (Weighted Sum Approach). Such approaches could offer a complementary perspective on country performance by incorporating stakeholder preferences or weighting schemes, potentially enriching the understanding of how different dimensions contribute to tourism competitiveness. While beyond the scope of this paper, integrating MCDM techniques in future work could further support nuanced decision-making and strategy development at both national and regional levels.

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

Cottrell S, Cottrell J R (2015) The state of tourism in the Baltics. Scandinavian Journal of Hospitality and Tourism 15(4): 321–326. https://doi.org/10.1080/15022250.2015.1081798

Della Corte V, Sciarelli M (2012) Can coopetition be source of competitive advantage for strategic networks. Corporate Ownership and Control 10(1): 363–379. https://doi.org/10.22495/cocv10i1c3art5

Duda R O, Hart P E, Stork D G (2000) Pattern Classification. John Wiley & Sons, New Jersey

Durand F (2022) What types of cultural cooperation exist in European cross-border areas?. Geografiska Annaler: Series B, Human Geography 104(4): 307–326. https://doi.org/10.1080/04353684.2021.2015245

Chim-Miki A F, Batista-Canino R M (2017a) The coopetition perspective applied to tourism destinations: a literature review. Anatolia 28(3): 381–393. https://doi.org/10.1080/13032917.2017.1322524

Chim-Miki A F, Batista-Canino R M (2017b) Tourism coopetition: An introduction to the subject and a research agenda. International Business Review 26(6): 1208–1217. https://doi.org/10.1016/j.ibusrev.2017.05.003

Chim-Miki A F, da Costa R A, Oliveira-Ribeiro R (2024) Dimensions of tourism coopetition for social value creation. Anatolia 35(3): 457–469. https://doi.org/10.1080/13032917.2023.2247800

Kožić I, Sorić P, Sever I (2019) Interdependence of international tourism demand for Mediterranean countries: impact of demand shocks. International Journal of Tourism Research 21(1): 97–107. https://doi.org/10.1002/jtr.2244

Kropinova E (2021) Transnational and cross-border cooperation for sustainable tourism development in the Baltic Sea Region. Sustainability 13(4): 2111. https://doi.org/10.3390/su13042111

Lorgnier N, Su C J (2017) Considering coopetition strategies in sport tourism networks: A look at the nonprofit nautical sports clubs on the northern coast of France. Value co-creation in sport management 14(1): 87–109. https://doi.org/10.1080/16184742.2013.876436

Spiriajevas E (2013) The development of international tourism in Lithuania: a comparative analysis of regional aspects. Baltic Region 1: 82–90. https://doi.org/10.5922/2079-8555-2013-1-8

Tafel M C, Szolnoki G (2021) Relevance and challenges of wine tourism in Germany: A winery operators' perspective. International Journal of Wine Business Research 33(1): 60–79. https://doi.org/10.1108/IJWBR-11-2019-0059

Timothy D J, Saarinen J (2013). Cross-border cooperation and tourism in Europe. Trends in European tourism planning and organisation. Channel View, pp 64–76.

Vašaničová P (2025) Cross-border tourism coopetition among european countries: Insights from the travel and tourism development index. Value proposition to tourism coopetition: cases and tools. Emerald Group Publishing, Leeds, pp 25–39

World Economic Forum (2024) Travel & Tourism Development Index 2024: Insight Report. https://www.weforum.org/publications/travel-tourism-development-index-2024/

Received: 2025-06-12 Last revision: 2025-07-16 Accepted: 2025-07-16

A Dynamic Two-Stage DEA Assessment of Energy Transition Progress in the European Union

Markéta Matulová[1], Martin Flegl[2]

Abstract. This study evaluates the transformation of EU Member States' energy systems between 2013 and 2023 using a combination of Malmquist Index (MI) and two-stage Data Envelopment Analysis (DEA) framework. The model captures how countries convert resource-intensive and externally dependent energy structures into economic output and, subsequently, into renewable energy adoption across key consumption sectors. In the first stage, input variables include energy consumption per capita and import dependency, whereas outputs reflect the system's economic and environmental outcomes—GDP per capita (desirable) and greenhouse gas emissions per capita (undesirable). In the second stage, GDP per capita is treated as an input enabling investment, while outputs are sectoral renewable energy shares in electricity, heating and cooling, and transport. This structure reflects a sequential interpretation of energy transition: the capacity to generate economic output from energy use, and the effectiveness with which this capacity is reinvested into sustainable energy deployment. The model is applied to data from all EU-27 countries in 2013 and 2023 to assess energy transition progress. The results highlight significant cross-country variation in performance and improvement. Some countries demonstrate high efficiency in both stages, while others either struggle to decouple energy use from emissions or fail to translate economic gains into renewable expansion. The two-stage approach offers nuanced insights into the mechanisms of energy transition and helps identify where policy or investment gaps persist.

Keywords: data envelopment analysis, energy transition, greenhouse gas emissions, malmquist index, renewable energy

JEL Classification: C44, O13
AMS Classification: 90C90, 90A30

1 Introduction

The accelerating shift toward a low-carbon energy system has become a defining strategic priority for the European Union. The EU has set the objective of achieving climate neutrality by 2050, with intermediate targets outlined in the "Fit for 55" and REPowerEU packages (European Commission, 2022). In this context, the advancement of national energy systems increasingly depends on how quickly and efficiently countries adopt renewable energy sources (RES), improve energy efficiency, and decarbonize their electricity, heating, and transport sectors (European Commission, 2021; IEA, 2024). As of 2023, renewable energy accounted for 24.5% of gross electricity consumption in the EU, with wind and solar power emerging as the leading contributors to this transformation (Eurostat, 2024). The accelerating shift toward a low-carbon energy system has become a defining strategic priority for the European Union.

In this paper, we assess the progress of EU Member States' energy transition towards integrated renewable energy sources using a quantitative framework grounded in the methodology of Data Envelopment Analysis (DEA). DEA has emerged as a valuable tool for assessing sustainability and energy-environmental performance across countries, regions, and sectors. For example, Jalali Sepehr et al. (2019) evaluated energy efficiency in 132 countries between 2007 and 2014, finding that even low-income countries could achieve high relative efficiency. Zurano-Cervelló et al. (2019) assessed the sustainability of electricity mixes in 28 EU countries, concluding that a higher share of hydropower, wind, and solar, and a lower reliance on fossil fuels and bioenergy, significantly improves sustainability performance. Similarly, Iqbal et al. (2019) identified Australia, China, Japan, Saudi Arabia, and Poland as leading performers in a comparative efficiency assessment of 20 industrial economies. Their findings confirm that differences in energy intensity and consumption structures play a crucial role in shaping performance

[1] Masaryk University, Faculty of Business and Administration, Brno, Czechia, marketa.matulova@econ.muni.cz.
[2] Tecnologico de Monterrey, School of Engineering and Sciences, Mexico City, Mexico, martin.flegl@tec.mx.

rankings. Additional evidence from OECD countries supports the link between energy efficiency and broader aspects of sustainable development, including financial and economic stability (Ziolo et al., 2020). A more recent DEA-based study by Karasek et al. (2023) ranked Ireland as the most energy-efficient EU country in both 2013 and 2020, noting a general decrease in EU-wide efficiency during the same period.

While most early applications relied on single-stage models, recent DEA studies increasingly use two-stage or network DEA approaches to better reflect the structure of energy systems and transition processes (Zhou et al., 2008). In particular, a two-stage DEA framework allows researchers to capture intermediate products or transformation pathways. For instance, Chen et al. (2018) applied a two-stage network DEA to environmental efficiency, decomposing production processes into desirable and undesirable outcomes. Liu et al. (2023) developed a network DEA model to meausre renawable energy efficiency in 30 China's provinces. The developed model enabled evaluating ineraction between economic, social and environmental subsytems. Zhang and Zhnag (2025) evaluated power system efficiency dividing the process into fossil and non-fosil energy power generation to identify sources of overall inefficiencies.

Another key strand of research uses dynamic DEA and Malmquist Index (MI) decomposition to analyze performance changes over time. Emrouznejad and Yang (2016) suggested a MI-based framework for measuring eco-efficiency with CO_2 emissions in Chinese manufacturing industries. Similar approaches were used in the studies performed by Sueyoshi and Goto (2011) or Sheng et al. (2016). Mardani et al. (2017) document in their review many other longitudinal analyses that have been applied in Europe and globally, identifying both efficiency improvements and technological advances. The studies demonstrate that DEA, particularly in its dynamic and multi-stage forms, provides a flexible and robust framework for analyzing the complexity of energy transitions.

The remainder of the paper is structured as follows: Section 2 outlines the conceptual framework, the selection of variables, and the two-stage DEA model applied to assess the performance of EU Member States; Section 3 presents empirical results and provides an interpretation of key patterns and country-level differences; Section 4 concludes with a summary of findings and implications for energy transition policy and future research.

2 Methodology

2.1 Conceptual framework

To evaluate the structural and dynamic aspects of energy transition in the EU, we adopt a two-stage Data Envelopment Analysis (DEA) framework that conceptually separates the macroeconomic capacity of an energy system from its ability to convert that capacity into renewable energy deployment.

Stage 1: From energy consumption and dependence to economic and environmental performance

The first stage captures the core functionality of the energy system—its capacity to support economic development while minimizing environmental burden. Inputs include:

- **Energy consumption**: A proxy for the scale and intensity of energy use in the economy.
- **Import dependency**: Reflects systemic vulnerability and lack of domestic energy sovereignty.

Outputs in this stage reflect both desirable and undesirable outcomes:

- **GDP** (desirable output): Represents economic productivity and the ability to generate resources for future investment.
- **Greenhouse gas emissions (GHG)** (undesirable output): Captures environmental cost and carbon intensity of the energy system.

This stage thus evaluates how efficiently an energy system converts raw energy flows—given its external dependencies—into economic value while controlling environmental harm.

Stage 2: From economic capacity to renewable energy deployment

The second stage focuses on the transformation of available economic capacity (GDP per capita) into tangible progress in clean energy transition. Here, **GDP per capita** functions as an input, reflecting the potential for investment in energy infrastructure and innovation. Outputs include the sectoral shares of renewable energy:

- **RES in electricity generation (RES-E)**
- **RES in heating and cooling (RES-H&C)**
- **RES in transport (RES-T)**

2.2 Data and methods

The data used in this study were obtained from the official energy statistics published by the European Commission: Directorate-General for Energy (2025). Summary of the selected variables and their corresponding descriptive statistics are provided in Table 1. To ensure consistency in the analytical framework, we transformed absolute values of energy consumption, GDP, and greenhouse gas emissions into per capita terms. This adjustment was necessary to avoid bias in efficiency estimation, as our model includes several ratio-based indicators that would be distorted by differences in country size. GDP values are expressed in constant 2015 prices and exchange rates to ensure comparability across countries and over time.

		2013				2023			
Variable	Units	min	mean	std	max	min	mean	std	max
Energ_cons	toe/cap	1.59	3.42	1.49	8.08	1.62	3.02	1.03	5.99
Import_dep	%	12.74	56.79	26.39	100.0	3.66	59.21	24.32	100.00
GHG	t CO2 eq./cap	5.30	9.29	3.60	21.10	4.10	7.12	1.84	11.80
GDP	ths EUR/cap	6.09	26.20	19.04	96.06	8.84	31.76	20.87	95.07
RES-T	%	0.48	5.26	3.27	15.32	0.92	10.03	6.12	33.64
RES-E	%	1.58	25.72	17.61	68.91	10.74	42.58	20.81	87.76
RES-H&C	%	4.00	25.34	15.13	61.71	7.86	34.66	17.83	67.07

Table 1 Summary statistics of input and output variables

We use an output-oriented BCC models in both stages of the analysis, as the aim is to expand desirable outputs while accounting for heterogeneity in country size and resource conditions. To incorporate undesirable outputs in the first stage, we follow the transformation approach proposed by Seiford and Zhu (2002). This method applies a monotone decreasing transformation to undesirable outputs (GHG emissions), allowing the standard DEA framework to model both the expansion of desirable outputs and the contraction of undesirable ones. Specifically, we transform the undesirable output y^- using a linear decreasing function $\tilde{y}^- = M - y^-$, where M exceeds the maximum observed value. This transformation ensures the model remains linear and classification invariant, preserving interpretability and comparability across decision-making units (DMUs). Overall efficiency of countries' energy transition progress is computed as the product of the scores from the individual stages.

To assess the evolution of energy system performance over time, we apply the Malmquist productivity index to compare efficiency between the years 2013 and 2023. The Malmquist index is particularly well suited for capturing dynamic changes, as it decomposes total productivity change into two components: Efficiency change (EffCH) reflects how much a decision-making unit has moved closer to or further from the best-practice frontier between two periods; and Technological change (TechCH) captures the shift in the frontier itself, indicating whether the performance standard has improved due to innovation, policy reforms, or structural transitions. A value greater than 1 indicates productivity improvement over the period, while a value less than 1 signals decline. Let's denote by θ_j^i overall DEA efficiency of a unit in period i, relative to the frontier in period j, where indexes $i, j \in \{1,2\}$ denote the individual periods 2013 or 2023. Then $\mathrm{MI} = \mathrm{EffCH} \times \mathrm{TechCH}$, where

$$\mathrm{EffCH} = \frac{\theta_2^2}{\theta_1^1}, \mathrm{TechCH} = \sqrt{\frac{\theta_1^2}{\theta_2^2} \times \frac{\theta_1^1}{\theta_2^1}} \quad (1)$$

3 Results

The DEA efficiency scores for the year 2023 are presented in Figure 1, illustrating the leading performers among the countries analyzed.

The results revealed that **Estonia** (EE) and **Sweden** (SE) emerge as clear frontrunners, achieving full efficiency scores in both stages of the model. This indicates that these countries simultaneously operate highly efficient energy-economic systems and successfully transform their economic capacity into high levels of renewable energy deployment across sectors. More nuanced patterns are observed for countries exhibiting strong divergence between

the two stages. For instance, **Luxembourg** (LU) and **Ireland** (IE) achieve full efficiency in Stage 1, meaning they successfully decouple energy use from economic growth while maintaining low per capita emissions. However, both countries score relatively low in Stage 2, with Luxembourg reaching only 0.27 and Ireland 0.46. This suggests that despite their favorable macroeconomic and environmental structures, their ability to convert economic resources into sectoral renewable energy deployment remains limited, potentially reflecting challenges in scaling renewables due to geographical constraints, economic structure, or policy design. In contrast, countries such as **Bulgaria** (BG) and **Croatia** (HR) exhibit the opposite pattern. They perform poorly in Stage 1, indicating relatively high energy intensity and GHG emissions per capita (Bulgaria at 0.41 and Croatia at 0.53). However, both countries achieve full efficiency in Stage 2, reflecting highly efficient deployment of renewables relative to their available GDP per capita. This may reflect targeted policy efforts to support renewable development (especially in heating, cooling, and transport) despite broader structural inefficiencies in the energy-economic system. These mixed patterns highlight the importance of separating structural energy performance from renewable deployment capacity when evaluating energy transition dynamics.

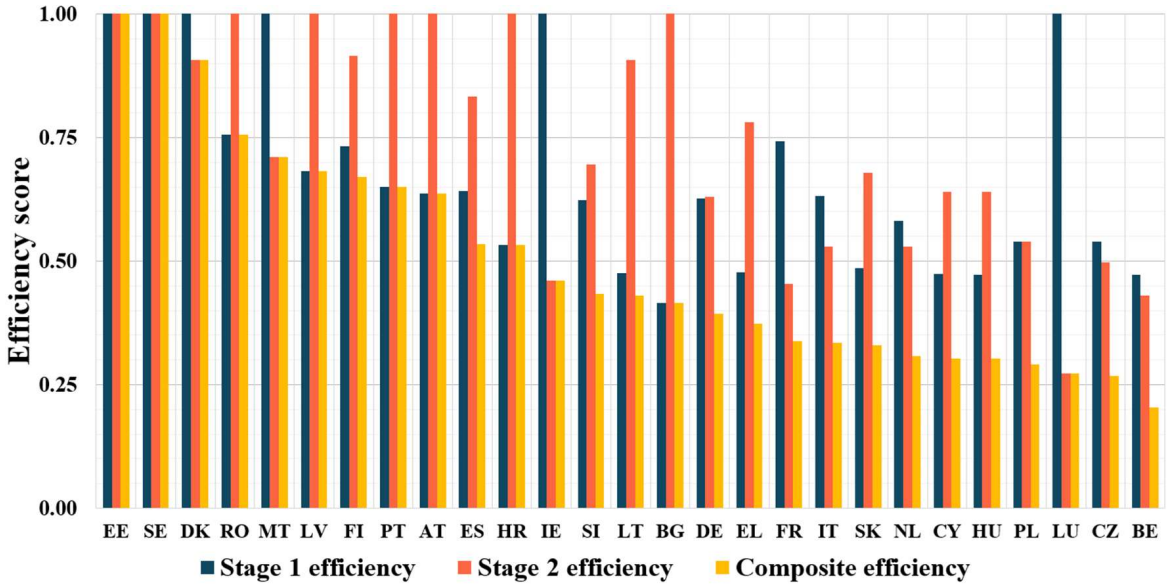


Figure 1 EU countries ordered by the overall energy transition efficiency, 2023

Importantly, countries that rank highest in static DEA efficiency for 2023 do not necessarily perform as well when examined dynamically. For many of these countries, high static efficiency largely reflects a favorable starting position, which limits further potential for improvement. In contrast, countries with more room to catch up may exhibit stronger productivity growth over time. **Poland** (PL) and **Romania** (RO) are particularly illustrative examples of this dynamic as we can see in Figure 2 displaying the decomposition of the Malmquist productivity index, providing additional insights into the sources of productivity change over time.

Poland (PL) demonstrates strong improvements in both stages of the analysis, reflected by high Malmquist Index values: the country achieved a total Malmquist Index of approximately 1.66. Notably, a substantial share of this improvement was driven by efficiency gains: the overall efficiency change (EffCH) component reached approximately 1.84 (1.26 and 1.46 in the individual stages), while technological change (TechCH) contributed only modestly, at around 0.91 (0.86 and 1.05 in the individual stages). This indicates that Poland's performance improvement was primarily the result of structural catch-up and better utilization of existing capacities, rather than frontier-shifting technological advancements.The results showing more prominent progress in the Stage 2, which aligns with substantial national efforts to restructure the energy system: The country has seen rapid growth in renewable electricity, especially photovoltaic systems, which rose from virtually zero in 2010 to over 950 ktoe by 2023, programmes such as Mój Prąd (for rooftop PV), Czyste Powietrze (for clean heating), and an active auction system. Those strategies have contributed to both efficiency improvements (EffCH) and technological change (TechCH). Poland's legal and regulatory framework was further enhanced in 2023, accelerating investment and implementation of RES projects (Observ'ER, 2025).

Romania (RO) followed a somewhat different trajectory, combining an already relatively high share of renewables in electricity with further improvements in both energy and emissions efficiency. The total Malmquist Index reached approximately 1.33. In Romania's case, efficiency change (EffCH) amounted to 1.283 (1.28×1.00), while

technological change (TechCH) contributed 1.035 (0.90×1.15). These results suggest that while Romania benefited from a relatively stable technological frontier, most of the progress came from substantial internal structural adjustments and improved utilization of existing capacities. The balanced performance in both stages reflects a combination of improved economic output per unit of energy and emissions, and a high share of renewables: In 2023, renewables represented 47.4% of Romania's electricity generation, among the highest in the EU, the deployment of wind and solar has been supported by green certificates and public investment schemes. Romania has thus benefited from both energy intensity reduction and from proactive investment in cleaner technologies (Observ'ER, 2025).

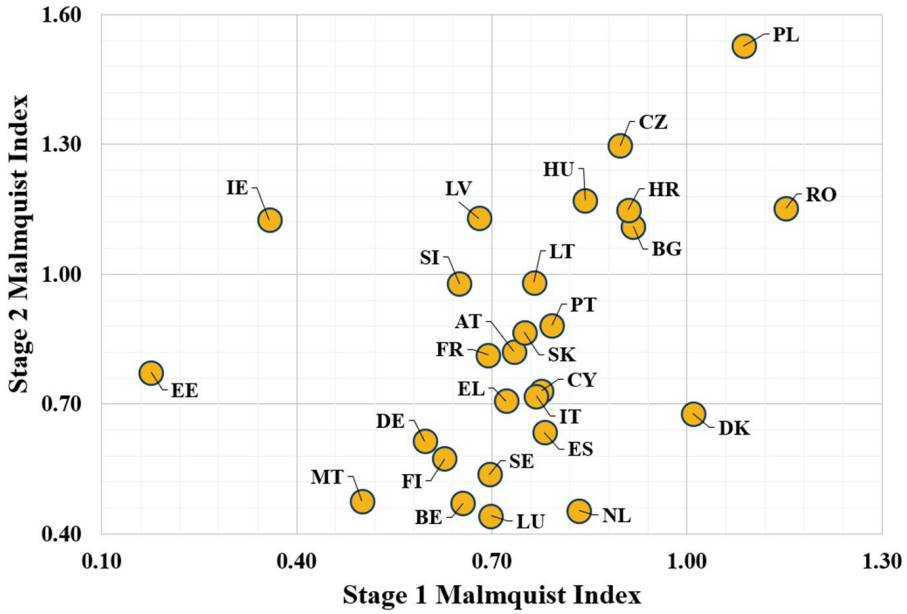


Figure 2 EU countries by the Malmquist index for the individual stages

4 Conclusions

This study applied an enhanced two-stage DEA model combined with Malmquist index decomposition to evaluate the energy transition performance of EU countries between 2013 and 2023. The two-stage framework allows for a separation of efficiency in energy-economic structure (Stage 1) and the capacity to translate economic performance into sectoral renewable deployment (Stage 2). The analysis revealed substantial heterogeneity in the performance across countries. Overall, the observed patterns confirm the importance of analyzing both static and dynamic aspects of the energy transition. While static DEA efficiency highlights a country's current positioning relative to the frontier, dynamic analysis offers a richer understanding of the direction and magnitude of change over time. The two-stage DEA-Malmquist framework thus helped reveal both countries' structural strengths and the momentum behind their energy transition efforts.

References

Chen L, Lai F, Wang Y-M, et al (2018) A two-stage network data envelopment analysis approach for measuring and decomposing environmental efficiency. Computers & Industrial Engineering 119:388-403. https://doi.org/10.1016/j.cie.2018.04.011

Emrouznejad A, Yang G-l (2016) A framework for measuring global Malmquist–Luenberger productivity index with CO2 emissions on Chinese manufacturing industries. Energy 115:840-856. https://doi.org/10.1016/j.energy.2016.09.032

European Commission (2021) 'Fit for 55': delivering the EU's 2030 Climate Target on the way to climate neutrality. https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/fit-55-delivering-eus-2030-climate-target-way-climate-neutrality Accessed 8 Apr 2025.

European Commission (2022) REPowerEU: energy policy in EU countries' recovery and resilience plans https://www.consilium.europa.eu/en/policies/repowereu/ Accessed 8 Apr 2025.

European Commission: Directorate-General for Energy (2025) EU energy statistical pocketbook and country datasheets. https://energy.ec.europa.eu/data-and-analysis/eu-energy-statistical-pocketbook-and-country-datasheets_en. Accessed 10 Apr 2025

Eurostat (2024) Renewable energy statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Renewable_energy_statistics. Accessed 12 Apr 2025

IEA, IRENA, UNSD, World Bank, WHO (2024) Tracking SDG 7: The Energy Progress Report. World Bank, Washington DC.

Iqbal W, Altalbe A, Fatima A, Ali A, Hou Y (2019) A DEA Approach for Assessing the Energy, Environmental and Economic Performance of Top 20 Industrial Countries. Processes 7:902. https://doi.org/10.3390/pr7120902

Jalali Sepehr M, Haeri A, Ghousi R (2019) A cross-country evaluation of energy efficiency from the sustainable development perspective. International Journal of Energy Sector Management 13:991–1019. https://doi.org/10.1108/IJESM-03-2018-0005

Karasek A, Fura B, Zajączkowska M (2023) Assessment of Energy Efficiency in the European Union Countries in 2013 and 2020. Sustainability 15:3414. https://doi.org/10.3390/su15043414

Liu J, Zhneg Y, Hu X, Yu S (2023) Assessing renewable energy efficiency to identify improvement strategies: A network data envelopment analysis approach. Energy for Sustainable Development 76:101308. https://doi.org/10.1016/j.esd.2023.101308

Mardani A, Zavadskas EK, Streimikiene D, Jusoh A, Khoshnoudi M (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

Observ'ER (2025) Policy and statistic reports. In: EurObserv'ER. https://www.eurobserv-er.org/policy-and-statistic-reports/. Accessed 10 Apr 2025

Seiford LM, Zhu J (2002) Modeling undesirable factors in efficiency evaluation. European Journal of Operational Research 142:16-20. https://doi.org/10.1016/S0377-2217(01)00293-4

Sheng Y, Wu Y, Shi X, Zhang D (2015) Energy trade efficiency and its determinants: A Malmquist index approach. Energy Economics 50:306-314. https://doi.org/10.1016/j.eneco.2015.05.019

Sueyoshi T, Goto M (2013) DEA environmental assessment in a time horizon: Malmquist index on fuel mix, electricity and CO2 of industrial nations. Energy Economics 40:370-382. https://doi.org/10.1016/j.eneco.2013.07.013

Zhang Q, Zhang R (2025) An extended mixed-network DEA method to analyze the power supply system with shared resources. Energy 324:136110. https://doi.org/10.1016/j.energy.2025.136110

Zhou P, Ang BW, Poh KL (2008) A survey of data envelopment analysis in energy and environmental studies. European Journal of Operational Research 189:1-18. https://doi.org/10.1016/j.ejor.2007.04.042

Ziolo M, Jednak S, Savić G, Kragulj D (2020) Link between Energy Efficiency and Sustainable Economic and Financial Development in OECD Countries. Energies 13:5898. https://doi.org/10.3390/en13225898

Zurano-Cervelló P, Pozo C, Mateo-Sanz JM, Jiménez L, Guillén-Gosálbez G (2019) Sustainability efficiency assessment of the electricity mix of the 28 EU member countries combining data envelopment analysis and optimized projections. Energy Policy 134:110921. https://doi.org/10.1016/j.enpol.2019.110921

Received: 2025-04-13 Last revision: 2025-06-06 Accepted: 2025-07-07

Chaotic Optimization with Different Chaos Maps for Tuning XGBoost: Application to Gas Demand Forecasting

Vojtěch Vávra[1]

Abstract. Accurate forecasting of gas consumption plays a crucial role in ensuring the stability and efficiency of energy systems. This paper explores the application of chaotic optimisation for tuning the hyperparameters of the XGBoost algorithm, with a particular focus on how different chaos maps influence forecasting performance. To enhance the model's predictive capabilities, several chaotic maps including Henon, Kent, Tent, Logistic, and Random evaluated within a metaheuristic optimisation framework.

The methodology is applied to the short-term prediction of aggregated daily gas consumption in the Czech Republic, with the Root Mean Square Error (RMSE) used as the primary evaluation metric. The study aims to assess the ability of chaos-based strategies to improve optimisation behaviour by helping the algorithm escape local optima and discover more accurate solutions.

Keywords: Chaotic Optimization, Chaos Maps, XGBoost, Hyperparameter Tuning, Gas Consumption

JEL Classification: C44, C53
AMS Classification: 90C59, 68T20

1 Introduction

Gas demand forecasting is a pivotal component in ensuring the operational efficiency and economic stability of contemporary energy systems. In the context of escalating market volatility and an overarching emphasis on sustainability, the necessity for precise short-term forecasts has become paramount. Forecasts of daily gas consumption plays a key role in market trading decisions and pipeline organisation.

Machine learning models, in particular gradient-boosted decision trees such as XGBoost, have been shown to be highly effective tools for time series forecasting. However, it should be noted that the performance of these models is contingent upon the optimisation of hyperparameters. The identification of the optimal set of hyperparameters is frequently a non-trivial task due to the presence of non-convex, high-dimensional search spaces, where standard tuning methods such as grid or random search may struggle to provide reliable results.

The present paper puts forward the notion of employing chaotic optimization algorithms as a potential solution to this challenge. Chaotic systems have been shown to exhibit efficiency to explore hyper-parameter grid in optimized way. Chaos maps cover different approaches that could be utilized to obtain different results and explore the search space in nonlinear way. Since models in daily run could run into different circumstances wider pallet of models is an interesting approach to succeed in the fast modern society.

The principal contributions of this study are as follows: The implementation of a chaotic optimization framework tailored for XGBoost hyperparameter tuning was the first element of the study. This was followed by a comparative analysis of several chaos maps, including Logistic, Tent, Henon, Kent, and a random baseline. Finally, the application of this framework was made to real-world gas demand data from the Czech Republic.

The remainder of the paper is structured as follows. The second section of this text provides a comprehensive overview of the extant literature concerning gas forecasting, XGBoost, and chaotic optimisation. The third section of this text provides a comprehensive overview of the methodological approach that has been employed, in addition to the proposed optimisation framework. The subject of Section 4 is the description of the chaos maps that are the subject of consideration. The fifth section is devoted to a comprehensive exposition of the dataset and the experimental setup. The sixth section of this text is dedicated to the presentation of the results and the ensuing discussion. Finally, Section 7 concludes the paper and suggests directions for future research.

[1] Prague University of Economics and Business, Faculty of informatics and statistics, nám. W. Churchilla 1938/4, 130 67 Praha 3, vojtech.vavra@vse.cz

2 Background and Related Work

Forecasting methods for gas consumption often incorporate exogenous factors such as temperature, day of week effects, and holidays. Traditionally, statistical models like ARIMA were used. More recently, machine learning models such as neural networks, SVR, and tree-based methods have gained popularity due to their capacity to model nonlinear dependencies and interactions Fan and Hyndman 2019.

Specifically for the Czech Republic a study Svoboda et al 2021 discovered different approaches for short-term gas prediction including random forests and ARIMA. However, hyperparameter optimization is overlooked by this study since only light gradient boosting is performed.

In this study is applied XGBoost as it is widely incorporated in regression tasks and time series analysis. Its success is partly due to its efficient implementation and built-in regularisation. Despite these strengths, its performance is strongly influenced by hyperparameters Chen and Guestrin 2016. Proper tuning is essential but difficult, as the parameter space is often high-dimensional and non-convex.

Chaos theory provides an alternative to random-based search strategies. Chaos maps generate sequences that appear random but are actually deterministic. They exhibit properties such as sensitivity to initial conditions and ergodicity, making them well suited for global optimisation. Previous research has integrated chaos maps into algorithms like PSO, GA, and DE, showing improvements in convergence and solution diversity Wang et al 2015.

3 Methodology

The optimisation problem is defined as minimising the RMSE of XGBoost forecasts on daily gas demand. Let θ represent the vector of hyperparameters and $f(X;\theta)$ the model output. The objective becomes:

$$\theta^* = \arg\min_{\theta} \mathrm{RMSE}(y, f(X;\theta))$$

Each candidate solution is represented as a set of hyperparameters. A chaotic map is used to evolve initial values over iterations. The sequence values are rescaled to fall within valid hyperparameter ranges. During each iteration, new candidate vectors are generated and evaluated; updates are accepted if performance improves. The process continues until a predefined number of iterations is reached.

3.1 Search Space

This study employs chaotic optimisation algorithms to tune a selected subset of XGBoost hyperparameters that influence model performance. These include all the parameters listed in Table 1. Column Step is ± maximal allowed change of parameter (neighborhood). Hyperparameters are set in Range.

The objective is to identify the optimal combination of these parameters to minimise forecasting error in short-term gas demand prediction. The search ranges for each parameter are defined based on prior studies and domain knowledge.

Variable	Range	Step	Description
n_estimators	Int [300, 1000]	50	Number of trees
max_depth	Int [3, 10]	1	Maximum tree depth
learning_rate	Float [0.0001, 0.05]	0.005	Learning rate
subsample	Float [0.8, 1.0]	0.05	Fraction of rows sampled per tree
colsample_bytree	Float [0.8, 1.0]	0.05	Fraction of columns sampled per tree
reg_alpha	Float [$\ln(10^{-5})$, $\ln(0.1)$]	0.1	L1 regularization (log-scale)
reg_lambda	Float [$\ln(10^{-5})$, $\ln(1.0)$]	0.1	L2 regularization (log-scale)
gamma	Float [0, 0.3]	0.05	Minimum loss reduction to make a split (pruning)
min_child_weight	Float [$\ln(1)$, $\ln(3)$]	0.1	Minimum sum of instance weight in a child (log-scale)

Table 1 Overview of Search Space for XGBoost hyperparameters.

3.2 Chaos maps

Chaotic maps are deterministic systems that exhibit pseudo-random behavior characterized by sensitivity to initial conditions and topological mixing. These properties make them suitable for global optimization, where they can help avoid premature convergence by enhancing exploration of the search space.

This study considers five such maps: Logistic, Tent, Henon, Kent, and Random (baseline). Their formulations and characteristics are shown in Table 2.

Chaos map	Formula	Description
Logistic	$x_{n+1} = rx_n(1 - x_n)$, $x_n \in (0, 1),\ r \in [3.57, 4]$	Classical chaotic map; simple, nonlinear, and sensitive to initial conditions.
Tent	$x_{n+1} = \begin{cases} \frac{x_n}{\mu}, & x_n < \mu \\ \frac{1-x_n}{1-\mu}, & x_n \geq \mu \end{cases}$ with $\mu = 0.5$	Piecewise linear; strong ergodicity and uniform distribution.
Henon	$\begin{cases} x_{n+1} = 1 - ax_n^2 + y_n \\ y_{n+1} = bx_n \end{cases}$ with $a = 1.4,\ b = 0.3$	Two-dimensional; generates complex chaotic attractor; rich dynamics.
Kent	$x_{n+1} = \sin(\pi x_n)$, $x_n \in (0, 1)$	Smooth 1D map; continuous and highly sensitive to initial values.
Random	$x_n \sim \mathcal{U}(0, 1)$	Pure stochastic sampling; used as a non-chaotic baseline.

Table 2 Overview of chaotic maps and baseline used for sample generation.

3.3 Objective function

The performance of forecasting models in this study was evaluated using several standard error metrics. Each metric captures a different aspect of prediction accuracy and contributes to a more comprehensive understanding of model behavior in Table 3.

Metric	Formula	Description
MAE	$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$	Mean Absolute Error: Average of absolute errors, less sensitive to outliers.
MSE	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	Mean Squared Error: Penalizes large errors more due to squaring.
RMSE	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$	Root Mean Squared Error: Square root of MSE, in the same units as target.
MAPE	$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n} \left\lvert \frac{y_i - \hat{y}_i}{y_i} \right\rvert$	Mean Absolute Percentage Error: Percentage-based error metric. May be unstable when $y_i \approx 0$.
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$	Coefficient of Determination: Proportion of variance explained by the model.

Table 3 Overview of performance metrics used for model evaluation.

4 Data and Experimental Setup

The dataset used in this study consists of aggregated daily gas consumption data for the Czech Republic, covering the period from 2016-07-08 to 2024-11-25. Data was obtained from OTE, a.s. 2025. The target variable is the total volume of gas delivered to end consumers each day.

In addition to gas demand, several explanatory variables are included: average daily air temperature, apparent temperature, wind speed, direct solar radiation and other weather inputs. From day specific values (non)working

day, seasonality, heating period, days degrees and historical lagged consumption values are utilized. Weather data was obtained from Open Meteo by Zippenfenig 2023. Important note is that gas consumption is highly correlated with weather situation in the Czech Republic. Calendar information was constructed based on Czech public and school holidays, weekends, and working days. Heating period and days degrees was created as it is defined by Czech Hydrometeorological Institute 2025.

The primary metric used for model evaluation was the Root Mean Square Error (RMSE), due to its sensitivity to large errors and common usage in energy forecasting tasks. Other metrics were also computed since each tels it's different story.

Maximum allowed iterations for whole optimization process was set to 100 due to calculation time efficiency. One calculation took around 12 minutes (in total 10 hours). This is acceptable for daily retrain of models and one calculation is in range of reasonable amount of time to wait.

During the training phase, cross-validation was performed with 10 folds, using a date period from 2016-07-08 to 2023-11-25. The trained model was then used to predict the period from 2023-11-25 to 2024-11-25.

All experiments were conducted using Python 3.11 on a laptop with a 6-core Intel Core i7-10750H processor and 16 GB of memory.

5 Results and Discussion

Across all models, the differences in average performance shown in Table 4 are relatively small, indicating stable and consistent results regardless of the training data generator. The **Kent map** achieves the best average performance, with the lowest MAE (10,092), RMSE (12,783) and mean error (7,323), as well as the highest R^2 (0.9839). The Kent map shows slightly inferior performance metrics across all categories, though still within a narrow margin, suggesting no significant degradation.

These results suggest that all chaotic sequences provide strong predictive performance with only marginal differences and that the choice of chaotic map may slightly affect precision, particularly in terms of absolute error.

train	MAE	RMSE	Mean Error	R-squared	Model Score
Henon	10139	12833	7415	0.9838	0.9838
Kent	10229	12921	7424	0.9835	0.9835
Logistic	10161	12885	7378	0.9836	0.9836
Random	10148	12863	7416	0.9837	0.9837
Tent	10092	12783	7323	0.9839	0.9839

Table 4 Average value for each metric.

The Henon and **Tent maps** stand out when considering the best-obtained results shown in Table 5. The Henon map achieves the lowest MAE (9,716), while the Tent map delivers the lowest mean error (6,900) and the highest R-squared value (0.9848). The Random and Logistic sequences also produce competitive results, particularly with regard to RMSE and Model Score.

The improvement in the best-case values compared to the average suggests the potential for fine-tuning or stochastic variability in performance, which could be exploited through repeated runs or hyperparameter optimisation.

train	MAE	RMSE	Mean Error	R-squared	Model Score
Henon	9716	12363	6919	0.9849	0.9849
Kent	10039	12790	7125	0.9839	0.9839
Logistic	9901	12648	7084	0.9842	0.9842
Random	9743	12537	7015	0.9845	0.9845
Tent	9833	12430	6900	0.9848	0.9848

Table 5 Best obtained values per metric.

6 Conclusion and Future Work

This study examined how different chaotic maps affect the initialisation and training of ML algorithm XGBoost for short-term gas consumption forecasting in the Czech Republic. The experimental results showed that all of the chaotic sequences that were evaluated led to high-performing models, with only marginal differences between them. This is due to a high correlation between weather and consumption data.

The Tent map produced the best average performance across most metrics, and the Henon and Tent maps delivered the most accurate predictions in the best-case scenario. Even the random baseline produced competitive results, confirming that the modelling framework is robust with respect to the choice of data initialization strategy for case of CZE gas consumption.

These results emphasise the potential of chaos-based data generators in supporting model stability and enhancing the accuracy of energy forecasting. This suggests their broader applicability in other machine learning domains where data diversity and sensitivity to initialisation are critical factors.

Several directions will be pursued for future work. Firstly, the framework could be extended to other energy systems or domains, such as electricity demand forecasting, renewable energy forecasting and industrial load modelling, where similar environmental dependencies exist. Secondly, other machine-learning algorithms could be implemented with this framework to receive different types of prediction to robust the prediction range for the short prediction period to create even more powerful ensemble model.

Acknowledgements

The research is supported by the Internal Grant Agency, Faculty of Informatics and Statistics, Prague University of Economics and Business – project no. F4/10/2024.

References

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp 785–794

Czech Hydrometeorological Institute (2025) Heating period. https://www.chmi.cz/historicka-data/pocasi/otopna-sezona

Fan S, Hyndman RJ (2019) A review of load forecasting techniques for smart grid and buildings. Wiley Interdisciplinary Reviews: Energy and Environment 8(3)

OTE, a.s. (2025) Yearly market report 2025. https://www.ote-cr.cz/en/statistics/yearly-market-report?set_language=en&date=2025-01-01

Svoboda R, Kotik V, Platos J (2021) Short-term natural gas consumption forecasting from long-term data collection. Energy 218:119430

Wang G, Guo S, Dong J (2015) A chaotic gravitational search algorithm for global optimization. Knowledge-Based Systems 84:82–94

Zippenfenig P (2023) Open-meteo.com weather api. https://doi.org/10.5281/zenodo.7970649, https://open-meteo.com/

Received: 2025-05-20 Last revision: 2025-07-08 Accepted: 2025-08-19

Lyrical Themes in Contemporary Heavy Metal Albums: A Case Study

Josef Kunhart[1], Jiří Pilný[2], Simona Huguet[3], Jan Bartoška[4]

Abstract. Heavy metal has been a leading force in the world of rock music for over five decades, and it is still going strong after all these years. Heavy metal bands are characterized by monumental sound which is supported by bass, drums and distorted guitars. Since its inception, heavy metal lyrics have been associated with a certain degree of controversy, mystery, and provocation. In this conceptual paper, we propose a method to identify the most common lyrical themes in the genre. The proposed method uses a semantic network and analytic network process to calculate the importance of individual lyrical themes, based on the lyrical themes of the leading bands in the heavy metal genre. The method provides a tool for evaluating lyrics on selected albums by a decision maker. We demonstrate the proposed concept in a case study. In the case study, we examine and evaluate the lyrics of contemporary heavy metal albums by bands that have a significant influence on the scene.

Keywords: analytic network process, case study, decision making, heavy metal, lyrical themes, semantic networks

JEL Classification: C44
AMS Classification: 90C15

1 Introduction

Heavy metal is a genre of rock music that originated in the late 1960s and early 1970s in the United Kingdom with hard rock and later heavy metal bands such as Black Sabbath, Deep Purple, and Led Zeppelin, and afterwards spread into the United States and the rest of the world (Larson, 2004). Heavy metal further developed in the 1970s and 1980s, mostly shedding blues influences, incorporating punk rock elements, and increasing overall speed and heaviness of the music with bands such as Judas Priest, Iron Maiden, Accept, and Motörhead (Bowe, 2009). The genre evolved into many subgenres, most notably thrash, speed, death, and black metal, and later progressive and symphonic metal (Weinstein, 2000). Nowadays, the genre is standing strong and continues to evolve. Heavy metal music is characterized by mid-to-fast-tempo riffs, by thumping basslines, crunchy riffs, extended lead guitar solos, and clean, often high-pitched vocals and anthemic choruses (Christe and Guitton, 2007). In the case of subgenres, some characteristics may be absent as the subgenres differ in song structure, instrumentation and playing style.

Heavy metal lyrical themes have always been associated with a certain degree of controversy and include occult themes, religion, violence, sex, substance abuse, and rock and metal music (Hansen and Hansen, 1991; Conway and McGrain, 2016; Cope, 2016). Weinstein (2000) divides the lyrical themes into two categories: the Dionysian theme and the Chaotic theme. The Dionysian theme embodies party, "sex, drugs and rock and roll" and a general enjoyment of life while the Chaotic theme covers dark subjects such as hell, death, and injustice. However, this classification cannot be used for all heavy metal subgenres. In example, power metal lyrics focus on fantasy and mythology themes. In 1995, Mannheim and Walser analyzed 88 metal songs to find out whether the songs' lyrics focus mainly on violence, substance abuse, and satanism. The authors discovered that the main theme in the lyrics was "longing for intensity" and negative themes or controversial themes were uncommon (Mannheim and Walser, 1995). Cheung and Feng (2019) summarized previous lyrics analyses and performed a study based on the attitude system on a corpus of 1152 heavy metal songs. According to the authors, heavy metal lyrics include themes such as freedom, social injustice and oppression, marginalization of religion and ideologies, unpleasant and disgusting objects, death, rebellion, and building heavy metal identity. The overall direction of the lyrical themes in heavy metal music generally contrasts with the lighthearted direction of lyrical themes in pop music (Weinstein, 2000).

In this conceptual paper, we propose a method for evaluating and ranking the most common lyrical themes of the leading bands in the genre. The method is based on semantic networks and analytic network process (ANP). We

[1] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, kunhart@pef.czu.cz.
[2] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, pilnyj@pef.czu.cz.
[3] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, huguet@tf.czu.cz.
[4] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, bartoska@pef.czu.cz.

have not found any publications in Scopus, Web of Science, and APA PsycNet scientific databases that analyze heavy metal lyrics with the help of multi-criteria decision analysis methods or methods similar to our proposal.

Therefore, we formulate two research questions that address this research gap:

- RQ1: What are the common lyrical themes of leading heavy metal bands?
- RQ2: What are the most popular lyrical themes of contemporary heavy metal bands?

The rest of the paper is organized as follows: In materials and methods, we describe the data sources, methods, and tools that we used. In Results, we present the results of the lyrical theme analysis and the results of the case study. In the discussion, we discuss our research and its limitations. In the conclusions, we summarize our paper and suggest future research options.

2 Materials and methods

2.1 Metal archives

Metal Archives (MA, 2025), or in full name Encyclopaedia Metallum: The Metal Archives, is an online website, database, and encyclopedia of musical artists who perform heavy metal music, including its subgenres. The database includes artists, discographies and albums, song lyrics, and user reviews. The website was launched in 2002, and is community based, non-commercial, and advertisement-free. Every registered member may contribute to the content of MA. The website utilizes a set of rules to ensure high quality content and exclude non-metal bands. Each submission of a new artist, album, or album review is carefully reviewed by a moderator. By the time of writing the paper, the database includes over 180 000 approved bands and over 130 000 album reviews. For our research, we utilize MA statistics, band details, and album lyrics. Every metal band detail includes information about the lyrical themes the band uses, album descriptions with lyrics, and albums reviews. The lyrical themes represent the community's opinion on the content of bands' lyrics.

2.2 Semantic networks

A semantic network is a notation for graphical representation of knowledge in patterns of interconnected nodes and links (also vertices and edges), forming a graph (Sowa, 2006). In a semantic network, nodes represent concepts, such as words or phrases, and links represent semantic relations between the connected nodes. Semantic networks can be used for representing human knowledge in computers and automated processing of this information. The links are tuples (source, semantic relation, target) that encode knowledge (Budel et al., 2023). Semantic networks provide a suitable tool for organizing large quantity of related information. We used semantic networks to create a semantic network of lyrical themes that models the relations between these themes based on information about lyrical themes in the band details in MA.

2.3 Analytic network process

The analytic network process (ANP) is a method used in multi-criteria decision analysis. ANP utilizes a network with goal and control criteria organized in clusters, and alternatives for modeling the problem (Saaty, 2001; Saaty, 2004). By default, ANP uses pairwise comparisons to evaluate the weights of the components in the network and rank the alternatives in the decision. The pairwise comparisons are associated with the links and semantic relations in semantic networks. The choices between alternatives are made by the decision maker (Saaty, 2001). An ANP network is represented by a super matrix which includes all nodes in the network. The result of the comparisons in the network is an unweighted super matrix, which is subsequently normalized and synthesized into a weighted matrix (Saaty, 2004). We used ANP to calculate the weights of lyrical themes in the semantic network. In our work, we replaced the pairwise comparisons with direct calculations of the weights from the lyrical theme scores from the lyrical theme analysis and directly assembled a super matrix that comprises all lyrical theme nodes.

3 Results

3.1 Evaluation of lyrical themes

To evaluate the lyrical themes of the leading bands in the genre, we selected the most reviewed eligible heavy metal bands and analyzed their lyrical themes. We constructed a semantic network for lyrical themes, which we transformed into a simple analytic network process network to calculate lyrical theme weights. The four steps for evaluating lyrical themes are displayed in figure 1.

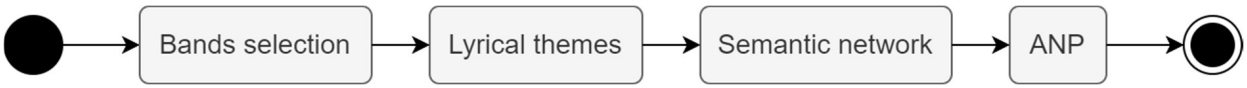


Figure 1 Steps for evaluating lyrical themes of leading heavy metal bands (source: own)

First, we selected the first ten eligible heavy metal bands, according to the Top 100 list of bands in MA by the number of reviews submitted by the users, sorted in descending order (MA, 2025). For the purposes of our research, we considered eligible any metal bands that could be associated with heavy metal genre or its closely related subgenres, specifically power, speed, thrash, progressive, and symphonic metal. We excluded rock and hard rock bands as well as more distant and extreme metal genres such as death metal or black metal. The selected bands are listed in table 1, along with their lyrical themes.

Band	Lyrical themes
Iron Maiden	Literature, History, Life, Death, War, Mythology, Society, Religion, Science fiction
Metallica	Introspection, Anger, Corruption, Deceit, Death, Life, Metal, Literature, Films
Megadeth	Society, Politics, History, Death, Religion, New world order, Addiction, Love
Black Sabbath	Doom, Drugs, Life, Death, Religion, Fantasy
Judas Priest	Fantasy, History, Allegory, Independence, Metal, Science fiction
Slayer	Death, Satan, Religion, Murder, War, Politics
Dream Theater	Introspection, Disorders, Philosophy, Emotions, Addiction, Spirituality
Overkill	Death, Violence, Rebellion, Metal, Society, War, Corruption, Humor
Helloween	Fantasy, Life, Humanity, Good/Evil, Humor, Religion, Politics, Glory, Metal
Blind Guardian	Legends, Mythology, Fantasy, Literature

Table 1 Most reviewed heavy metal bands on Metal Archives (source: own)

Second, we analyzed lyrical themes of the selected bands. We extracted all lyrical themes for selected bands into a table. We reviewed the lyrical themes and merged and renamed several lyrical themes with close meanings (Dark Fantasy → Fantasy, Myths → Mythology, Rock/Metal → Metal, Anti-religion → Religion, Epic tales → merged into Legends). Subsequently, we calculated a score for each lyrical theme. The score equals the number of occurrences of the lyrical theme among the selected bands, yielding a possible score between 0 and 10. The maximum possible score equals the number of selected bands. See table 1 for lyrical themes associated with each band. The table reflects merged lyrical themes.

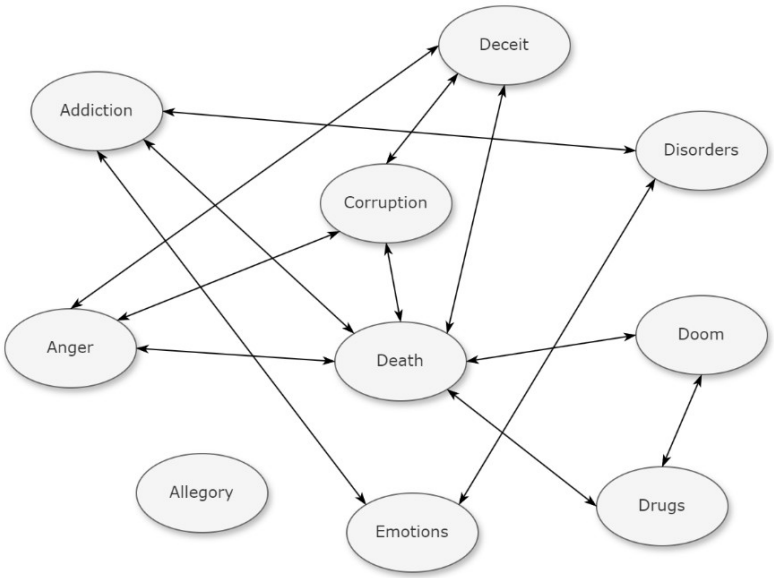


Figure 2 Partial section of the semantic network for lyrical themes (source: own)

Third, we constructed a semantic network for lyrical themes. The nodes in the network represent concrete lyrical themes, the links represent connections between the lyrical themes. We connected every two lyrical themes if they appear together in the band's detail (see table 1) as they relate to the same band's lyrics. See figure 2 for an illustrative partial section of the semantic network. The displayed part of the network includes the first ten lyrical

themes (nodes), i.e. Addiction, Allegory, Anger etc. All relations in the network are bi-directional. We do not list the entire semantic network due to its complexity and large number of links. The Allegory node is not connected to any of the first ten nodes in the figure. However, it is connected to other nodes in the entire semantic network.

Lyrical theme	Weight	Lyrical theme	Weight	Lyrical theme	Weight
Death	0.142	Corruption	0.023	Rebellion	0.006
Religion	0.097	Science fiction	0.020	Violence	0.006
Metal	0.085	Disorders	0.020	Allegory	0.006
Fantasy	0.075	Emotions	0.020	Independence	0.006
Life	0.074	Philosophy	0.020	Love	0.006
Politics	0.044	Spirituality	0.020	New world order	0.005
Literature	0.043	Mythology	0.009	Murder	0.005
Introspection	0.043	Anger	0.007	Satan	0.005
History	0.041	Deceit	0.007	Legends	0.005
Addiction	0.041	Films	0.007	Doom	0.005
Society	0.040	Glory	0.007	Drugs	0.005
War	0.039	Good/Evil	0.007		
Humor	0.023	Humanity	0.006		

Table 2 Lyrical themes with calculated weights (source: own)

Finally, we calculated the final weights of lyrical themes using a simple ANP network. We calculated an unweighted super matrix from the scores that we subsequently normalized to obtain a weighted super matrix with final lyrical theme weights. The final weights of lyrical themes are displayed in table 2. The table includes all lyrical themes along with their calculated weights, rounded to three decimal places. The lyrical themes are sorted in descending order, from the most important to the least important theme. With these results, we answer the first research question, "What are the common lyrical themes of leading heavy metal bands?". The calculated weights represent reference values, based on the leading bands in the heavy metal genre that will be used for evaluating selected album lyrics. See the case study for the details.

3.2 Case study

We provide a simple case study to demonstrate the proposed concept on a selection of heavy metal albums. In the case study, we let the decision maker analyze the lyrics of contemporary heavy metal albums from 2024. The three steps for evaluating lyrics in the case study are displayed in figure 3.

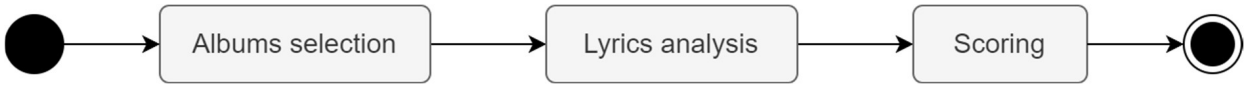


Figure 3 Steps for analyzing the album lyrics in the case study (source: own)

First, we selected a list of heavy metal albums. The selection forms a representative sample of contemporary heavy metal albums. We selected eight albums from 2024 with an overall positive user rating, i.e. user reviews in MA (2025) over 75 % at the time of writing the paper. The selected albums are presented in table 3. The table includes band names, album titles, and years. In general, albums are selected by the decision maker, who would like to evaluate the lyrical themes on his/her choice of albums.

Second, we let the decision maker analyze, and rate the lyrics of the selected albums through the lens of the lyrical themes. We provided the decision maker with an artificial intelligence tool (AI, 2024) to simply summarize the lyrics on the albums. The decision maker rated each lyrical theme on each album with a score ranging from 0 to 3. 0 means no occurrence of the given lyrical theme, 1 if the lyrical theme was only mentioned in the lyrics, 2 if the lyrical theme was used multiple times in a few songs, and 3 if the lyrical theme represents a key concept on the album, and is used in most song lyrics.

Band	Album	Year
Accept	Humanoid	2024
Bruce Dickinson	The Mandrake Project	2024
Dream Evil	Metal Gods	2024
Flotsam and Jetsam	I Am the Weapon	2024
Judas Priest	Invincible Shield	2024
Powerwolf	Wake Up the Wicked	2024
Riot V	Mean Streets	2024
Sonata Arctica	Clear Cold Beyond	2024

Table 3 Selection of heavy metal albums in the case study (source: own)

Finally, we performed scoring, i.e. calculated the final weights of the lyrical themes for the selected albums. We used reference weights of the lyrical themes from the previous chapter and the scores provided by the decision maker as inputs for calculating the final weights. We added up scores for each lyrical theme and multiplied by the reference weight for given theme to get the unweighted values. We then normalized the unweighted values to obtain final weights. See formula 1 for the equation. In the equation, w_i represents final weight, s_i stands for total score provided by the decision maker, and r_i stands for reference weight for a lyrical theme with index i.

$$w_i = \frac{s_i\, r_i}{\sum_1^n s_i\, r_i} \tag{1}$$

The results are displayed in table 4. The table includes the most popular lyrical themes in the selected albums along with their calculated weights, rounded to three decimal places. The table includes the lyrical themes with weight over 0.020, sorted in descending order. With these results, we answered the second research question, "What are the most popular lyrical themes of contemporary heavy metal bands?"

Lyrical theme	Weight
Death	0.176
Life	0.146
Fantasy	0.112
Metal	0.084
Religion	0.072
Society	0.069
Introspection	0.064
History	0.030
Spirituality	0.030
Emotions	0.023
Philosophy	0.023

Table 4 Most popular lyrical themes in contemporary heavy metal albums (source: own)

4 Discussion

In this conceptual paper, we proposed a simple method for analyzing lyrical themes on heavy metal albums. The method can be also used to evaluate the lyrics of individual songs without any necessary adjustments. Moreover, the decision maker does not need to know the implementation details of the semantic network, and calculations of the reference weights with the use of ANP. He/she is only required to analyze the lyrics and provide scores for the individual lyrical themes as input for the proposed method. The final weights of the lyrical themes are calculated using an excel file provided by the authors (this part will likely be automated if the method is implemented as software).

Considering the limitations of the proposal, we have included only ten leading heavy metal bands for the reference lyrical themes, thus limiting the number of available lyrical themes. Since this proposal is a conceptual paper, we by no means aimed for an exhaustive list of lyrical themes. With more bands and lyrical themes, the proposed

method and reference weights may (or may not) be improved. In fact, we observed only a few lyrical themes in the lyrics of the selected albums in the case study that were *really* missing in the list, namely Suffering/Pain/Fear, Fate/Destiny, and Personal choices. More bands and lyrical themes will increase the complexity of the proposed method, requiring more sophisticated semantic and ANP networks as well as more comprehensive input from the decision maker. In future research, we will try multiple variants and choose the best one. The method itself requires input from the decision maker, which may be influenced by the human factor and infer various biases in the decision-making process, i.e. anchoring or confirmation bias.

5 Conclusions

We developed a new method for analyzing lyrical themes of heavy metal albums, based on the importance of lyrical themes of leading heavy metal bands. This method allows for the evaluation of the lyrics of any selection of heavy metal albums. The albums are selected, and the lyrics are evaluated by a decision maker. In addition, we provided a case study to demonstrate the use of the method on a selection of contemporary heavy metal albums from 2024. In future research, we suggest improving the proposed concept and method by using a wider selection of reference bands and more sophisticated semantic and ANP networks with multiple levels and groups of lyrical themes. We may also introduce a simple software solution that implements the proposed method and conceals reference weights and their calculations under the hood.

References

AI (2024). AI Summarizer. https://summarizer.best/. Accessed 14 May 2025

Budel G, Jin Y, Van Mieghem P, Kitsak M (2023) Topological properties and organizing principles of semantic networks. Sci Rep 13:11728. https://doi.org/10.1038/s41598-023-37294-8

Conway KP, McGrain P (2016) Understanding Substance Use and Addiction Through the Lyrics of Black Sabbath: A Content Analysis. Substance Use & Misuse 51:1655–1663. https://doi.org/10.1080/10826084.2016.1191515

Cope AL (2016) Black Sabbath and the Rise of Heavy Metal Music, 1st edn. Routledge. https://doi.org/10.4324/9781315569499

Bowe BJ (2009) Judas Priest: Metal Gods. Enslow, Berkeley Heights, NJ

Cheung JO, Feng D (William) (2021) Attitudinal meaning and social struggle in heavy metal song lyrics: A corpus-based analysis. Social Semiotics 31:230–247. https://doi.org/10.1080/10350330.2019.1601337

Christe I, Guitton A (2007) Sound of the beast: The Complete Headbanging History of Heavy Metal. Flammarion, Paris

Hansen CH, Hansen RD (1991) Schematic Information Processing of Heavy Metal Lyrics. Communication Research 18:373–411. https://doi.org/10.1177/009365091018003005

Larson T (2004) History of Rock and Roll. Kendall/Hunt Pub., Dubuque, Iowa

MA (2025). Encyclopaedia Metallum: The Metal Archives. https://www.metal-archives.com/. Accessed 16 May 2025

Manheim JM, Walser R (1995) Running with the Devil: Power, Gender, and Madness in Heavy Metal Music. American Music 13:231. https://doi.org/10.2307/3052258

Saaty TL (2001) Decision making with dependence and feedback: The analytic network process; the organization and prioritization of complexity, 2. ed. RWS Publ, Pittsburgh, Pa

Saaty TL (2004) Decision making — The Analytic Hierarchy and Network Processes (AHP/ANP). J Syst Sci Syst Eng 13:1–35. https://doi.org/10.1007/s11518-006-0151-5

Sowa JF (2006) Semantic Networks. In: Nadel L (ed) Encyclopedia of Cognitive Science, 1st edn. Wiley

Weinstein D (2000) Heavy Metal: The Music and Its Culture, Da Capo Press, New York

Received: 2025-06-08 Last revision: 2025-07-02 Accepted: 2025-07-08

Consumer Preferences and Utility in Milk Chocolate Selection: Conjoint Analysis and TOPSIS Method

Zuzana Neničková[1], Lucie Waleczek Zotyková[2], Radmila Krkošková[3]

Abstract. This study examines consumer preferences in the context of milk chocolate selection using a conjoint analysis based on primary data collected through a questionnaire. The analysis focusses on four key attributes of the product: brand, price, weight, and content of certified cocoa. The results of the conjoint analysis provide the relative importance (weights) of individual criteria and estimate the utility associated with brand and certified cocoa content.

The resulting preference structure serves as the basis for applying the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method to rank the actual chocolate products currently available on the market. This multicriteria decision-making framework allows for a systematic comparison of alternatives with respect to an ideal product profile.

In addition, a sensitivity analysis is conducted to explore the robustness of the ranking outcomes. Specifically, we investigated the extent of the change in attribute weights required to alter the final product ranking. This provides insight into which attributes have the most influence on consumer choice under varying preference scenarios.

Keywords: conjoint analysis, preferences, TOPSIS

JEL Classification: C14, C44, C61, D01, D60, M31
AMS Classification: 62H12, 90B50, 91B06

1 Introduction

In 2024, a dramatic increase in the global market price of cocoa was observed. According to data from the World Bank (2025), the average price more than tripled compared to 2022. Naturally, this development was reflected in the prices of cocoa-based products, including milk chocolate. Consumers witnessed a rapid rise in the retail prices of this popular treat. In response to this substantial price increase, many manufacturers opted for shrinkflation – reducing the size of the packaging while maintaining (or even increasing) the price.

We decided to examine the influence of selected relevant factors on consumer behavior and to conduct a sensitivity analysis of these influences. The sensitivity analysis focused on changes in the weights (preference intensities) of the factors and the impact of such changes on the final ranking when choosing among three milk chocolate options available in Czech supermarkets. Conjoint analysis was used to determine the weights of the individual factors, and the TOPSIS method was applied to evaluate the available milk chocolate alternatives.

The sensitivity analysis was applied specifically to the weights of the factors (criteria). The weight of a single criterion was deliberately modified, and the weights of the remaining criteria were recalculated to maintain their mutual proportionality. These newly derived weights were then used within the TOPSIS method, and the resulting rankings were compared to the original ranking (based on the weights obtained from the conjoint analysis). It was found that changes in the weight of several factors – whether a complete increase or elimination – had no impact on the final ranking of the milk chocolate alternatives.

2 Utility and Contextual Value: Conceptual and Methodological Note

Utility is not absolute but contextual. Classical consumer theory assumes stable, ordinal preferences (Varian, 2014), but empirical and behavioral economics have shown that perceived utility is sensitive to situational framing (Tversky and Kahneman, 1981; Thaler, 2015). A simple example illustrates this: to a mother with a crying child

[1] Silesian University in Opava, School of Business Administration in Karvina, Department of Informatics and Mathematics, Univerzitní náměstí 1934/3, Karviná, 733 40, nenickova@opf.slu.cz.
[2] Silesian University in Opava, School of Business Administration in Karvina, Department of Informatics and Mathematics, Univerzitní náměstí 1934/3, Karviná, 733 40, waleczekzotykova@opf.slu.cz.
[3] Silesian University in Opava, School of Business Administration in Karvina, Department of Informatics and Mathematics, Univerzitní náměstí 1934/3, Karviná, 733 40, krkoskova@opf.slu.cz.

and no tissues, the act of someone offering a tissue may carry more perceived value than receiving a small amount of money. In such contexts, utility cannot be reduced to price – it is situational, relational, and time-sensitive (Ariely, 2008; Kahneman, 2011). This contextuality aligns with theories of reference-dependent preferences and constructed utility (Köszegi and Rabin, 2006), which posit that people often "construct" value during decision-making, based on what is salient in the moment. To capture such structure empirically, conjoint analysis is an appropriate method. It decomposes utility into partial contributions of individual attributes, allowing estimation of the implicit value people assign to various features or scenarios (Green and Srinivasan, 1990; Louviere et al., 2000).

The study of Tung and Chen (2016) integrates conjoint analysis with the TOPSIS method to strengthen the capability of the conjoint analysis to generate icon designs which satisfy multiple users' image perception requirements. In our approach, conjoint analysis is used to estimate the utility of individual levels of the criteria brand and certified cocoa and weights of criteria. These utility estimates and weights subsequently serve as input values for the TOPSIS method. Conjoint analysis and TOPSIS are applied as two consecutive steps rather than being integrated, as their integration is not required for the objectives of our study.

3 Methodology

Conjoint analysis was employed to calculate the weights of the individual criteria and to determine the utility values associated with each brand as well as the presence or absence of certified cocoa. Subsequently, the TOPSIS method was used to establish the ranking of three real-world milk chocolate alternatives.

3.1 Conjoint analysis

Conjoint analysis is a robust multivariate technique derived from mathematical psychology and psychometrics (Grünwald, 2011). It enables the estimation of part-worth utilities of individual product attributes by decomposing overall preferences expressed by respondents, typically using ratings or rankings. The approach is grounded in variance analysis and least squares estimation, depending on the nature of the response data.

In this study, a choice-based conjoint analysis (CBCA) is applied, which follows the discrete choice model of McFadden's random utility theory. Respondents choose between profiles composed of multiple attributes, from which their implicit utilities can be inferred. These utilities are then used to predict preferences, estimate market shares, and assess the economic value of specific attribute-level combinations (McFadden, 1974; Orme, 2006).

The design of the conjoint experiment follows an orthogonal array, ensuring independence among attribute levels and enabling unbiased estimation of main effects. To minimize the number of profiles while maintaining statistical efficiency, we employ Graeco-Latin square designs (Box et al., 1978). These allow each combination of two levels across different factors to occur exactly once, ensuring balance and orthogonality. The design supports both symmetric and asymmetric factorial structures, as defined by Addelman (1962), who formulated conditions under which all main effects can be estimated without confounding:

$$s_1 > s_2 > \cdots s_k, \text{and} \sum_{i=1}^{k} t_i \leq \frac{s_1^n - 1}{s_1 - 1} \tag{1}$$

where $s_1, s_2 \ldots, s_k$ are the numbers of levels of the factors, arranged in decreasing order, t_i denotes the number of factors that have s_i levels, n is the parameter related to the design (often the power in the prime power s_1^n).

This inequality sets the upper bound on the total number of factors $\sum t_i$ that can be estimated without confounding main effects in an orthogonal design.

After randomizing the generated plan, profiles were presented to respondents in the form of rank-order tasks. The relative importance I_j of each attribute was calculated as:

$$I_j = \frac{\max(U_j) - \min(U_j)}{\sum_j (\max(U_j) - \min(U_j))} \cdot 100 \tag{2}$$

where U_j denotes the utility estimates for levels of attribute j. To evaluate the internal consistency of the model, we use Pearson's r and Kendall's τ rank correlation coefficient. The latter is a robust non-parametric statistic suited for ordinal validation and is widely applied in ranking agreement analysis (Kendall & Babington Smith, 1939).

3.2 TOPSIS

The TOPSIS method (Hwang & Yoon, 1981) ranks a set of alternatives $A_1, A_2, \ldots, A_m$ based on multiple criteria $C_1, C_2, \ldots, C_n$ by identifying the option closest to the positive-ideal solution (PIS) and farthest from the negative-ideal solution (NIS).

To eliminate the influence of differing units of measurement across criteria, the decision matrix is normalized:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \tag{3}$$

where r_{ij} denotes the normalized value of the i-th alternative on the j-th criterion, and x_{ij} is the original value.

Let w_j be the weight of with the j-th criterion. The weighted normalized decision matrix V is then obtained by:

$$v_{ij} = w_j \cdot r_{ij} \tag{4}$$

Subsequently, the *positive-ideal solution* A^+ and *negative-ideal solution* A^- are identified. PIS comprises the best attainable performance values for each criterion: maximum values for benefit criteria and minimum values for cost criteria. Conversely, NIS includes the least desirable values – minimum for benefit and maximum for cost criteria:

$$\begin{aligned} A^+ &= \{(\max(v_{ij})|j \in J_1), (\min(v_{ij})|j \in J_2)\}, i = 1,2, \ldots, m \text{ and} \\ A^- &= \{(\min(v_{ij})|j \in J_1), (\max(v_{ij})|j \in J_2)\}, i = 1,2, \ldots, m \end{aligned} \tag{5}$$

where J_1 and J_2 represent the sets of benefit and cost criteria, respectively.

The separation measures, quantifying the Euclidean distances between alternatives and PIS and NIS, are computed:

$$D_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - A_j^+)^2} \text{ and } D_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - A_j^-)^2} \tag{6}$$

Finally, the relative *closeness coefficient* C_i for each alternative is calculated as:

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{7}$$

The value of C_i lies in the interval [0,1]. Alternatives are subsequently ranked in descending order of C_i. The alternative with the highest relative closeness is considered the most preferred option.

4 Application

Data were collected via an anonymous online questionnaire distributed in the Czech Republic through social media and community groups. Of approximately 2,000 reached individuals, 896 responded (44.8% response rate), and 833 completed questionnaires were included in the analysis. Sixty-three responses (7%) were excluded due to ineligibility or incomplete data. The survey included a conjoint task and questions on attitudes toward downsizing and price retention, reflecting perceptions of 'shrinkflation.' The sample is limited to Czech-speaking consumers within the Czech cultural and economic context..

Table 1 presents the attributes and their levels used in the conjoint analysis. Brand and price have three levels, while weight and certified cocoa have two. The levels were designed based on a preliminary study to reflect the Czech market conditions.

Attribute	Brand	Price [CZK]	Weight [g]	Certified cocoa
Attribute levels	Czech	49.9	80	Yes
	Swiss	59.9	90	No
	Slovak	65.9		

Table 1 Attribute values

Using SPSS (Statistical Package for the Social Sciences), an orthogonal design was generated to create 9 representative model profiles for comparison (Table 2). Orthogonal designs ensure that the attribute levels vary independently across profiles, allowing for unbiased estimation of the main effects of each attribute on respondents' preferences. This efficient design reduces the total number of profiles needed while maintaining statistical rigor, facilitating clear identification of the relative importance and utility values associated with each attribute level. Presenting respondents with these carefully constructed profiles enables precise inference of their implicit valuations, which is critical for robust conjoint analysis and market modeling.

Brand	Price [CZK]	Weight [g]	Certified cocoa
Swiss	59.9	80	Yes
Swiss	65.9	80	No
Slovak	49.9	80	No
Slovak	65.9	90	Yes
Slovak	59.9	80	Yes
Czech	65.9	80	Yes
Czech	49.9	80	Yes
Swiss	49.9	90	Yes
Slovak	59.9	90	No

Table 2 Plan Cards

The orthogonal design generated nine representative profiles, each combining different attribute levels in a statistically efficient manner. These profiles formed the basis for respondents' evaluations, allowing for the extraction of part-worth utilities. The resulting utility estimates and their relative importance values are summarized in the following Table 3.

		Utility Estimate	Std. Error
Brand	Czech	0.238	0.049
	Swiss	0.831	0.049
	Slovak	-1.069	0.049
Price	49.9 CZK	-3.693	0.262
	59.9 CZK	-4.434	0.314
	65.9 CZK	-4.878	0.346
Weight	80 g	1.847	0.588
	90 g	2.077	0.661
Certif. cocoa	Yes	0.485	0.037
	No	-0.485	0.037
(Constant)		6.348	0.686

Importance Values (Averaged Importance Score)	
Brand	45.270
Price	20.618
Weight	9.103
Certif. cocoa	25.009

Table 3 Utility and Importance Values

The conjoint analysis reveals that consumers place the highest importance on brand (45.3 %), favoring Swiss brands significantly over Slovak ones, with Czech brands seen moderately positively. Certified cocoa (25.0 %) is valued more than price (20.6 %), which shows typical price sensitivity. Product weight, although related to the concept of shrinkflation, has the lowest impact (9.1 %). These findings suggest that while consumers notice size reductions, they are more concerned about brand trust and perceived ethical value in the context of rising prices and input cost pressures in the chocolate market.

	Value	Sig.
Pearson's R	0.998	<0.001
Kendall's tau	1.000	<0.001

Table 4 Utility and Importance Values

The very high values of Pearson's R (0.998) and Kendall's τ (1.000), both statistically significant at $p < 0.001$, indicate excellent model performance in reproducing respondents' stated preferences (Table 4). The strength of the correlations reflects the robustness of the conjoint design and the consistency of the responses across a large sample (n = 833), supporting the reliability of the estimated part-worth utilities.

Three individual real-world milk chocolate variants (denoted as A_1, A_2 and A_3) were evaluated using the TOPSIS method. The utility values associated with the brand and the presence or absence of certified cocoa were derived from the results of the conjoint analysis. The utility of the Czech brand was 0.238, the utility of the Swiss brand was 0.831, and the utility of the Slovak brand was -1.069. The utility associated with the presence of certified cocoa was quantified at 0.485, whereas the utility of its absence was -0.485.

The weights of the individual factors were also determined through conjoint analysis. The criteria brand, weight, and certified cocoa are considered benefit-type criteria, whereas the criterion price is classified as a cost-type criterion. Table 5 shows the data for TOPSIS and the results.

Criterion	Character	Weights	A_1	A_2	A_3
Brand [utility]	benefit	0.453	0.238	0.831	-1.069
Price [CZK]	cost	0.206	59.9	65.9	49.9
Weight [g]	benefit	0.091	90	90	80
Certif. cocoa [utility]	benefit	0.250	0.485	0.485	-0.485
C_i			0.725	0.995	0.045
Ranking			2	1	3

Table 5 TOPSIS – data and resulting ranking

Subsequently, a sensitivity analysis was conducted. The influence of changes in the weights of individual criteria on the resulting ranking of alternatives by TOPSIS was examined. At any given moment, a specific criterion was considered as the one primarily being altered; its weight was experimentally adjusted – both increased and decreased. The weights of the remaining criteria were modified secondarily in such a way that their original relative proportions within the residual weight sum were preserved:

$$\frac{u'_i}{\sum u'_i} = \frac{w'_i}{\sum w'_i} \tag{8}$$

where w'_i is the original weight of the secondarily adjusted factor and u'_i is the new (adjusted) weight of the secondarily adjusted factor.

It was found that a change in the ranking occurred in only two out of eight cases: when the weight of the most important criterion, brand, was decreased, and when the weight of the third most important criterion, price, was increased. In the case of the criteria weight and certified cocoa, a change in ranking occurred only when the weight of these criteria reached a value of one. This implies that the evaluation was no longer multi-criteria, as the weights of all other criteria were zero in such a scenario. In both cases where the ranking of alternatives changed, it was observed that alternative A_3consistently remained in third place. Only the order of alternatives $A_1$and A_2 was affected. Results of the sensitivity analysis are in Table 6, where primarily altered weight represents the minimum (for increasing) or maximum (for decreasing) weight required to change the ranking of alternatives.

Criterion	Type of change	Primarily altered weight	Changed ranking (A_1, A_2, A_3)
Brand	Decreasing	0.094022	1, 2, 3
	Increasing	1.000000	---
Price	Decreasing	0.000000	---
	Increasing	0.767858	1, 2, 3
Weight	Decreasing	0.000000	---
	Increasing	1.000000	1.5, 1.5, 3
Certif. cocoa	Decreasing	0.000000	---
	Increasing	1.000000	1.5, 1.5, 3

Table 6 Sensitivity analysis results

5 Conclusions

This study investigated how brand, price, weight, and certified cocoa influence consumer preferences for milk chocolate amid rising cocoa prices and shrinkflation. Using conjoint analysis and TOPSIS, brand emerged as the most influential factor, followed by certified cocoa and price, with weight being least significant. A strong model fit (Pearson's R = 0.998; Kendall's τ = 1.000) supports the robustness of the results. Sensitivity analysis showed rankings remained stable unless the weight of brand decreased or price increased significantly.

Despite disapproval of shrinkflation, product weight had minimal effect on choice, revealing a gap between attitudes and behavior. Certified cocoa influenced preferences mainly when heavily weighted, suggesting socially driven but not dominant decision factors. These findings highlight the central role of subjective utility in consumer choices and underscore the strategic importance of brand perception in pricing and product design.

Acknowledgements

This paper 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 2025.

References

Addelman S (1962) Orthogonal Main-Effect Plans for Asymmetrical Factorial Experiments. Technometrics 4:21–46.

Ariely D (2008) Predictably Irrational: The Hidden Forces That Shape Our Decisions. HarperCollins Publishers.

Box G E P, Hunter W G, Hunter J S (1978) Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. Wiley, Hoboken.

Green P E, Srinivasan V (1990) Conjoint analysis in marketing: New developments with implications for research and practice. Journal of Marketing 54:3–19.

Grünwald O (2011) Introduction to Modeling of Buying Decisions. Acta Polytechnica 51:45–52.

Hwang C L, Yoon K (1981) Multiple Attribute Decision Making: Methods and Applications. Springer-Verlag, Heidelberg.

Kahneman D (2011) Thinking, Fast and Slow. Farrar, Straus and Giroux.

Kendal M G, Babington Smith B (1939) The Problem of m Rankings. The Annals of Mathematical Statistics 10:275–287. https://doi.org/10.1214/aoms/1177732186

Köszegi B, Rabin M (2006) A model of reference-dependent preferences. The Quarterly Journal of Economics 121:1133–1165. https://doi.org/10.1093/qje/121.4.1133

Louviere J J, Hensher D A, Swait J D (2000) Stated choice methods: Analysis and applications. Cambridge University Press. https://doi.org/10.1017/CBO9780511753831ResearchGate

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics. Academic Press.

Orme B K (2006) Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research. Research Publishers LLC.

Thaler R H (2015) Misbehaving: The Making of Behavioral Economics. W. W. Norton & Company.

Tung T C, Chen H Y (2016) Integrating conjoint analysis with TOPSIS algorithm to the visual effect of icon design based on multiple users' image perceptions. Eurasia Journal of Mathematics, Science and Technology Education 13: 1025-1040.

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211: 453–458. https://doi.org/10.1126/science.7455683PubMed+2Science+2SCIRP+2

Varian H R (2014) Intermediate Microeconomics: A Modern Approach. W. W. Norton & Company.

Worldbank.org. World Bank Commodities Price Data. https://thedocs.worldbank.org/en/doc/18675f1d1639c7a34d463f59263ba0a2-0050012025/related/CMO-Pink-Sheet-May-2025.pdf. Accessed on May 20, 2025.

Received: 2025-06-06 Last revision: 2025-06-27 Accepted: 2025-07-02

Relationship Between Military, Education and Health Expenditures in NATO Countries

Aleš Olejníček[1], Jakub Odehnal[2], Jiří Neubauer[3]

Abstract. Defense can be considered an example of a public good provided by the state and financed through the state budget. The size of military expenditure as part of government expenditure is influenced by a number of factors, such as the security and the economic environment of the country. In the case of NATO countries, there is a long-standing requirement to allocate at least 2 % of gross domestic product to defense. The current security situation in Europe is driving NATO countries to reconsider this political commitment towards a future increase in military spending above the required 2 % of GDP. The increased military budget can be financed in the form of an expansion of budget revenues, an increase of expenditures and thus an increase of the deficit and indebtedness, or in the form of a reduction of other state budget items (e.g. education expenditures) leading to a crowding-out effect of non-military expenditures and the existence of an opportunity cost. The aim of the paper is to analyze the existence of a relationship between military expenditures and social expenditures and to confirm the possible crowding-out effect of these expenditures by military expenditures. A panel data analysis between 2013 and 2021 was used to analyze the selected NATO countries. Variables characterizing military expenditure as a share of GDP, education and health expenditure as a share of GDP were obtained from the World Bank database. As control variables, (tax revenues, households consumption GDP, Gini Index and population) were used in the econometric model.

Keywords: military expenditures, crowding out, panel data

JEL Classification: H41, H5
AMS Classification: 62-07**,** 62H30

1 Introduction

The allocation of limited resources is a perennial challenge for governments across the globe, as they seek to balance the demands of competing priorities within their budgetary allocations. Governments are tasked with the provision of a wide range of goods and services, including national defense, technical infrastructure, healthcare and education. A significant proportion of these services, due to their characteristics, such as non-excludability or non-rivalrous consumption, are unlikely to be adequately provided by the private sector and therefore require government intervention. In the contemporary context of global uncertainty and a shifting geopolitical landscape, numerous nations are undertaking measures to augment their defense expenditures. The increase in military expenditure naturally engenders complications in the implementation of other significant public investments, thereby underscoring the persistent and escalating discord between fiscal priorities. Recent trends indicate that rising defense spending is facing inevitable pressure to increase spending on social programs (healthcare, education, and social security). This is due to demographic changes, such as an ageing and declining population. This dual burden on government budgets necessitates a thorough examination of how decisions regarding resource allocation in one area influence the availability of resources in another. The assurance of defense as a foundational state function is inextricably linked to the issue of the effective and rational distribution of resources available to the state and the economy, with a view to meeting the social and individual needs of citizens.

Economists have posed and persist in posing questions such as: "*Who benefits from defense and who pays for it?"* It is evident that no expense is exonerated from consideration, and national defense is not an exception to this. The present study will examine which categories of expenditure will experience a decline or slower growth in the event of an increase in defense spending. The question of whether military expenditure will take precedence over spending on other sectors, such as education, healthcare, and culture, is a crucial one. The previously mentioned ques-

[1] University of Defense, Faculty of Military Leadership, Kounicova 65, 662 01, Brno, Czech Republic, ales.olejnicek@unob.cz.
[2] University of Defense, Faculty of Military Leadership, Kounicova 65, 662 01, Brno, Czech Republic, jakub.odehnal@unob.cz
[3] University of Defense, Faculty of Military Leadership, Kounicova 65, 662 01, Brno, Czech Republic, jiri.neubauer@unob.cz

tions are directly related to the so-called *"production possibility frontier,"* which is a fundamental economic concept that describes the maximum amount of products and services that a national economy can produce with a given amount of inputs, technology, and efficiency in a given period of time. The frontiers of production possibilities and defense security are interlinked; defense decisions can affect the economy and the structure of production. Simultaneously, the prevailing state of the economy has the capacity to influence the resources available for defense and the state's capacity to respond to threats and crises. The relationship between defense production and the provision of civilian needs is illustrated by the well-known "guns and butter model". This simplified model is employed to illustrate the notion that economies are characterized by the scarcity of resources, including labor, raw materials and capital, and are compelled to determine the allocation of these resources between defense production (guns) and civilian production for consumption by the population (butter).

1.1 Review of literature dealing with the problem of crowding out between military and social expenditure (health and education spending)

Russet (1969) was among the first authors to explicitly address the relationship between military and civilian spending. Russet's analysis delineates the advent of opportunity costs in the context of national defense. He posits that in circumstances where the resources available for distribution between civilian and military expenditure are severely constrained (for example, in the event of unhealthy public finances or a financial or economic crisis), the resulting relationship will be one of negative correlation. That is to say, an increase in military expenditure will result in a decrease in civilian expenditure. However, should there be interest in increasing both items, then civilian spending will grow in tandem with military spending. In contemplating a final option, he considers a complementary relationship between the two types of spending, in which case the negative correlation will be very weak. The findings of his research in France and the United Kingdom demonstrate a robust negative correlation with civilian expenditure. However, evidence from the United States indicates that an increase in military expenditure does not necessarily result in a decrease in spending on health and education. In a 1982 study, Russett found no evidence of a crowding-out effect between military and civilian spending in the United States from 1945 to the start of the arms race initiated during Ronald Reagan's presidency (Russett, 1982). Furthermore, Kamlet, Mowery, and Su (1988) and Wildavsky (1988) confirm that this effect is evident during the Reagan era in the White House. Huang and Minz (1991) conducted an analysis of the prevailing literature on defense spending and economic performance in industrialized countries. This analysis posits the hypothesis that an increase in military expenditure has a negative impact on investment, which in effect reduces growth. Consequently, this has a harmful effect on the ability of governments to increase social spending. As Eichenberg (1992) asserts, there is a direct negative effect of defense spending on civilian spending. Building on these findings, Huang and Mintz (1991, 1992) conducted a detailed theoretical analysis of the relationship between military spending and spending on "social welfare". The prevailing consensus, stemming from analyses of the "guns and butter" model, posits that the impact of military expenditure on civilian spending is predominantly short-term. It is emphasized that there is no reason to expect short-term effects to prevail over long-term ones. The short-term effect of military expenditure, which is associated with the stimulation of demand and the reduction of unemployment, is accompanied in the long term by a decline in investment and a slowdown in economic growth. However, the authors also point out that confirmation or refutation of the crowding-out effect may be further influenced by the degree of externalities that military spending and research can generate.

In their 2011 paper, Obinger and Schmitt propose a theory that is supported by empirical evidence. This theory hypothesizes that competition between regimes, such as the Cold War and the arms race, had a positive impact on the growth of the welfare state. From their perspective, the augmentation in military expenditure was accompanied, in the long term, by an increase in expenditure on social programs and the expansion of the public sector. Bauwens (2012) examines the Middle East and North Africa due to the high level of security instability in the region. The study concludes with the assertion that there is evidence of a negative crowding-out effect between defense spending and human capital in these regions. It is argued that the opportunity costs of defense spending are higher for education than for healthcare. The mitigation of the crowding-out effect of military expenditure is determined by the level of economic development of the country in concern. In their 2015 analysis, Lin, Ali, and Lu examined the correlation between military expenditure and expenditure on health and education in 29 OECD countries over the period from 1988 to 2005. The results of the study demonstrate the positive crowding-out effect, which states that an increase in military expenditure is accompanied by an increase in expenditure on health and education. The empirical evidence suggests a correlation between increased social spending in economically advanced OECD countries and a greater willingness to engage in such expenditures. Furthermore, it has been demonstrated that

increased defense spending can contribute to human capital formation, as members of the armed forces are well cared for, physically well trained, and well qualified.

Carter, Ondercin, and Palmer (2021) have highlighted that research into the crowding-out effect between military and social welfare spending has been inadequately designed. It is imperative to consider the manner in which governments finance their expenditures. A close examination of the interconnectivity between military spending, social spending, tax revenues, debt, and the money supply is therefore essential. The results of their analysis suggest that the existence of a crowding-out effect between military and civilian spending depends on the manner in which politicians seek to maintain or increase the level of social welfare. The relationship between these factors and the size of the tax burden on the population and businesses, government debt, and the money supply is of particular importance. An analysis conducted by Wang (2022) demonstrates that the crowding-out effect between military expenditure and expenditure on education and healthcare is not a short-term phenomenon, i.e. it does not manifest itself immediately. The author posits that the manifestation of the crowding-out effect requires a minimum period of six years. The government can take corrective measures (tax increases, bonds) to delay the impact of this phenomenon. In the countries under scrutiny (namely the United Kingdom and the United States), it was determined that crowding out between military expenditure and education spending is more probable than crowding out between military expenditure and healthcare spending. Since the 1960s, there has been an observable increase in both social spending and military spending. The findings of the research call into question the assumption that the defense sector exerts a dominant influence on the allocation of resources within the public budget. The conclusions of the study presented by Njifen and Anemann (2023) reflect the diversity of current knowledge in the field, as revealed by the literature review presented by the authors. In sub-Saharan African countries, the results demonstrated no evidence of a crowding-out effect between defense and education spending. The correlation between military expenditure and educational spending is significant and positive, thus validating the hypothesis of complementarity between social and military spending. A similar correlation is observed between military expenditure and education spending in the central and southern subregions of Africa.

The following conclusions can be drawn from the literature review:

- despite the utilization of different models, estimation tools and procedures, and various data sets, the findings from the study of the crowding-out effect of military and social spending lead to ambiguous conclusions;
- in consideration of the knowledge acquired pertaining to the phenomenon, it is evident that the influence of military expenditure on social expenditure can be either direct or indirect, with a favorable (increasing) or unfavorable (decreasing) nature;
- several situations may arise in relation to military and social spending:
  - military spending crowds out social spending;
  - military spending does not crowd out social spending (here, two variants are possible, the first meaning an increase in military spending alongside social spending, the second representing an increase in military spending with no impact on social spending, which remains the same or grows at the same rate)
- The resulting relationship between military and social spending is influenced by the focus of the research itself, i.e., whether the impacts examined are direct or indirect, short-term or long-term, whether the overall impact of military spending or its individual components (salaries, operations, investments) is assessed, and, last but not least, whether external effects of military spending are considered.

2 Data a Methods

The study analyzed the impact of military spending on social spending in a sample of 14 countries between 2013 and 2021. Data on military spending and data on social spending (including healthcare and education spending) were drawn from the World Bank database (World development Indicators). Figure 1 shows the selected countries and the average values of the individual variables for the review period.

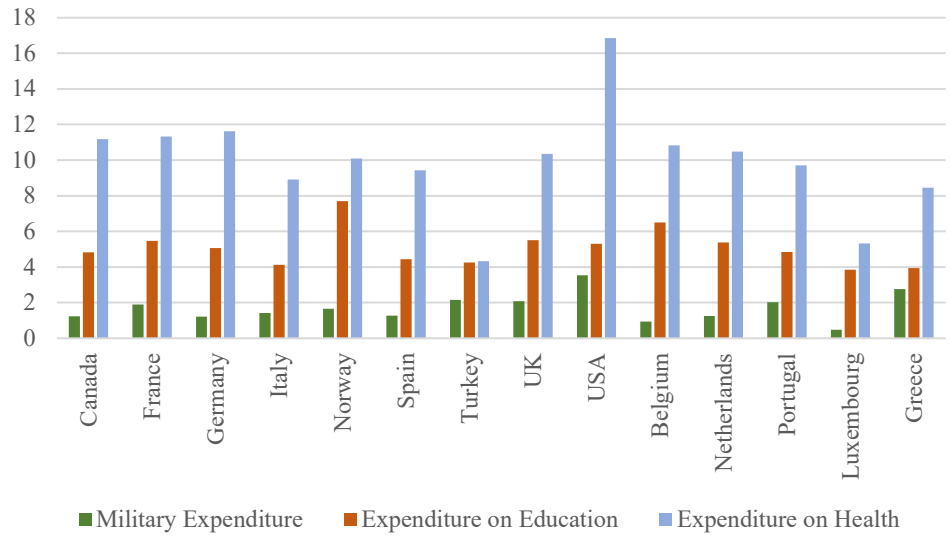


Figure 1 Average values of the analyzed variables (% of GDP)

The variable measuring the average size of military expenditure (as a share of GDP) is a key indicator for assessing the fulfilment of the political commitment of the Alliance countries to spend at least 2% of GDP on defense. From the assessment of the traditional NATO countries, it is clear that the highest value is observed in the case of the USA and Greece in the analyzed period. At the same time, it can be seen that most of the traditional countries have not fulfilled this commitment in the long term. Military expenditure higher than 2% is only evident in the case of the USA, Turkey, Greece, Portugal and the UK. In the case of education spending, it is evident that Norway has the highest share of GDP and Luxembourg the lowest. Thus, Norway is one of the countries with the highest expenditure on education (measured as a share of GDP), e.g. allocating more than 8% of GDP to education in 2020. The lowest value is seen in the case of Turkey in 2017 (3.5% of GDP). In the case of expenditure on health, the highest average value can be observed for the USA at 16% of GDP. The lowest value is seen in the case of Turkey, which spends less than 5% of GDP on health in the long term.

To analyze the relationship between military spending and education, as well as health spending and other variables, we employ panel data analysis methods. A concise overview of the employed models follows.

The *pooled* panel data model is given by

$$y_{it} = \alpha + \beta' X_{it} + u_{it}, \tag{1}$$

where $i = 1,2, \dots, n$ is the individual index (group, country, …), $t = 1,2, \dots, T$ is the time index and u_{it} is a random zero mean disturbance term, X_{it} is a $k \times 1$ vector of independent variables, β_{it} is a $k \times 1$ vector of parameters. This is a standard linear model *pooling* all data across i and t.

If one needs to model individual heterogeneity, it can be assumed that the error term has two separate components $u_{it} = \mu_i + \varepsilon_{it}$, where μ_i is specific to the individual and does not change over time. We get the model

$$y_{it} = \alpha + \beta' X_{it} + \mu_i + \varepsilon_{it}. \tag{2}$$

The error term ε_{it} is usually assumed independent of both the regressors X_{it} and the individual component μ_i. If the individual component is correlated with the regressors, it is customary to treat the μ_i as next n parameters to be estimated. This is called the *fixed effect* model (Hsiao 2014, Wooldridge 2002). Let us denote $\alpha_i = \alpha + \mu_i$, we obtain the model

$$y_{it} = \alpha_i + \beta' X_{it} + \varepsilon_{it}. \tag{3}$$

If the individual component μ_i is uncorrelated with the regressors, the model is termed *random effect*, μ_i are not treated as fixed parameters, but as random drawings from a given probability distribution.

When analysing panel data, we often encounter the problem of autocorrelation in residuals. One possible solution is the application of dynamic models. A *dynamic* linear panel data model can be written in the form

$$y_{it} = \rho y_{i,t-1} + \beta' X_{it} + \mu_i + \varepsilon_{it}. \quad (4)$$

The main idea on which the difference estimator is based is to get rid of the individual effect via differencing. First-differencing yields

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \beta' \Delta X_{it} + \Delta \varepsilon_{it}. \quad (5)$$

The error term $\Delta\varepsilon_{it}$ is autocorrelated and also correlated with lagged dependent variable $\Delta y_{i,t-1}$. Generalized method of moments (GMM) approach is used to get estimates of given equation, see Arellano and Bond (1991). Calculations were performed in R software using package "plm".

3 Results and discussion

The basic and control variables used in the models analyzing the relationship between military and social spending are following: Military expenditure, (Milex, % of GDP); Expenditure on education (Education, % of GDP); Expenditure on health (Health, % of GDP); Tax revenue (Tax, % of GDP); Households and NPISHs final consumption expenditure (Household, % of GDP); Gini index (Gini, %); Population (Population, Millions).

In analyzing the panel data, we initially employed fixed and random effects models. The Hausman test values (p-value < 0.001) were used to determine the selection of the fixed effects model (see Table 2). The model designated as *"Fixed effect - full"* encompasses all variables, whereas the *"Fixed effect"* model contains solely those that are statistically significant at the 0.05 level of significance. The R^2 values of these models are 0.324 and 0.322, respectively. The coefficients for the variables Education and Health are positive and statistically significant, indicating that a crowding-out effect is not present. However, an analysis of the residuals indicated that the residuals of the fixed effects model are autocorrelated. The Breusch-Godfrey/Wooldridge test, the Durbin-Watson test, and the Wooldridge test for panel data, see Croissant and Millo (2008), were utilized to determine the presence of autocorrelation. The p-values for all tests were found to be less than 0.05. Consequently, we implemented a dynamic GMM model. The parameter estimates of the full model and the final model are presented in Table 1. The parameter for the Education and Health variables is again positive; however, only the parameter for the Health variable is statistically significant at the 0.05 level. The parameter for the lagged Milex and Gini is still significant at the 0.10 level of significance.

Variable	Fixed effect-full	p-value	Fixed effect	p-value	GMM full	p-value	GMM	p-value
lag(Milex,1)					0.400356	0.2053	0.480189	*0.08562*
Education	0.16654	**0.01669**	0.15887	**0.01875**	0.11973	0.11957	0.06017	0.41320
Health	0.11796	**0.00116**	0.11685	**0.00121**	0.09273	*0.09326*	0.12225	**0.00234**
Tax	0.00997	0.61249			-0.00352	0.81101		
Household	-0.02598	**0.04231**	-0.02743	**0.02747**	-0.00538	0.73349		
Gini	0.00614	**0.00542**	0.00590	**0.00598**	0.00214	0.39717	0.00344	*0.09800*
Population	-0.02528	**0.02785**	-0.02560	**0.02526**	0.00480	0.77230		

Table 1 Estimated models

These findings offer several important implications. First, the absence of a statistically significant negative relationship between military and social spending suggests that the feared crowding-out effect does not materialize in the short to medium term for the analyzed NATO countries. In fact, the positive and significant coefficient for health expenditure in the GMM model implies that increases in military budgets may coincide with or even support higher health spending. This could reflect broader fiscal expansions or strategic complementarities, such as investments in military medical infrastructure that benefit the civilian sector. The positive but statistically insignificant coefficient for education in the final model suggests a more complex or delayed relationship. It is possible that education budgets are less sensitive to short-term shifts in defense spending or are protected by political or institutional mechanisms. The significance of the lagged military expenditure variable highlights the persistence of defense spending patterns, likely due to long-term procurement cycles and strategic commitments. Meanwhile,

the Gini index's marginal significance may indicate that income inequality plays a role in shaping public spending priorities, potentially influencing the balance between defense and welfare allocations.

4 Conclusion

The results obtained do not confirm a crowding-out effect between military and social spending (i.e. spending on health and education). Therefore, it can be concluded that the relationship between military and social spending is positive; as military spending increases, so does social spending. Furthermore, the results challenge the traditional "guns versus butter" trade-off and support the notion that, under certain fiscal and political conditions, NATO countries can simultaneously expand both defense and social sectors. However, it remains to be seen whether other factors indicated in the introduction to the article are at play here, and further research is needed to explore the long-term dynamics of models and the role of fiscal policy in mediating these relationships.

Acknowledgements

This article resulted from research supported by the Ministry of Defense of the Czech Republic named: Long-term Development of Organization - Land Operations (LANDOPS).

References

Arellano M, Bond S (1990) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The Review of Economic Studies. 58, pp 277–297.

Bauwens M A (2012) Guns versus Butter: A Budgetary Trade-Off in the MENA Region? Lund University: School of Economics and Management. Diploma thesis. pp 1-39.

Carter J, Ondercin H L, Palmer G (2021) Guns, Butter, and Growth: The Consequences of Military Spending Reconsidered. Political Research Quarterly, 74(1) pp 148–165. https://doi.org/10.1177/1065912919890417

Croissant Y, Millo G (2008) Panel Data Econometrics in R: The plm Package. Journal of Statistical Software, 27(2), pp 1–43.

Eichenberg R (1992) Do we know yet who pays for defense? Conclusion and synthesis. In Chan S, Mintz A. *Defense, Welfare and Growth* (pp 231-241). London Routledge.

Hsiao, Ch (2014) Analysis of Panel Data. 3rd edition. Cambridge University Press.

Mintz A, Huang Ch (1991) Guns Versus Butter: The Indirect Link. American Journal of Political Science, 35 (3), pp 738-757. http://www.jstor.org/stable/2111564

Mintz A, Huang Ch (1992) Guns Versus Butter: Conceptual and Methodological Issues. Policy Studies Review, 11 (3), pp 347-356. http://www.jstor.org/stable/2111564

Njifen I (2023) Military expenditures and human capital development in sub-Saharan Africa: a system GMM approach. Development Studies Research, 10 (1), pp 1-14. https://doi.org/10.1080/21665095.2022.2163678

Obinger H, Schmitt C (2011) Guns and Butter? Regime Competition and the Welfare State during the Cold War. World Politics, 63 (2), pp 246-270. https://www.jstor.org/stable/23018786

Russett B M (1969) Who Pays For Defense? The American Political Science Review, 63 (2), pp 412-426. https://www.jstor.org/stable/1954697

Russett B M (1982) Defense Expenditures and National Well-being. The American Political Science Review, 76 (4), 767-777. https://www.jstor.org/stable/1962969

Wang Y (2022) Revisiting the guns-butter trade off: a wavelet analysis of the US and Britain. Policy Studies, https://doi.org/10.1080/01442872.2022.2076822

Wooldridge, J M (2002) Econometric Analysis of Cross Section and Panel Data. Cambridge, Massachusetts: The MIT Press.

Received: 2025-06-05 Last revision: 2025-07-30 Accepted: 2025-07-30

Using Multidimensional Comparative Analysis to Assess the Health Risks of the Elderly Population

Pavla Jindrová[1], Hana Boháčová[2], Dominik Miškář[3]

Abstract. Multidimensional comparative analysis deals with the methods and techniques of comparing objects with multiple characteristics. This analysis can be widely applied in economics. In our paper, multidimensional comparative analysis will be used after assessing the health risks of the elderly population. Recently, the relevance of this topic has increased as people in all countries of the world are living to an increasingly older age. This trend is particularly evident in developed economies, including EU countries. If we focus on health risks, the situation of older people is different from that of younger people, because older people have been exposed to various, often harmful, influences and substances over a longer period, which have weakened the body's defenses and often caused chronic diseases. These health risks also place a heavy burden on healthcare systems, especially as the proportion of older people in the population increases in developed economies. The aim of this article is to identify and quantify risk factors associated with health status, mortality and health care costs for the elderly population in the Czech Republic and selected EU countries using multidimensional comparative analysis. We also use regression and cluster analysis.

Keywords: elderly population, health risks, multidimensional comparative analysis.
JEL Classification: C38, I14

AMS Classification: 62P20, 62H30

1 Introduction

We can see an ageing population in all developed countries. This is the result of better living conditions, lifestyles and health care, as well as greater social protection. The proportion of elderly people in the population is increasing. According to OECD (2024) the proportion of people aged over 65 in the EU has risen from 16% in 2000 to 21% in 2023 and is projected to reach 29% by 2050. The share of people aged over 65 in 2023 was particularly high in Italy and Portugal with nearly 25% of the population in that age group, while Ireland and Luxembourg had the lowest proportion with 15%. In the Czech Republic the share of elderly population has risen from 14% in 2000 to 20% in 2023 and is projected to reach 27% by 2050.

The health risks faced by the elderly population are, in some ways, specific to them and different from those faced by younger people. Their health status develops differently from that in earlier years of life as a result of long-term exposure to various negative influences.

In this paper, we will analyze selected health risk factors of the elderly population through multivariate comparative analysis, regression and cluster analysis.

2 Elderly population and their health risks

The elderly population is defined as the population aged 65 and over. This group of people faces specific health risks. The ageing population is the most important factor in their health risks. This is related to better living conditions and lifestyles, as well as the higher quality and availability of healthcare and financing methods, and greater social protection.

[1] University of Pardubice, Faculty of Economics and Administration, Institute of Mathematics and Quantitative Methods, Studentská 95 532 10 Pardubice, Czech Republic, Pavla.Jindrova@upce.cz.
[2] University of Pardubice, Faculty of Economics and Administration, Institute of Mathematics and Quantitative Methods, Studentská 95 532 10 Pardubice, Czech Republic, Hana.Bohacova@upce.cz.
[3] University of Pardubice, Faculty of Economics and Administration, Institute of Mathematics and Quantitative Methods, Studentská 95 532 10 Pardubice, Czech Republic, Dominik.Miskar@upce.cz.

The importance of this issue is illustrated by several papers that address these selected issues. For example (Pascual-Saez et al., 2017), (Pacáková and Kopecká, 2018), (Jindrová, 2019), (Oliver et al. 2021), (Kocot et al., 2022), (Jindrová and Boháčová, 2023), (Kim et al., 2023).

The quality and availability of health care are other factors related to the health risks of the elderly population. A healthy lifestyle is important for good health too. In this area, the risk factors include smoking, alcohol consumption, poor diet and obesity and lack of physical activities. Aging of the people also influences the main causes of death which change with increasing age. For this age category, deaths due to diseases of the circulatory system predominate mostly. Cancer is in second place. In the older population the dementia is one of the main causes of death. Elderly people were most at risk of dying from Covid-19, especially during pandemic, but also after it. Older age is also associated with a higher risk of death from chronic diseases such as diabetes mellitus.

3 Methods and data

According to the goals of analysis we have used multidimensional comparative analysis, Spearman's correlation and cluster analysis.

The data from matrix $\boldsymbol{X} = (X_{ij})$ are analyzed for $i = 1, \ldots, n$ and for $j = 1, \ldots, m$, when n is number of countries (in our case $n = 27$) and m is number of variables in the main group of variables (in our case $m = 15$). At the beginning of multidimensional comparative analysis, the type of each variable X_j, $j = 1, \ldots, 15$ should be defined, which means that the first step is to identify which variables are stimulants and which are destimulants. A variable is considered a stimulant if higher values have a positive effect on the analyzed processes. A variable is called a destimulant if lower values are considered more favorable. The second step of our analysis is to normalize the variables, as they are measured in different units. After this process, the normalized variables are comparable. Normalization has been performed according to formula (1) for stimulants and to formula (2) for destimulants:

$$b_{ij} = \frac{\max\limits_j x_{ij} - x_{ij}}{\max\limits_j x_{ij} - \min\limits_j x_{ij}} \tag{1}$$

$$b_{ij} = \frac{x_{ij} - \min\limits_j x_{ij}}{\max\limits_j x_{ij} - \min\limits_j x_{ij}} \tag{2}$$

The synthetic indicator for each of the *i-th* country has been calculated as the average of the b_{ij}. According to the formulas (1), (2) obviously implies that the lower the value of synthetic variable, the higher the level of the multidimensional object.

The match of the ordering of the countries by each pair of synthetic variables can be quantified using Spearman's rank correlation coefficient. The objective of cluster analysis is to classify multidimensional objects into a small number of mutually exclusive groups based on the similarities among the objects.

According to the goals of this paper, Eurostat data (2025) for the most recent year was used to compare the situation in EU countries. 15 variables have been used. They have been divided into four main groups which correspond with the health risk factors for elderly population – health status, health care, risk factors, and selected main causes of death for elderly population. Here is a list of these used variables:

Health status:

HS1 - Life expectancy at the absolute value of 65 (2022)

HS2 - Healthy life years at the absolute value of 65 (2022)

Health care:

HC1 - Health care expenditure (total); Purchasing power standard (PPS) per inhabitant (2022)

HC2 - Physicians (practicing); Per hundred thousand inhabitants (2022 or nearest)

HC3 - Nurses (practicing); Per hundred thousand inhabitants (2022 or nearest)

Risk factors:

RF1 - Obese by body mass index (BMI), 65 years or over; Percentage (2019)

RF2 - Everyday alcohol consumption, 65 years or over; Percentage (2019)

RF3 - Daily smoker of tobacco products, 65 years or over; Percentage (2019)

RF4 - Daily consumption of fruit and vegetables - 0 portions, 65 years or over; Percentage (2019)

RF5 - Performing health-enhancing physical activity, 65 years or over; Percentage (2019)

Causes of death:

CD1 - Diseases of the circulatory system - 65 years or over; Standardized death rate (2022)

CD2 - Malignant neoplasms - 65 years or over; Standardized death rate (2022)

CD3 - COVID-19, virus identified, 65 years or over; Standardized death rate (2022)

CD4 - Dementia - 65 years or over; Standardized death rate (2022)

CD5 - Causes of death - Diabetes mellitus - 65 years or over; Standardized death rate (2022)

4 Results and discussion

Multidimensional comparative analysis has been used for comparing the situation among EU countries via health status of elderly people, health care, risk factors, and selected main causes of death of elderly people.

The synthetic variable allows us to replace the set of variables into one aggregated variable. By this it allows us to transform multidimensional problems into one dimensional. We construct four synthetic variables SV1, SV2, SV3, and SV4 – each for four main areas above. The synthetic variable SV1 made up of 2 indicators of health status for elderly people, when variables HS1-HS2 were stimulants. The synthetic variable SV2 made up of 3 indicators of health care, when all these 3 indicators were stimulants. The synthetic variable SV3 made up of 5 indicators of risk factors of elderly people, where RF1-RF4 were destimulants and RF5 was stimulant. SV4 made up of 5 variables of selected main causes of death of elderly people and all of them were destimulants. These synthetic variables will make it possible to compare the monitored countries and arrange them from the best to the worst situation in the monitored areas.

Based on the values of the relevant synthetic variable $SV_j = (y_{ij})$, for $i = 1, \ldots, 27$ and $j = 1, 2, 3, 4$, the countries were grouped according to their quality of life with respect to this area. The groups of countries were formed as follows (built using mean $\bar{y}_j$ and standard deviation s_j of relevant synthetic variable):

- high quality of area: $y_{ij} \leq \bar{y}_j - s_j$
- better than medium quality of area: $\bar{y}_j - s_j < y_{ij} \leq \bar{y}_j$
- lower than medium quality of area: $\bar{y}_j < y_{ij} \leq y_j + s_j$
- the lowest quality of area: $\bar{y}_j + s_j < y_{ij}$

The classification of EU countries based on the value of synthetic variable SV1 (Health status) is shown in Figure 1.

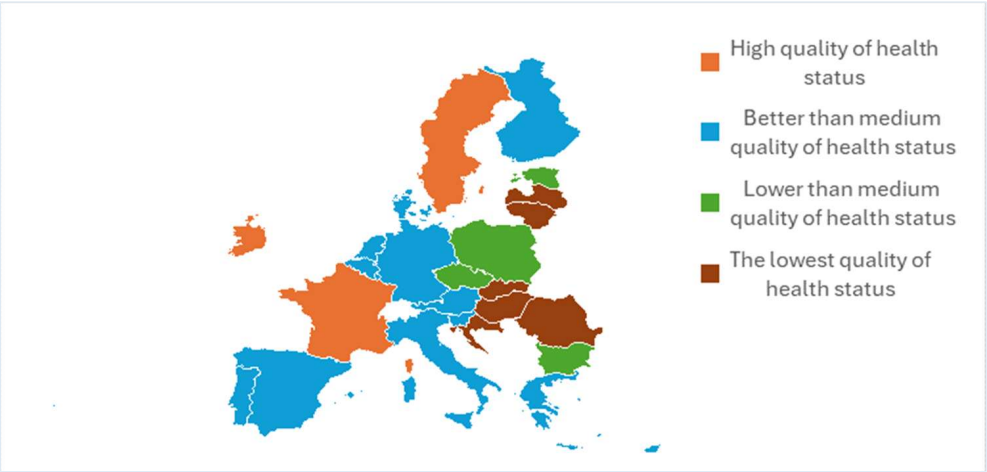


Figure 1 The classification of EU countries based on the value of synthetic variable SV1 (Health status)

In Figure 1 we can see that Sweden, France, Malta, and Ireland are the countries with the highest quality of health status for people at 65. Second group of countries with better than medium quality of health status for people at 65 contains the remaining EU countries that do not belong to the post-socialist bloc and Slovenia. Post-socialist countries, except Slovenia, are classified as the countries with lower than medium quality of health status, i.e. Estonia, Poland, Czechia, and Bulgaria, and the countries with the lowest quality of this analyzed area, i.e. Lithuania, Hungary, Croatia, Slovakia, Latvia, and Romania.

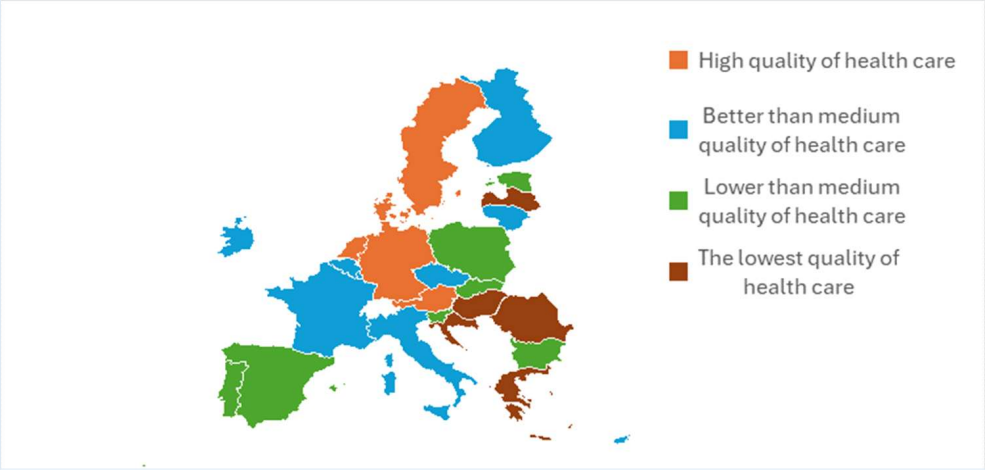


Figure 2 The classification of EU countries based on the value of synthetic variable SV2 (Health care)

The classification of EU countries based on the value of synthetic variable SV2 (Health care) is shown in Figure 2. The best situation in the health care area among EU countries can be seen in Austria, Germany, Sweden, Denmark, and the Netherlands. Czechia, together with Belgium, Ireland, Malta, Luxembourg, Cyprus, Finland, France, Italy, and Lithuania, belongs to the group of countries with better than medium quality of health care. Countries with lower than medium quality of health care are Spain, Bulgaria, Slovenia, Portugal, Estonia, Poland, and Slovakia. Croatia, Hungary, Latvia, Greece, and Romania are the countries with the lowest quality of health care.

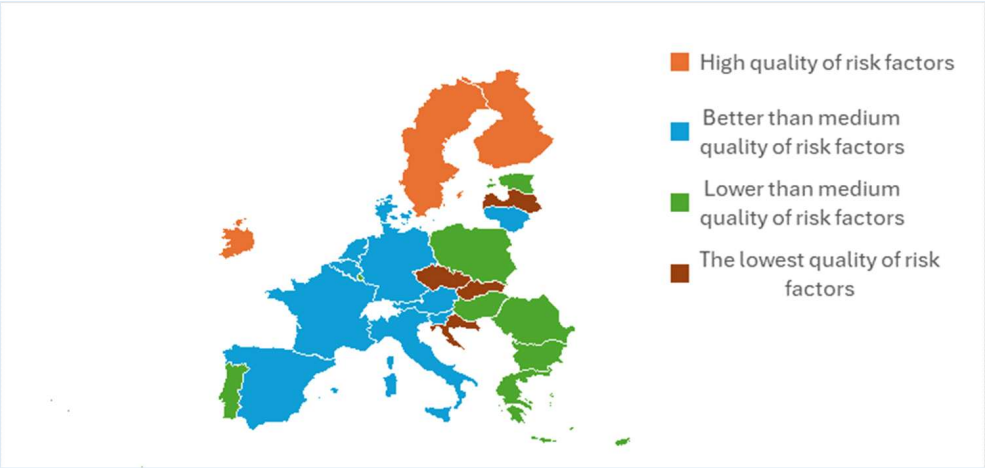


Figure 3 The classification of EU countries based on the value of synthetic variable SV3 (Risk factors)

The classification of EU countries based on the value of synthetic variable SV3 (Risk factors) is shown in Figure 3. The high quality of risk factors can be seen in Sweden, Finland, and Ireland. Czechia, together with Croatia, Slovakia, Latvia, and Malta, belong to the group of countries with the lowest quality of the situation in risk factors for elderly population.

The classification of EU countries based on the value of synthetic variable SV4 (Causes of death) is shown in Figure 4. The best situation in EU countries via selected main causes of death of elderly population can be seen in France, Luxembourg, Belgium, and Spain. Czechia belongs to a group of countries with lower than medium quality of this area.

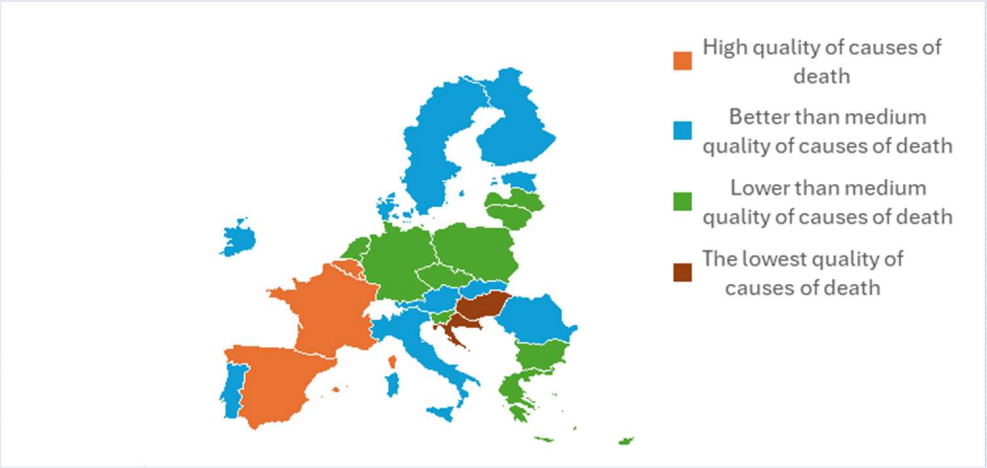


Figure 4 The classification of EU countries based on the value of synthetic variable SV4 (Causes of death)

Spearman's correlation coefficient was used to test the relationships between pairs of synthetic variables SV1-SV4. Its values are statistically significant for $\alpha = 0.01$ between all pairs of variables except the pairs SV2 (Health care) and SV4 (Causes of death) (p-value = 0.1161) – see Tab. 1.

	SV1 (Health status)	SV2 (Health care)	SV3 (Risk factors)	SV4 (Causes of death)
SV1 (Health status)	1.0000	0.6398	0.5690	0.6172
SV2 (Health care)	0.6398	1.0000	0.6270	0.3095
SV3 (Risk factors)	0.5690	0.6270	1.0000	0.5299
SV4 (Causes of death)	0.6172	0.3095	0.5299	1.0000

Table 1 Table of Spearman's correlation coefficients

The synthetic variables SV1-SV4 were analyzed using cluster analysis. This analysis revealed similarities among EU countries in our four analyzed areas. Ward's clustering method with Euclidean distances was used. Figure 5 shows the results of this analysis in dendrogram.

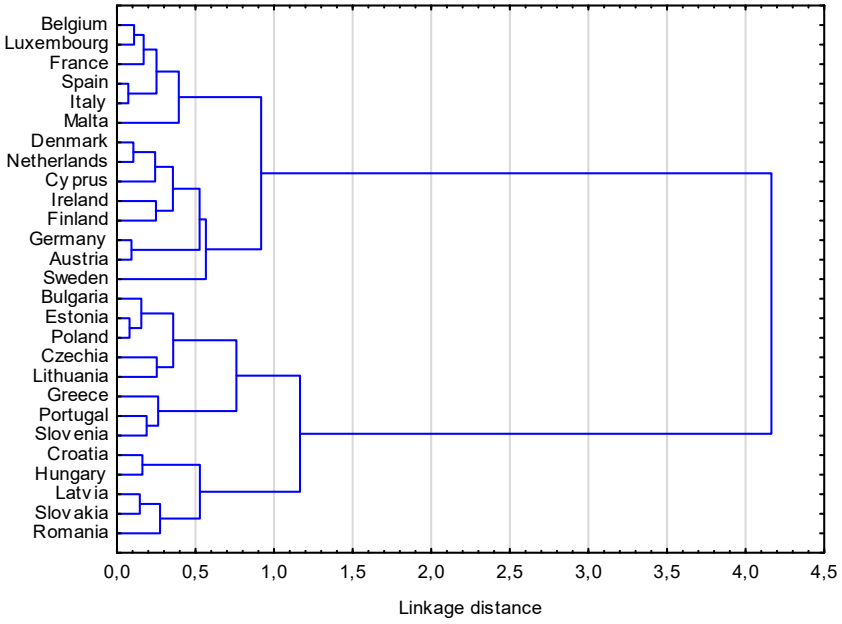


Figure 5 Dendrogram

From the dendrogram in Figure 5, it is evident that the countries are divided into two significantly different clusters. The first cluster consists of the countries from the upper part of the figure - these are the countries of Western and Southern Europe (without Greece and Portugal) and the Scandinavian countries. If we look for these countries on the cartograms in Figures 1, 2, 3 and 4, then with some exceptions they are for the most part from the groups of countries with high or better than medium quality of analyzed areas - classified according to the values of four synthetic variables. The second cluster contains the countries of Central and Eastern Europe, together with Greece and Portugal. On the other hand, these countries are mostly included in the groups of countries with lower than medium or the lowest quality of analyzed areas in the cartograms. These mean bigger values of synthetic variables.

5 Conclusion

The situation regarding health risk factors for the elderly population in EU countries was analyzed using 15 variables. Multivariate analysis was used to analyze differences in the four areas under study: health status, healthcare, risk factors and the main causes of death in the elderly population. Spearman's correlation coefficient was used to verify the dependence among these areas. Cluster analysis was used to compare the situation in the analyzed areas across EU countries. Our results confirmed the significant differences between post-socialist countries and other EU members. These differences can be characterized by the original EU member states showing better values within all four synthetic variables - better health status, more effective healthcare, lower incidence of risk factors in the population of people 65 years old and over, and lower mortality from selected diseases.

References

Eurostat database (2025) (Accessed on 10 May 2025). https://ec.europa.eu/eurostat/web/main/data/database

Jindrová P (2019) Impact of health expenditures and health resources on health status in the European union members states. 13th International Days of Statistics and Economics. 607-616. DOI: 10.18267/pr.2019.los.186.0

Jindrová P, Boháčová H (2023). Assessing the health system efficiency by using avoidable mortality indicators in European countries, 2019. Regional Statistics. 13(6): 1124-1144. DOI:10.15196/RS130605

Kim J, Petalcorin CC, Park D, Tian S(2023) Determinants of the Elderly Share of Population: A Cross-Country Empirical Analysis. Soc Indic Res 165, 941–957. DOI:10.1007/s11205-022-03052-y

Kocot E, Denkowska S, Fijorek K. (2022). Are Measures of Health Status for the Total Population Good Proxies for the Health of the Older Population in International Comparison Studies?. *International Journal of Environmental Research and Public Health*, *19*(13), 7559. DOI: 10.3390/ijerph19137559

OECD/European Commission (2024), Health at a Glance: Europe 2024: State of Health in the EU Cycle, OECD Publishing, Paris, https://doi.org/10.1787/b3704e14-en

Oliver A, Sentandreu-Mañó T, Tomás JM, Fernández I, Sancho P (2021) Quality of life in European older adults of SHARE wave 7: Comparing the old and the oldest-old. Journal of Clinical Medicine. 10(13):2850.. 10(13). DOI:10.3390/jcm10132850

Pacáková V, Kopecká L (2018) Inequalities in Health Status Depending on Socio-Economic Situation in the European Countries. E+M Ekonomie a Management. 21(2): 4-20. DOI:10.15240/tul/001/2018-2-001

Pascual-Saez M, Cantarero-Prieto D, Castañeda D (2017) Public health expenditure, GDP and the elderly population: a comparative study. International Journal of Social Economics. 44(10): 1390-1400. DOI:10.1108/IJSE-03-2016-0106

Received: 2025-06-13 Last revision: 2025-07-11 Accepted: 2025-07-15

Determinants of the Herfindahl-Hirschman Index Affecting Market Power and Concentration in the Czech Insurance Sector

Jana Heckenbergerová[1], Simona Tučková[2]

Abstract. This paper examines the determinants of the Herfindahl-Hirschman Index (HHI) in the Czech insurance industry, focusing on market power and concentration. The HHI values describe market concentration in the Czech insurance industry and reveal a moderately concentrated market with significant players. The 2012-2022 dataset from the Czech Association of Insurance Companies was selected for numerical analyses. Utilizing econometric methods, we assess the impact of key variables. They are divided into two categories: a) insurance sector-related indicators such as the number of insurance companies, claims settled, financial positioning, etc.; b) macroeconomic indicators such as GDP, inflation, government debt, etc. The findings suggest that factors such as the number of insurance companies and financial placement significantly affect market concentration. The paper concludes with recommendations for regulatory and strategic initiatives to increase competition and market efficiency in the Czech insurance industry.

Keywords: Herfindahl-Hirschman Index, market power and concentration, correlation analysis, economic indicators

JEL Classification: G22, L11, L13, C23
AMS Classification: 91B82, 91B70

1 Introduction – Insurance industry

The insurance industry is an integral part of today's developed world's market economy. It is a sector of the economy characterized by the elimination of financial risk. The meaning of the term "insurance" differs across countries. In the Czech Republic (Ducháčková, 2009), it refers to insurance, reinsurance, and intermediation activities in the field of commercial insurance, as well as related activities. The insurance industry has an irreplaceable place in a market economy. Not only the economy, but all activities of all economic entities are affected by risks and insurance (Vávrová, 2014). It is very important for the economy that economic entities, especially companies and investors, are willing to take risks. This allows the economy to innovate and develops the dynamism of the economy. The meeting of supply and demand for insurance protection takes place in the insurance market. In the insurance market, insurance companies offer their specific monetary services in the form of insurance and clients demand insurance protection. Insurance companies protect households and businesses through their insurance activities by reducing the financial consequences of risks. They thus contribute to maintaining the financial stability of businesses and a certain standard of living for individuals, thereby supporting overall economic stability. Commercial insurance companies are required by law to create technical reserves, which are temporarily free funds from premiums received that are intended to cover liabilities that insurance companies know or estimate and will have to pay in the future in the form of insurance claims (Böhm and Mužáková 2010). With these funds, they enter the capital market as significant investors, thereby increasing economic activity.

1.1 Market power and concentration

In order to promote healthy economic development, support competitiveness, ensure financial stability, and increase consumer protection, it is beneficial to examine the market power operating in a selected sector. Market power is the ability of a firm or organisation to influence the price and quantity of goods or services in the market. An insurance company that has market power is in a position to distort competition (Nejezchleb and Hajná, 2013). Another possible definition of a firm's market power is the level of control over the price of output (Kolář and Veselá, 2006). According to classical economics, market power is when a firm - an insurance company - can price

[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, simona.tuckova@upce.cz

discriminate or set prices above marginal cost. Indicators of market power are by concentration, large market share and the number of firms operating in a given market (Zemplinerová, 1999). On the basis of these indicators, the activities of insurance companies can be classified as imperfect competition, specifically as oligopoly in terms of market power (Brčák et al., 2020). Several indicators are used to measure market power, Nejezchleb and Hajná (2013) point out that it is advisable to take a comprehensive approach when examining market power, the interpretation of one indicator may not lead to the correct conclusion. In addition to the basic indicators, it is also useful to look at the evolution of indicators over time, barriers to entry into the industry, what regulation exists in a given market, and so on. The most commonly used indicators to measure market power are **market share, concentration ratio (CR) and the Herfindahl-Hirschman index (HHI)**.

1.2 Herfindahl-Hirschman Index

The Herfindahl-Hirschman Index (HHI) is named after two authors, O. C. Herfindahl and A. O. Hirschman, who independently used it to calculate market concentration.

The formula for the calculation is as follows (Hirschman, 1945; Herfindahl, 1950):

$$HHI = \sum_{i=1}^{n} s_i^2 \qquad (1)$$

where s_i represents the market share of the i-th insurance company in percent and n is the total number of insurance companies on the market.

HHI represents the sum of the squares of the market shares of all insurance companies. As Naldi and Flamini (2014) mention, the calculation of HHI is considerably more difficult than the calculation of the previous indicator, mainly because it is necessary to know the market shares of all market participants in a given market. According to calculated HHI values, market concentration can be classified into three categories. Following the most recent domestic framework proposed by Brčák et al. (2020): a) HHI values below 1000 indicate an unconcentrated market; b) values between 1000 and 1800 reflect moderate concentration; c) values above 1800 signal a highly concentrated market. Nevertheless as highlighted in Brezina et al. (2014), the classification varies across jurisdictions and academic sources. For instance, the U.S. Federal Trade Commission (2010) defines moderately concentrated markets as those with HHI values between 1500 and 2500, while the European Commission (2004) adopts a range of 1000 to 2000.

2 Analyses of Czech Insurance Sector

For the purposes of the analysis of the Czech insurance market, a literature review of foreign studies was first carried out, an overview of which is shown in Table 1.

Country and authors	Year of study	Type of market/research specifications	Subject of research	Results of the study
12 EU Countries (Berry-Stölzle et al., 2011)	2003 - 2007	Property and liability insurance	Testing the hypotheses: structure-behavior-performance (SCP), relative market power (RMP), efficient structure (ES)	Support for the efficient structure (ES) hypothesis, more efficient firms can charge lower prices than less efficient competitors, allowing them to gain higher market share and higher profits, a win-win situation for consumers and firms
24 EU Countries (Janků and Badura, 2021)	1997 - 2019	Non-life insurance	Impact of market concentration on the claims ratio (financial stability of the industry)	A non-linear relationship between market concentration and the claims ratio has been demonstrated - as concentration increases, the claims ratio decreases (positive process)

Jordan and Saudi Arabia (Almajali and Zainuddin, 2023)	2009 - 2019	The entire insurance market	Testing the hypothesis: increased market concentration increases market profitability due to cooperation between dominant firms	The more insurance companies increase their market share, the more competitive they will be and their profits will increase
Thailand (Sukpaiboonwat, 2014)	2006-2011	Life and non-life insurance	Assessing market concentration and the impact of financial statements on HHI and CR values	The life insurance market is more concentrated than the non-life insurance market
USA (Shim, 2017)	1992 - 2010	Property and liability insurance	Assessment of the market concentration - insurer financial stability relationship	A highly concentrated market has negative effects on an insurer's financial stability, referred to as "concentration-fragility"
USA (Sharma and Rotthoff, 2019)	2001 - 2012	Life, property and health insurance	Finding the relationship: market concentration - jobs, market concentration - wages, market concentration - total premiums, market concentration - annual wages in the insurance industry	Market concentration - jobs: statistically significant negative relationship Market concentration - wages: statistically significant negative relationship Market concentration - total premiums: statistically significant positive relationship Market concentration - annual wages: no statistically significant relationship found

Table 1 Results of the literature review

Many authors addressed the topic of market power in the insurance market several decades ago. Based on the results of the literature review in Table 1, the most common indicator of market concentration was the Herfindahl-Hirschman Index, which will be used for the following analysis in the paper and the concentration coefficient will be used as an additional indicator.

2.1 Datasets and indicators

The analysis of market concentration within the Czech insurance sector was conducted using the Herfindahl-Hirschman Index (HHI), calculated on the basis of annual market shares derived from gross written premiums (Vlková, 2024). The underlying data were sourced from the annual reports of the Czech Insurance Association (ČAP), which aggregates information from insurers representing over 98% of the domestic market. The study covers the period from 2012 to 2022, during which the number of ČAP member insurers declined from 32 to 23, reflecting ongoing consolidation within the sector. The HHI was computed as the sum of squared market shares of all insurers operating in the market, providing a robust measure of market concentration.

The results illustrated in Figure 1 indicate that the Czech insurance market exhibits characteristics of a moderately concentrated structure throughout the observed period, with HHI values consistently ranging between 1167 and 1591. The average HHI of 1339 places the market within the "moderately concentrated" category as defined by standard antitrust thresholds. Notably, a marked increase in concentration was observed between 2018 and 2020, coinciding with a reduction in the number of active insurers and several significant mergers, including the integration of Generali and Česká pojišťovna. By 2022, the top four insurers accounted for nearly 69% of the market, while the top ten controlled over 91%, suggesting a growing dominance of a few key players. Despite this trend, the market has not yet reached the threshold of high concentration, although the trajectory warrants continued regulatory attention.

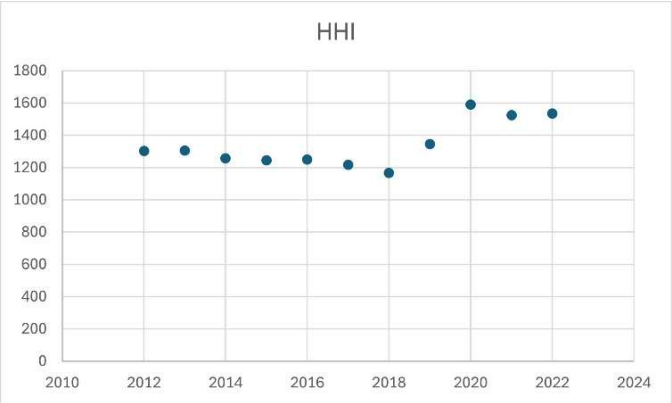


Figure 1 Herfindahl-Hirschman Index 2012-2022 dataset

The correlation analysis incorporates six insurance sector-specific variables that are hypothesized to influence market concentration, as measured by the Herfindahl-Hirschman Index (HHI). These include:

(1) the number of insurance companies operating in the market (NICo), reflecting the structural competitiveness of the sector;

(2) the number of settled insurance claims (NICl), which serves as a proxy for market activity and operational scale;

(3) insurers' net profit or loss for the accounting period (ICoP), capturing financial performance and potential market dominance;

(4) the volume of technical reserves (TR), representing the financial obligations insurers must cover and their capacity for long-term investment;

(5) the insurance coverage rate (ICv), defined as the ratio of gross written premiums to GDP, indicating the relative significance of the insurance sector within the national economy;

and (6) the total volume of financial placements resp. financial positioning (FP), which reflects the investment behavior and asset management strategies of insurers.

These variables collectively provide a comprehensive view of the operational, financial, and structural dimensions of the insurance industry. The Czech insurance market over the period 2012–2022 exhibited several structural and financial dynamics, as captured by six sector-specific indicators illustrated in Figure 2. First, the number of active insurance companies declined steadily, a trend attributable not only to market exits but also to consolidation through mergers and acquisitions — notably, the integration of Pojišťovna České spořitelny into Kooperativa and the absorption of AXA by UNIQA. Second, the volume of settled insurance claims fluctuated within a range of approximately 2,6 to 3 million annually, with a peak in 2014. A notable contraction occurred in 2020 and 2021, plausibly linked to pandemic-related mobility restrictions, which likely suppressed claim incidence, particularly in motor and travel insurance.

Third, the net profit of insurers displayed a generally upward trajectory, with exceptional peaks in 2019 and 2021. The latter was partially driven by extraordinary intra-group transactions, as reported by the Czech Insurance Association (ČAP). Fourth, technical reserves — a key measure of insurers' long-term solvency and investment capacity — remained relatively stable, averaging around CZK 330 billion, though a decline to CZK 314 billion was observed in 2022. Fifth, the insurance penetration rate (premiums as a percentage of GDP) declined from 3,8% in 2012 to 2,9% in 2022, reflecting a structural weakening of the life insurance segment. Finally, the total volume of financial placements — representing insurers' investment portfolios — peaked in 2018 at CZK 345 billion but subsequently contracted to CZK 269,8 billion by 2022. This decline may be attributed to inflationary pressures, increased claims payouts, or shifts in asset allocation strategies.

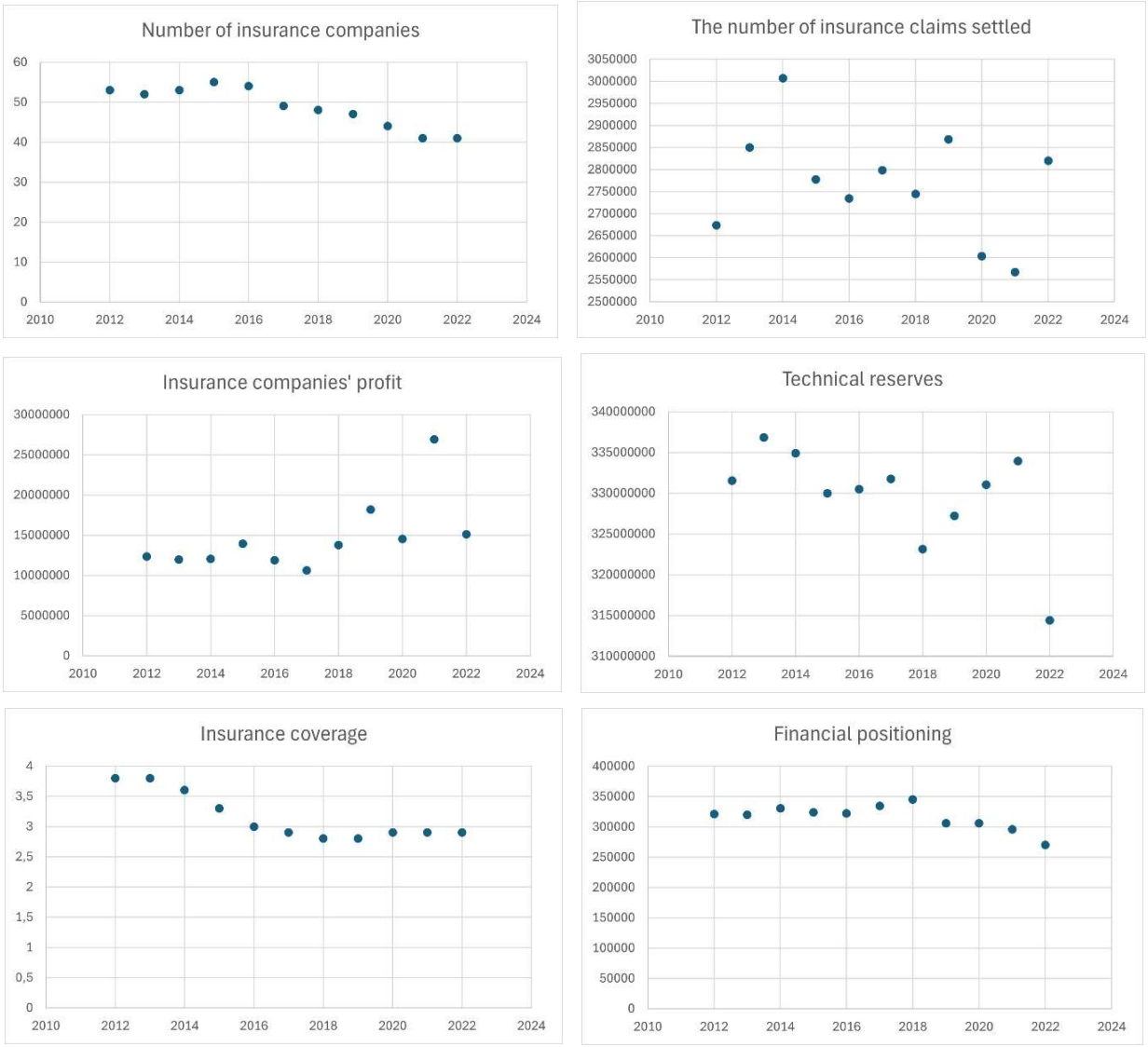


Figure 2 Insurance sector-related indicators

In addition to sector-specific metrics, the model incorporates three macroeconomic variables to account for broader economic conditions that may affect market concentration. These are:

(1) gross domestic product (GDP), which captures the overall size and performance of the national economy and may influence both demand for insurance and the scale of market operations;

(2) the inflation rate (Infl), which affects the real value of premiums and claims, and may alter consumer behavior and insurer profitability;

and (3) the level of the government debt (GoD), which serves as an indicator of fiscal sustainability and macroeconomic stability, potentially influencing investor confidence and regulatory priorities.

These macro-level indicators are essential for contextualizing the insurance market within the broader economic environment and for identifying systemic factors that may drive concentration dynamics. The macroeconomic environment in the Czech Republic over the period 2012–2022 was characterized by a combination of cyclical fluctuations and structural shifts, which are reflected in the evolution of three key indicators: gross domestic product (GDP), inflation, and government debt. Real GDP exhibited a generally upward trend, with the exception of a contraction in 2020, attributable to the COVID-19 pandemic and associated economic disruptions. By 2022, GDP had recovered and surpassed pre-pandemic levels, reflecting both domestic demand resilience and external trade recovery.

Inflation remained relatively stable and within the Czech National Bank's target range until 2019, after which it began to accelerate. The most pronounced surge occurred in 2022, when inflation reached 15,1%, driven by global supply chain disruptions, energy price shocks, and geopolitical tensions following the Russian invasion of Ukraine.

This inflationary spike marked the highest annual rate since the early 1990s and had significant implications for real income, consumption patterns, and monetary policy.

Public debt, measured in absolute terms, remained relatively contained until 2019, after which it increased sharply. The debt stock rose from approximately CZK 1,6 trillion in 2016 to nearly CZK 2,9 trillion in 2022. This escalation was primarily the result of expansive fiscal measures adopted in response to the pandemic, including income support schemes and healthcare expenditures, as well as subsequent structural deficits. The trajectory of public debt underscores the growing fiscal pressures facing the Czech government and the potential implications for long-term macroeconomic stability.

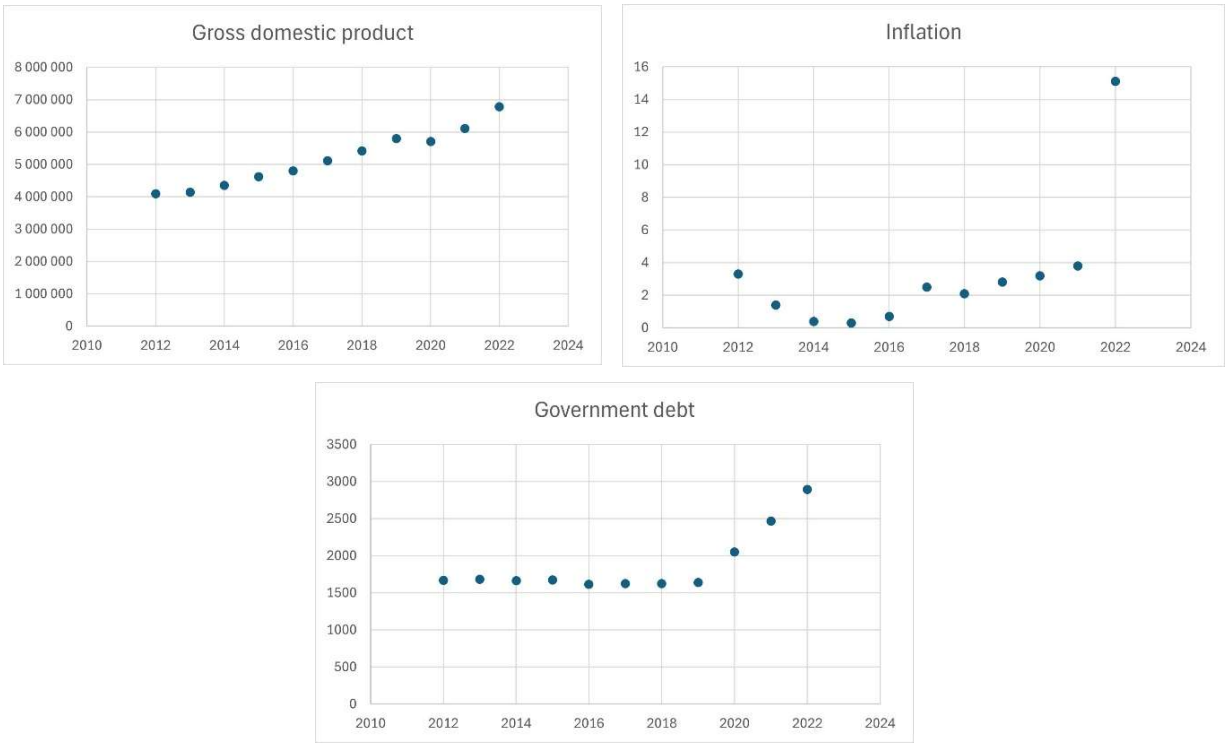


Figure 3 Macroeconomic indicators

2.2 Correlation analysis of HHI and Insurance sector-related indicators

The relationship between two quantitative variables can be assessed using correlation analysis, which provides insight into the strength and direction of association. The Pearson correlation coefficient (denoted as R) quantifies the degree of linear dependence between two continuous variables, under the assumption that bivariate distribution is approximately normally distributed. In contrast, the Spearman rank correlation coefficient (R_s) evaluates the strength of a monotonic relationship and is computed based on the ranked values of the variables. While R captures linear association, R_s is a non-parametric measure that does not require any distributional assumptions, making it suitable for data that deviate from normality or contain outliers.

Both coefficients are dimensionless and bounded within the interval [−1, 1]. A value of −1 indicates a perfect negative association, whereas a value of 1 signifies a perfect positive association. A coefficient of 0 implies the absence of a statistically detectable linear (in the case of R) or monotonic (in the case of R_s) relationship. However, it is important to emphasize that a zero correlation does not necessarily imply statistical independence, as non-linear or more complex forms of dependence may still be present (Walker, 2013).

Before proceeding with correlation analysis, it is necessary to assess whether the Herfindahl-Hirschman Index (HHI) and the selected explanatory indicator jointly a bivariate normal distribution. This verification was carried out using the Mardia test for multivariate normality, implemented in the statistical software PAST. The null hypothesis (H_0) assumes that the joint distribution of HHI and the selected indicator is bivariate normal, while the alternative hypothesis (H_1) posits a deviation from normality. The test results are interpreted with respect to two significance levels: at $\alpha = 0{,}05$, the null hypothesis is weakly rejected, indicating marginal evidence against normality; at $\alpha = 0{,}01$, the rejection is considered strong, suggesting substantial deviation from the assumption of normality. The outcomes of the Mardia tests for the insurance sector-related indicators are presented in Table 2.

Indicator	Mardia tests	p-value	normality decision
NICo	Skewness	0,7321	cannot be rejected
	Sample corrected	0,5437	cannot be rejected
	Kurtosis	0,2004	cannot be rejected
	Doornik and Hansen omnibus	0,1748	cannot be rejected
NICl	Skewness	0,8224	cannot be rejected
	Sample corrected	0,6757	cannot be rejected
	Kurtosis	0,1944	cannot be rejected
	Doornik and Hansen omnibus	0,6989	cannot be rejected
ICoP	Skewness	0,1202	cannot be rejected
	Sample corrected	0,02472	**weakly rejected**
	Kurtosis	0,9505	cannot be rejected
	Doornik and Hansen omnibus	0,004953	**strongly rejected**
TR	Skewness	0,4258	cannot be rejected
	Sample corrected	0,2076	cannot be rejected
	Kurtosis	0,6723	cannot be rejected
	Doornik and Hansen omnibus	0,1508	cannot be rejected
ICv	Skewness	0,6136	cannot be rejected
	Sample corrected	0,3946	cannot be rejected
	Kurtosis	0,1065	cannot be rejected
	Doornik and Hansen omnibus	0,1211	cannot be rejected
FP	Skewness	0,1197	cannot be rejected
	Sample corrected	0,02454	**weakly rejected**
	Kurtosis	0,8454	cannot be rejected
	Doornik and Hansen omnibus	0,004183	**strongly rejected**

Table 2 Bivariate normality test of HHI and insurance related indicators

Normality was violated for two indicators: insurers' net profit or loss for the accounting period (ICoP) and financial placements (FP). Consequently, the strength of their association with the Herfindahl-Hirschman Index (HHI) was assessed using the non-parametric Spearman rank correlation. For the remaining indicators — NICo, NICl, TR, and ICv — no significant deviation from normality was detected, allowing for the application of Pearson's correlation coefficient. Statistical significance was evaluated using *p*-values, with values below 0,05 indicating a significant correlation. Numerical results from the PAST software are presented in Table 3. The strength of correlation is interpreted according to Evans (1996): values between 0,00–0,19 are considered very weak, 0,20–0,39 weak, 0,40–0,59 moderate, 0,60–0,79 strong, and 0,80–1,00 very strong.

Indicator	Correlation type	Correlation coefficient	p-value	dependence description
NICo	Pearson	-0,78352	0,0043301	negative, strong, significant
NICl	Pearson	-0,46833	0,14626	negative, moderate, not significant
ICoP	Spearman	0,62727	0,038845	positive, strong, significant
TR	Pearson	-0,20555	0,54428	negative, weak, not significant
ICv	Pearson	-0,26842	0,42483	negative, weak, not significant
FP	Spearman	-0,91818	0,000066615	negative, very strong, significant

Table 3 Correlation coefficients and power of dependence of HHI and insurance related indicators

Among the insurance sector indicators, the number of insurance companies (NICo) exhibits a strong negative association with market concentration, which aligns with theoretical expectations: as the number of market participants declines—often due to mergers or market exits—concentration naturally intensifies. The number of insurance claims (NICl) shows a moderate but statistically insignificant negative relationship, suggesting that higher claim volumes may be associated with more competitive environments, though the evidence is inconclusive. Insurers' net profit or loss (ICoP) is positively and significantly correlated with HHI, implying that firms operating

in more concentrated markets may benefit from greater pricing power or economies of scale, leading to enhanced profitability. Technical reserves (TR) and insurance coverage (ICv) both display weak and non-significant negative correlations, indicating that these structural balance sheet variables are not closely tied to market concentration dynamics in the observed period. In contrast, financial placements (FP) show a very strong and statistically significant negative relationship with HHI, suggesting that insurers in less concentrated markets tend to manage larger investment portfolios—possibly due to greater competition driving diversification and capital allocation strategies. These findings collectively highlight the nuanced interplay between market structure and firm-level financial behavior in the Czech insurance sector.

2.3 Correlation analysis of HHI and Macroeconomic related indicators

The same methodological framework was applied to examine the relationship between the Herfindahl-Hirschman Index (HHI) and selected macroeconomic indicators. Table 4 summarizes the results of the Mardia tests for multivariate normality, while Table 5 presents the correlation coefficients and their statistical significance.

Indicator	Mardia tests	p-value	normality decision
GDP	Skewness	0,9014	cannot be rejected
	Sample corrected	0,8069	cannot be rejected
	Kurtosis	0,1619	cannot be rejected
	Doornik and Hansen omnibus	0,9305	cannot be rejected
Infl	Skewness	0,03505	**weakly rejected**
	Sample corrected	0,003308	**strongly rejected**
	Kurtosis	0,5441	cannot be rejected
	Doornik and Hansen omnibus	0,0001331	**strongly rejected**
GoD	Skewness	0,1264	cannot be rejected
	Sample corrected	0,02684	**weakly rejected**
	Kurtosis	0,9208	cannot be rejected
	Doornik and Hansen omnibus	0,004021	**strongly rejected**

Table 4 Bivariate normality test of HHI and macroeconomic indicators

Indicator	Correlation type	Correlation coefficient	p-value	dependence description
GDP	Pearson	0,65771	0,02784	positive, strong, significant
Infl	Spearman	0,65455	0,02887	positive, strong, significant
GoD	Spearman	0,77273	0,0052995	positive, strong, significant

Table 5 Correlation coefficients and power of dependence of HHI and macroeconomic indicators

All three macroeconomic variables — gross domestic product (GDP), inflation (Infl), and government debt (GoD) — exhibit strong and statistically significant positive correlations with HHI. The positive association between GDP and market concentration may reflect the tendency of larger, more developed economies to support the dominance of a few well-capitalized insurers capable of scaling operations efficiently. Similarly, the strong correlation between inflation and HHI could indicate that inflationary environments disproportionately benefit larger insurers, who are better equipped to manage cost pressures and adjust pricing strategies. Finally, the positive relationship between government debt and market concentration may suggest that in fiscally strained environments, regulatory or market conditions favour consolidation, possibly due to increased risk aversion or capital requirements that smaller insurers struggle to meet. These findings underscore the relevance of macroeconomic context in shaping the structural dynamics of the insurance sector.

2.4 Multiple linear regression model

To further investigate the joint influence of insurance sector and macroeconomic indicators on market concentration, a multiple linear regression (MLR) model was estimated with the Herfindahl-Hirschman Index (HHI) as the dependent variable. The model includes all nine explanatory variables previously analyzed in bivariate correlation tests. This approach allows for the simultaneous control of interdependencies among predictors and provides insight into their partial effects on HHI. The model yields a high coefficient of determination ($R^2 = 0,987$), indicating

that nearly 99% of the variation in HHI is explained by the included variables. However, the adjusted R^2 (0,867) and the relatively low F-statistic significance (p = 0,265) suggest potential overfitting or multicollinearity, especially given the small sample size (N = 11). These limitations imply that while the model fits the data well in-sample, its inferential power should be interpreted with caution.

MLR model statistics		Indicator	Coeff.	Std.err.	t	p	R^2
Dependent variable:	HHI	Constant	1867,5	17541	0,10647	0,93248	
N:	11	NICo	-11,517	52,629	-0,21883	0,86285	0,6139
Multiple R:	0,9933	NICl	-0,00039	0,0005	-0,77216	0,5814	0,21933
Multiple R2:	0,98665	ICoP	-0,00002217	0,000011574	-1,9152	0,30634	0,31818
Multiple R2 adj.:	0,86653	TR	0,00000154	0,000025745	0,059819	0,96196	0,042252
		ICv	388,08	726,15	0,53444	0,68754	0,072049
ANOVA		FP	-0,00548	0,006894	-0,7954	0,57224	0,72689
F:	8,2134	GDP	0,000287	0,000811	0,35351	0,78368	0,43258
df1, df2:	9, 1	Infl	-42,124	30,866	-1,3647	0,40258	0,34859
p:	0,26483	GoD	0,043588	0,2896	0,15051	0,9049	0,65765

Table 6 HHI Multiple linear model statistics, coefficients and significance

The estimated coefficients reveal that none of the predictors are statistically significant at conventional levels, as all *p*-values exceed 0.05. Nonetheless, the signs and magnitudes of the coefficients offer some interpretative value. The number of insurance companies (NICo) and financial placements (FP) both exhibit negative coefficients, consistent with earlier findings suggesting that market consolidation and reduced investment activity are associated with higher concentration. The coefficient for insurers' profit (ICoP) is also negative, which contrasts with the bivariate result and may reflect interaction effects or suppression by other variables. Technical reserves (TR) and insurance coverage (ICv) show weak positive effects, while GDP and government debt (GoD) also enter the model with positive signs, suggesting that larger and more indebted economies may support more concentrated insurance markets. Inflation (Infl) displays a negative coefficient, potentially indicating that price instability may erode market dominance. Overall, while the model confirms the relevance of several predictors, the lack of statistical significance highlights the need for cautious interpretation and possibly a larger dataset for robust inference.

3 Concluding recommendations

This article analyzed the determinants of the Herfindahl-Hirschman Index (HHI) in the Czech insurance sector, with an emphasis on market power and concentration. The results show that the Czech insurance market is moderately concentrated, with HHI values between 1,167 and 1,591 in the period 2012-2022. Key factors influencing market concentration include the number of insurance companies, financial positioning, and macroeconomic indicators such as GDP, inflation, and government debt.

A strong negative correlation between the number of insurance companies and market concentration suggests that market consolidation leads to higher concentration. The positive correlation between insurance companies' net profits and HHI shows that insurance companies in more concentrated markets can benefit from greater pricing power and economies of scale. Macroeconomic factors such as GDP, inflation, and government debt also significantly influence market concentration, underscoring the importance of the broader economic context.

Based on these findings, the article recommends regulatory and strategic initiatives aimed at increasing competition and market efficiency in the Czech insurance sector. It is important to monitor consolidation trends and their impact on market dynamics in order to ensure healthy economic development and consumer protection.

To further deepen the analysis, it would be possible to consider extending the time horizon of the data examined and including other macroeconomic indicators, such as the impact of interest rates or the impact of monetary and fiscal policy on market concentration in the insurance sector. According to Sharmin and Arman (2017), it would also be appropriate to use a combination of the HHI index and the concentration of the three or four largest companies in the sector under review, or entropy index. Alternatively, it would also be possible to examine consumer behavior in the insurance market and how the increased demand for digital services today affects market dynamics and concentration in the insurance industry. The findings of the analysis can thus be further elaborated and the analysis deepened in several directions, as the topic has great potential for future research.

Acknowledgements

This contribution has been supported by the institutional support of the University of Pardubice, Czech Republic.

References

Almajali M, Zainuddin M (2023) Market Concentration and Profitability: Empirical Evidence from the Jordanian and Saudi Arabia Insurance Industries. Review of Economics and Finance

Berry-Stölzle T R, Weiss M A, Wende S (2011) Market Structure, Efficiency, and Performance in the European Property-Liability Insurance Industry. In: ARIA 2011 Meeting. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b0df226d40af84ae06c ce42d38f6b1e1eb69d5b7

Böhm A, Mužáková K (2010) Pojišťovnictví a regulace finančních trhů. Praha: Professional Publishing (in Czech)

Brčák J, Sekerka B, Severová L, Svoboda R (2020) Mikroekonomie: teorie a aplikace. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk (in Czech)

Brezina I, Pekár J, Čičková Z, Reiff M (2014) Herfindahl–Hirschman index level of concentration values modification and analysis of their change. Central European Journal of Operations Research, 24, pp. 49-72

Ducháčková E (2009) Principy pojištění a pojišťovnictví. Praha: Ekopress (in Czech)

European Commission (2004) Guidelines on the assessment of horizontal mergers under the Council Regulation on the control of concentrations between undertakings. Official Journal of the European Union, C 31, 5.2.2004, pp. 5–18. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=oj:JOC_2004_031_R_0005_01

Evans J D (1996) Straightforward statistics for the behavioral sciences. Pacific Grove, CA: Brooks/Cole Publishing

Herfindahl O (1950) Concentration in the Steel Industry, Unpublished Ph.D. dissertation, Columbia University, U.S.A.

Hirschman, A (1945) National Power and the Structure of Foreign Trade, University of California Press, Berkeley CA

Janků J, Badura O (2021) Non-linear Effects of Market Concentration on the Underwriting Profitability of the Non-life Insurance Sector in Europe. Working Paper Series

Naldi M, Flamini M (2014) The CR4 Index and the Interval Estimation of the Herfindahl-Hirschman Index: An Empirical Comparison. SSRN. https://ssrn.com/abstract=2448656

Nejezchleb K, Hajná Z (2014) Stanovení tržní síly a její význam v soutěžním právu. Online. Časopis pro právní vědu a praxi. ISSN 1210-9126.7 (in Czech)

Sharma G, Rotthoff K W (2019) Impact of Market Concentration on Employment and Wages: a Look at the Insurance Industry. Atlantic Economic Journal, pp. 169–178

Sharmin S, Arman M (2017) Measuring Market Concentration using Indices: An Empirical Analysis on Listed Financial Companies of Bangladesh. AIMS International Journal of Management, 17(4), 1513-1519. https://www.aims-international.org/aims17/17ACD/PDF/A412-Final.pdf

Shim J (2017) An Investigation of Market Concentration and Financial Stability in Property–Liability Insurance Industry. The Journal of Risk and Insurance

Sukpaiboonwat S (2014) Measuring the Degree of Market Concentration in Thailand Insurance Industry. https://www.semanticscholar.org/paper/Measuring-the-Degree-of-MarketConcentration-in-SukpaiboonwatPiputsitee/36c4fb3f8bbfc2429ea979d6ab7036314874c87

U.S. Department of Justice and Federal Trade Commission (2010) Horizontal Merger Guidelines. https://www.ftc.gov/sites/default/files/attachments/merger-review/100819hmg.pdf

Vávrová E (2014) Finanční řízení komerčních pojišťoven. Expert. Praha: Grada (in Czech)

Vlková S (2024) Insurance market. Diploma thesis. Univerzita Pardubice (in Czech)

Walker I (2013) Výzkumné metody a statistika. Praha: Grada Publishing, a. s. (in Czech)

Zemplinerová A (1999) Tržní koncentrace ve zpracovatelském průmyslu České republiky a antimonopolní politika. Politická ekonomie. DOI: 10.18267/j.polek.21. ISSN 0032-3233 (in Czech)

Received: 2025-06-03 Last revision: 2025-07-15 Accepted: 2025-07-29

Modelling Capital–Labor Substitution in the Era of Industry 4.0: Evidence from the Czech Manufacturing Sector

Alena Pozdílková[1], Martin Pozdílek[2], Martina Hedvičáková[3]

Abstract. The advent of Industry 4.0 is reshaping the labor, capital, and productivity landscape, posing both challenges and opportunities for economic policy and corporate strategy. A critical component of this transformation is understanding the extent to which labor is being substituted by capital—an insight essential for informed managerial decisions and the creation of innovation-driven policies. This paper employs the Cobb-Douglas production function as a robust analytical framework to quantify this dynamic within the Czech manufacturing sector. Through correlation analysis and least squares estimation, we construct production functions for the manufacturing industry. The explanatory power of each model is evaluated through determination coefficients, providing a comparative perspective on capital-labor relationship. The findings offer data-driven insights into the evolving capital-labor dynamics in the context of Industry 4.0, highlighting key implications for automation, technological investment, and the strategic development of workforce skills.

Keywords: Industry 4.0, Cobb-Douglas production function, manufacturing industry, labor, capital

JEL Classification: D81, C69
AMS Classification: 90C15

1 Introduction

The Cobb-Douglas production function is one of the best-known and most widely used analytical tools in economics. Since its formulation by Charles Cobb and Paul Douglas in the 1920s, it has been successfully applied in countless empirical studies – across industries, levels of aggregation, and time periods. Its appeal lies in its simplicity, interpretability of parameters, and empirically observed agreement between the estimated coefficients and the shares of production factors in value added. It is this "agreement" that used to be associated with neoclassical distribution theory, which the Cobb-Douglas function supported in many ways (Balk, 2024).

However, the contemporary economic literature evaluates this function much more critically. As a number of studies have shown (Balk, 2024; Felipe & McCombie, 2013; Temple, 2021), the observed accuracy of estimates may simply be a consequence of accounting identity, and not a reflection of a real technological relationship. Although the form of the function is theoretically derived from assumptions about the behavior of firms and production technologies, in real applications it often relies on simplifying assumptions such as constant returns to scale, constant factor shares or invariance of parameters over time. Despite this, the Cobb-Douglas function retains a strong position in practical economic analysis, mainly due to its comprehensibility and the availability of the necessary data.

At the same time, however, fundamental changes are taking place in production processes under the influence of digitalization and automation, collectively referred to as the fourth industrial revolution (Industry 4.0). This technological shift places new demands on modeling the relationship between inputs and outputs, as traditional production factors are expanded by data infrastructure, software solutions, artificial intelligence, or cyber-physical systems. In this context, the question is whether and to what extent the classic Cobb-Douglas function can capture structural changes in production caused by the implementation of the principles of Industry 4.0, and whether its parameters remain economically interpretable.

This study, therefore, seeks to build on the original tradition of applying the production function in sectoral analyses while reflecting the criticisms that have been leveled against this approach in recent decades. We focus on the Czech manufacturing industry, where we test the suitability and limits of the Cobb-Douglas specification in estimating the relationships between capital, labor, and output in the context of the transformation of firms towards

[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.
[3] University of Hradec Králové, Department of Economics, Rokitanského 62, Pardubice, martina.hedvicakova@uhk.cz.

the Industry 4.0 model (Bettiol et al., 2024; Pozdílek et al., 2024). Our goal is not to confirm the superiority of this functional form, but to evaluate its informational contribution and discuss whether its parameters say more about the production technology or the method of economic accounting.

2 Methodology

2.1 Cobb-Douglas production model

The Cobb-Douglas production function is a widely used tool in economics for modelling the relationship between inputs and output. Originally proposed by Swedish economist Knut Wicksell (1851–1926), it was later empirically tested and popularized by Charles Cobb and Paul Douglas in 1928. Their study, which analysed the growth of the U.S. economy from 1899 to 1922, introduced a simplified economic model in which output is determined solely by the levels of labor and capital. Despite its simplicity and the omission of many other relevant economic factors, the Cobb-Douglas function has shown a remarkable ability to accurately reflect real-world economic dynamics in the long run.

The Cobb-Douglas production function is used in the following form (Cobb & Douglas, 1928):

$$Q(L,K) = AK^{\alpha}L^{\beta}, \tag{1}$$

where the parameters are:

Q … total output,

K … capital input,

L … labor input,

A … total factor productivity,

α … elasticity of production relative to capital input,

β … elasticity of production relative to labor input,

A, α, β … positive constants, where $\alpha + \beta \approx 1$.

Unless the goal is to examine technological progress or compare productivity across economies, $A = 1$ is set to simplify the model to the basic relationship between capital, labor, and output. In theoretical models (e.g., the Solow model without technological progress), it is customary to start with $A = 1$ unless technology is analyzed separately. In our study, we will also use a model with $A = 1$.

2.2 Ordinary Least Squares (OLS) Method

The Ordinary Least Squares (OLS) method is a fundamental statistical technique particularly well-suited for analyzing data obtained through measurement. It can also be effectively applied to estimate the parameters of the Cobb-Douglas production function using empirical input data. To facilitate the use of OLS, it is necessary to linearize the model by taking the natural logarithm of both sides of the original Cobb-Douglas function:

$$ln(Q) = ln(A) + \alpha\, ln(K) + \beta l\, n(L). \tag{2}$$

For this transformation to be valid, the variables K, L, and Q must all be positive, which is generally the case in economic applications. Introducing the substitutions

$$a = ln(A)\,, x = ln(K)\,, y = ln(L)\,, z = ln\,(Q), \tag{3}$$

the equation takes the form of a linear regression model

$$z = a + \alpha x + \beta y. \tag{4}$$

This linear formulation allows us to apply the OLS method to estimate the parameters a, α, β. Once these estimates are obtained, the original parameters of the Cobb-Douglas production function can be recovered by exponentiating the parameter a and substituting all values back into the non-linear form of the function.

2.3 Coefficient of Determination

To assess the quality of the regression model, we use the coefficient of determination, commonly denoted as R^2. This metric quantifies the proportion of the variance in the dependent variable that is explained by the model. In its basic form, R^2 is defined as one minus the ratio of the residual variance (i.e., the variance of the differences between the model's predicted values and the actual observed values) to the total variance of the dependent variable. A higher R^2 value indicates a better fit of the model to the data. (Salh, 2015)

3 Analysis and Results

3.1 Data Description

To estimate the parameters of the Cobb-Douglas production function, the log-linear form of the equation was chosen, which allows the use of the ordinary least squares (OLS) method for linear regression. In this approach, output (Y) represents value added at current prices, which is a commonly used indicator of production, as it reflects the net output created by enterprises without double-counting of inputs. Investment in fixed capital was used as a proxy for capital input (K), which is a common approach in cases where detailed data on the capital stock are not available. Labor input (L) was represented by labor costs, which include not only the number of workers, but also partly their qualifications and remuneration for the Czech manufacturing industry. This selection of variables corresponds to the methodology commonly used in the empirical literature, for example, in the study (Hájková & Hurník, 2007), and is conditioned by the availability of data in the Czech environment. The data were drawn from publicly available sources of the Czech Statistical Office (2025) and the Ministry of Industry and Trade of the Czech Republic (2025), covering annual data for selected sectors in the period 2008-2023. This methodology allows to quantify the elasticity of output to labor and capital and to estimate the rate of technological progress.

We verified the correlation on the available input data, which turned out to be quite high (0.79), which is caused by the fact that larger companies invest more and pay more, and higher investments lead to higher employment and labor costs. In the model, the values are logarithmized, which reduces the correlation.

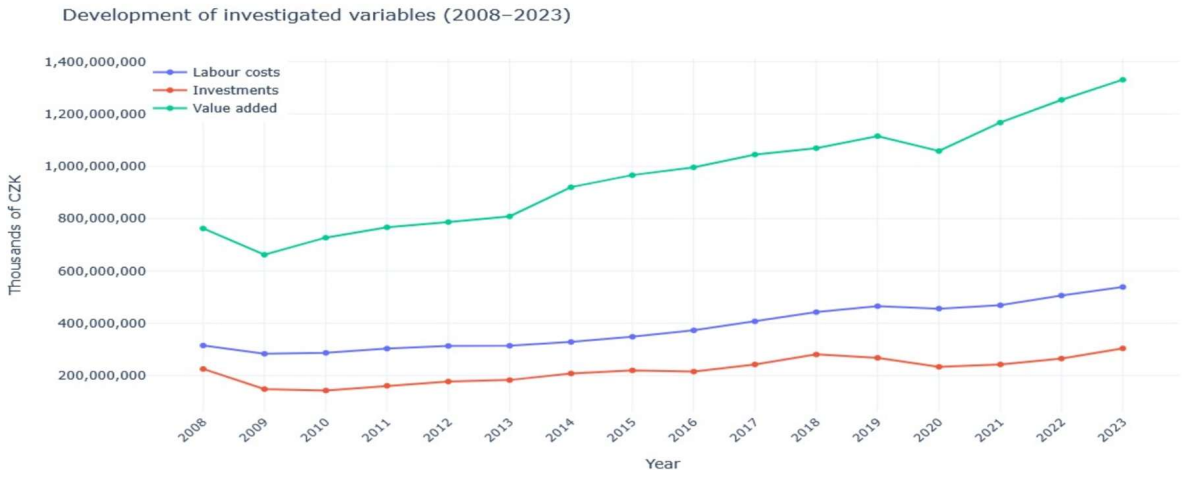


Figure 1 Development of the investigated variables in 2008 - 2023. Source: Own processing.

3.2 Results

Using the Ordinary Least Squares (OLS) method applied to the Cobb-Douglas production function, the following elasticities were obtained:

Parameter	Elasticity	Confidence interval
α = elasticity for investments	0.5096	<0.3078, 0.7114>
β = elasticity for labour costs	0.4383	<0.2518, 0.6248>

Table 1 Results of elasticity coefficients computed using OLS method. Source own processing.

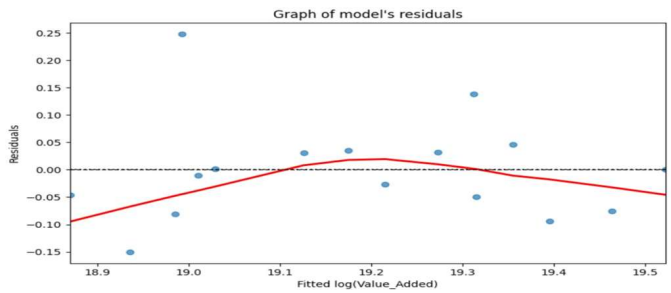


Figure 2 Graph of model residuals. Source: Own processing.

Computed results of α and β ($\alpha + \beta < 1$) shows, that the Cobb-Douglas production function exhibits decreasing returns to scale. Graph of model residuals is shown in the Figure 2.

For the developed Cobb-Douglas model, the coefficient of determination R^2 for the manufacturing industry was calculated as $R^2 = 0.8319$.

This value, which is close to 1, indicates a high degree of fit, suggesting that the model accurately represents the relationship between the inputs and output in the manufacturing sector.

The following Figure 2 illustrates the results of the Cobb-Douglas function. After applying the inverse logarithmic transformation and substituting the values into the Cobb-Douglas production function, Figure 3 displays the surface in its final form for the specific subject under study.

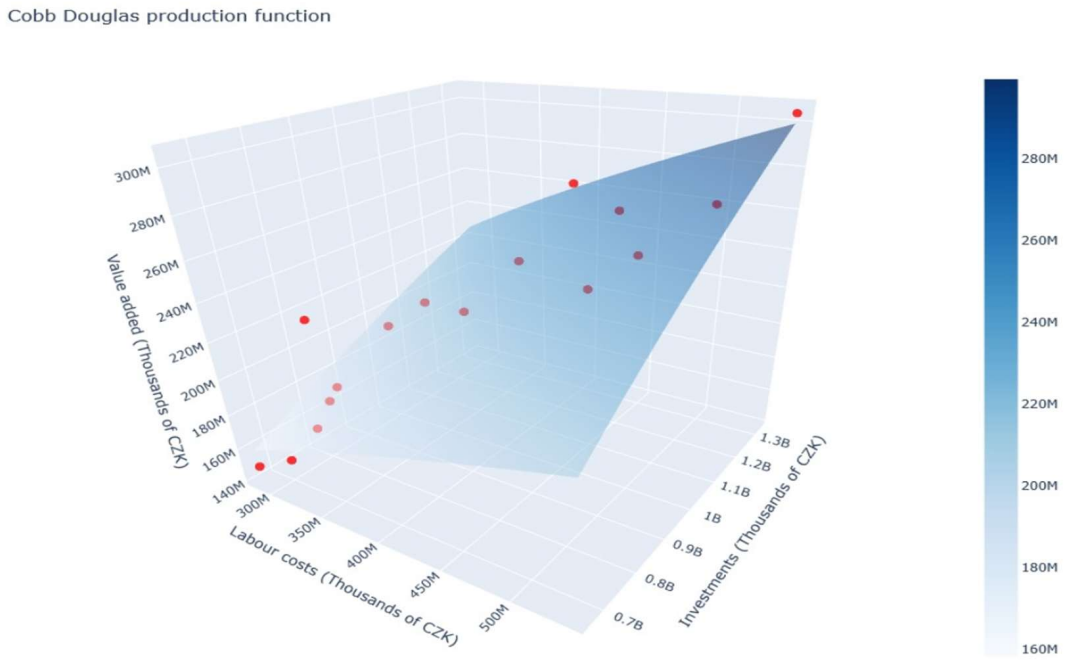


Figure 3 Results of the Cobb-Douglas function. Source: Own processing.

4 Discussions

The results of our study, based on the application of the Cobb-Douglas production function to data from the Czech manufacturing industry, show consistent relationships between capital, labor, and value added. The model parameters show the expected sign and magnitude – the share of labor and capital is approximately equal to their shares in output, which supports the basic assumptions of the marginalist theory of distribution. Nevertheless, it is important to interpret the results cautiously and in the broader context of the current scholarly debate on the suitability of this production function.

In particular, the conclusions of Gechert et al. (2019) indicate that the Cobb-Douglas function systematically deviates from empirically estimated values of the elasticity of substitution between labor and capital. Their meta-analysis of more than 3,000 estimates shows that the average elasticity of substitution, after correction for publication bias, is more likely to be around 0.3, rather than 1, implying that CES (Constant Elasticity of Substitution) functions could describe reality better.

Similarly, Sheng (Sheng, 2020) argues that the Cobb-Douglas function is often just a formal transcription of the national economic identity without any real informative value about the technological process. According to the author, the significance of marginal product cannot be attributed to the partial derivatives of this function without further ado, and moreover, the validity of constant returns to scale is not ensured by construction, but ex post due to the calculation method.

The discussion about the validity of the Cobb-Douglas approach is also relevant in the case of converging economies. The study (Hájková & Hurník, 2007, 2007b) examines the Czech economy and shows that although

deviations from the constant share of labor in the product (key assumption CD function) exist, the results are relatively robust to these changes. However, the authors themselves recommend using more general production functions as a tool for sensitivity analysis.

Foreign empirical applications, for example to Bangladesh (Husain & Islam, 2016), Oman (Zakir Hossain & Said Al-Amri, 2010) or the USA (Elsby et al., 2013) further show the diversity of parametric results, often detecting non-constant returns to scale and a decline in the labor share of output, a phenomenon that the Cobb–Douglas form is unable to adequately capture without extension.

Despite these theoretical and empirical reservations, the Cobb–Douglas production function remains a suitable tool for initial analysis in our study due to its simplicity and transparency. Moreover, the estimates were subjected to diagnostics – the Durbin–Watson and Breusch–Pagan tests do not indicate a violation of the basic assumptions of OLS. Nevertheless, we are aware that the model implicitly assumes constant elasticity of substitution, constant returns to scale, invariance of parameters over time, exogenously given technology.

These assumptions are not empirically tested in this study, which is an important direction for future research – for example, the application of translog or CES functions could reveal nonlinearities or time-varying relationships.

In the context of Industry 4.0, these limitations gain additional relevance. As manufacturing processes become increasingly automated, digitized, and integrated with advanced technologies (e.g., IoT, AI, robotics), the nature of capital and labor inputs changes significantly—not only in quantity but also in quality and substitutability. The standard Cobb-Douglas framework, assuming constant substitution elasticity, may thus fail to reflect the true dynamics of modern production systems. Future modeling approaches for Industry 4.0 environments should consider incorporating variable elasticities, endogenous technological change, and digital capital to better capture these structural shifts.

From a theoretical perspective, alternative approaches can also be considered – for example, models based on Lie groups (Lange & Papageorgiou, 2020; Smirnov & Wang, 2020) or Hamiltonian structures (Smirnov & Wang, 2021) represent advanced ways to describe growth mechanisms without restrictive assumptions.

5 Conclusion

The aim of this study was to empirically verify the validity of the Cobb-Douglas production function in relation to the Czech manufacturing industry in the period 2008–2023. Based on publicly available data from the Czech Statistical Office and the Ministry of Industry and Trade, a log-linear form of the production function was constructed and estimated using the ordinary least squares (OLS) method. Output was represented by value added, capital input by fixed capital investment, and labor input by labor costs.

The estimation results show that the elasticity of output to capital is 0.5096 and the elasticity to labor is 0.4383. The sum of these elasticities is less than one ($\alpha + \beta < 1$), which indicates the occurrence of increasing returns to scale in the given sector. The coefficient of determination R^2 =0.8319 indicates a very good agreement of the model with the empirical data, which confirms the suitability of using the Cobb-Douglas production function to describe the relationship between inputs and output in the Czech manufacturing industry.

The knowledge gained goes beyond theoretical discussion and can serve as a valuable basis for economic policy-making and strategic planning in industrial sectors. In particular, it can be useful in considering the efficient allocation of production factors and in assessing the impacts of technological progress on productivity. These insights are especially relevant in the context of Industry 4.0, which is transforming traditional manufacturing through automation, digitalization, and the integration of smart technologies. The observed capital elasticity suggests that investment in advanced technologies—such as robotics, AI-driven systems, and interconnected machinery—can play a critical role in enhancing productivity. However, the significant contribution of labor also highlights the ongoing importance of human capital, particularly in roles requiring adaptability, digital skills, and oversight of automated systems. Therefore, aligning industrial policy and workforce development with the principles of Industry 4.0 is essential for maintaining competitiveness and sustainable growth in the Czech manufacturing sector.

Acknowledgements

The paper was supported by institutional support of the University of Pardubice and by the internal project “SPEV – Economic Challenges and Opportunities within Industry 4.0 and the Societal Concepts of 5.0 and 6.0”, 2025, University of Hradec Králové, Faculty of Informatics and Management, Czech Republic. “

References

Balk BM. (2024) Why is the Cobb-Douglas production function so popular? Evolutionary and Institutional Economics Review, 21(1), 1–20. https://doi.org/10.1007/s40844-024-00279-x

Bettiol M, Capestro M, Di Maria E, Ganau R. (2024) Is this time different? How Industry 4.0 affects firms' labor productivity. Small Business Economics, 62(4), 1449–1467. https://doi.org/10.1007/s11187-023-00825-8

Cobb CW, Douglas PH. (1928) A Theory of Production. The American Economic Review, 18(1):139-65

ČSÚ. (2025) Český statistický úřad, hlavní makroekonomické ukazatele. https://www.czso.cz/csu/czso/hmu_cr

Elsby MW, Hobijn B, Şahin A. (2013) The Decline of the U.S. Labor Share. Brookings Papers on Economic Activity, 2013(2), 1–63. https://doi.org/10.1353/eca.2013.0016

Felipe J, McCombie JS. (2013) The Aggregate Production Function and the Measurement of Technical Change: ŠNot Even Wrong . Edward Elgar Publishing.

Gechert S, Havranek T, Irsova Z, Kolcunova D (2019) Death to the Cobb-Douglas Production Function? A Quantitative Survey of the Capital-Labor Substitution Elasticity, Working Papers 2019/8, Czech National Bank.

Hajkova D, Hurník J. (2007)Cobb-Douglas production function: The case of a converging economy. Czech Journal of Economics and Finance (Finance a uver). 57(9-10):465-76.

Hájková D, Hurník J (2007b) Cobb-Douglas Production Function: The Case of a Converging Economy," Czech Journal of Economics and Finance (Finance a uver), Charles University Prague, Faculty of Social Sciences, 57(9-10), 465-476,.

Husain S, Islam MS. (2016) A Test for the Cobb Douglas Production Function in Manufacturing Sector: The Case of Bangladesh. International Journal of Business and Economics Research. 5(5):149-54. https://doi.org/10.11648/j.ijber.20160505.13

MPO (2025). Ministerstvo průmyslu a obchodu České republiky. Panorama zpracovatelského průmyslu. https://www.mpo.cz/cz/panorama-interaktivni-tabulka.html

Lange F, Papageorgiou T (2020) Beyond Cobb-Douglas: Flexibly Estimating Matching Functions with Unobserved Matching Efficiency (No. w26972; s. w26972). National Bureau of Economic Research. https://doi.org/10.3386/w26972

Pozdílek M, Pozdílková A, Hedvičáková M (2024) Analysis of the Manufacturing Industry Using the Cobb-Douglas Production Function (J. Maci, P. Maresova, K. Firlej, & I. Soukal, Ed.; s. 366–375). https://doi.org/10.36689/uhk/hed/2024-01-032

Sheng Y (2020) The Success of the Cobb-Douglas Production Function Is An Illusion. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3666648

Smirnov RG, Wang K (2020) In search of a new economic model determined by logistic growth. European Journal of Applied Mathematics, 31(2), 339–368. https://doi.org/10.1017/S0956792519000081

Smirnov RG, Wang K (2021) The Cobb-Douglas Production Function Revisited. In D. M. Kilgour, H. Kunze, R. Makarov, R. Melnik, & X. Wang (Ed.), Recent Developments in Mathematical, Statistical and Computational Sciences (Roč. 343, s. 725–734). Springer International Publishing. https://doi.org/10.1007/978-3-030-63591-6_66

Salh SM (2015) Estimating R2 Shrinkage in Regression. International. Journal of Technical Research and Applications, 3, 1–6.

Temple J (2021) Jeff E. Biddle, Progress Through Regression: The Life Story of the Empirical Cobb-Douglas Production Function. OEconomia, 11–3, 571–576. https://doi.org/10.4000/oeconomia.11428

Zakir Hossain M, Said Al-Amri K (2010) Use of Cobb-Douglas production model on some selected manufacturing industries in Oman. Education, Business and Society: Contemporary Middle Eastern Issues, 3(2), 78–85. https://doi.org/10.1108/17537981011047925

Received: 2025-06-08 Last revision: 2025-08-10 Accepted: 2025-08-12

Detecting Regime Shifts in the Taylor Curve: A Comparison of Time-Varying Parameter and Markov Switching Models

Dominik Kavřík[1], Jan Rejthar[2]

Abstract. The Taylor curve characterizes the trade-off between the variances of inflation and the output gap, offering a theory-based framework for evaluating monetary policy. This paper compares two empirical approaches to modeling the time-varying nature of this relationship: a state-space model with time-varying parameters estimated via the Kalman filter, and a Markov switching regression that allows for discrete regime changes. Using quarterly macroeconomic data from the Czech Republic, both models are applied to detect periods of policy inefficiency, focusing on key episodes such as the 2008 global financial crisis, the Czech National Bank's currency interventions, and the high-inflation period following 2021. The time-varying parameter model captures gradual shifts in the policy trade-off, whereas the Markov switching model detects abrupt structural breaks—particularly during periods of heightened macroeconomic stress. The results suggest that the choice of the model affects conclusions about the effectiveness of monetary policy in different macroeconomic periods.

Keywords: monetary policy, regime switching, Taylor curve

JEL Classification: E31, E51, C13
AMS Classification: 91B62

1 Introduction

The Taylor curve (Taylor, 1979) represents the inverse relationship between the variances of inflation and the output gap, arising from a central bank's objective to minimize welfare loss. This theoretical concept, which stems from the second-order approximation of the welfare-maximizing objective function (Galí, 2015), suggests that under optimal monetary policy, a trade-off between these volatilities should be observed. The stability or evolution of this relationship can therefore serve as an indicator of policy efficiency, as explored in studies such as Olson et al (2012).

This paper estimates the dynamic Taylor curve relationship using two complementary approaches. The first is a time-varying parameter (TVP) model estimated via the Kalman filter, building on the methodology of Olson et al (2012), who applied this framework to analyze shifts in U.S. monetary policy. The TVP model captures gradual changes in the trade-off over time. The second approach is a Markov switching regression model (Goldfeld and Quandt, 1973), which allows for discrete regime changes in model parameters, capturing abrupt structural breaks.

The empirical analysis is based on quarterly macroeconomic data for the Czech Republic, specifically real gross domestic product (GDP) to estimate the output gap, and consumer price index (CPI) inflation. Conditional volatilities of both series are extracted using a Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model (Bollerslev, 1986), applied to filtered time series data after appropriate diagnostic testing.

The primary research question is whether these two modeling strategies can detect changes in policy efficiency during major economic events, such as the 2008 global financial crisis, or during periods of unconventional monetary policy—most notably the Czech National Bank's exchange rate commitment between 2013 and 2017 (Cecrdlova, 2021).

2 Methodology

This study evaluates the time-varying relationship between the volatilities of inflation and the output gap, as defined by the Taylor Curve. The analysis proceeds in three steps: Construction of conditional volatilities, specification of the time-varying parameter (TVP) model, and estimation of a Markov switching regression.

[1] Prague University of Economics and Business, Department of Econometrics, Winston Churchill Square 4, 13067 Prague, Czech Republic, dominik.kavrik@vse.cz
[2] Prague University of Economics and Business, Department of Econometrics, Winston Churchill Square 4, 13067 Prague, Czech Republic, jan.rejthar@vse.cz

Conditional Volatility

Conditional variances of inflation and the output gap, denoted by $h_{\pi,t}$ and $h_{\tilde{y},t}$ respectively, are estimated using the GARCH(1,1) model. In this two-step approach, a univariate ARIMA model is first fitted to each series, and the residuals are passed into a GARCH model to extract time-varying volatility. The order of the ARIMA model is determined by the Bayesian Information Criterion (BIC).

Time-Varying Parameter Model

Following the methodology of Olson et al (2012), the dynamic Taylor curve relationship is modeled using a state-space specification with time-varying coefficients:

$$h_{\pi,t} = \beta_{0,t} + \beta_{1,t} \cdot h_{\tilde{y},t} + u_t, \quad u_t \sim \mathcal{N}(0, \sigma_u^2), \tag{1}$$

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}), \tag{2}$$

where $\boldsymbol{\beta}_t = [\beta_{0,t}, \beta_{1,t}]'$ is a vector of time-varying coefficients and $\boldsymbol{Q}$ is a variance-covariance matrix of innovations $\boldsymbol{\eta}_t$ of the random walk process of parameters. Equation (1) is the measurement equation, linking the conditional variance of inflation to that of the output gap. Equation (2) is the state transition equation, assuming a random walk evolution for the parameters. Estimation is carried out using the Kalman (1960) filter .

Markov Switching Regression

To capture potential trade off violations, the analysis also includes a Markov switching regression, originally introduced by Goldfeld and Quandt (1973). This model allows the relationship between inflation and output gap volatilities to switch between discrete regimes governed by an unobserved Markov process. The specification used in the following analysis is given by:

$$h_{\pi,t} = \beta_0 + \beta_1^{(s_t)} \cdot h_{\tilde{y},t} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_{s_t}^2), \tag{3}$$

where $h_{\pi,t}$ and $h_{\tilde{y},t}$ denote the conditional variances of inflation and the output gap, respectively. The variable $s_t \in \{1, 2\}$ represents the latent regime at time t, evolving according to a first-order Markov process with transition probabilities $P(s_t = j \mid s_{t-1} = i)$ for $i, j \in \{1, 2\}$.

Each regime is characterized by its own slope coefficient $\beta_1^{(s_t)}$ and error variance $\sigma_{s_t}^2$. In this specification, the intercept β_0 is held fixed across regimes to isolate variation in the slope and volatility. The regimes can be interpreted as representing efficient and inefficient policy trade-offs, depending on the sign of $\beta_1^{(s_t)}$, where $\beta_1^{(s_t)} > 1$ indicates inefficient monetary policy and vice versa.

3 Data

The empirical analysis is based on quarterly macroeconomic data from the Czech Republic, spanning from 2000 Q1 to 2025 Q1. The dataset includes real gross domestic product (GDP), consumer price index (CPI). All data series are obtained from the Federal Reserve Economic Data (FRED) database provided by the St. Louis Federal Reserve.

Output Gap Estimation

The output gap is defined as the cyclical component of real GDP, extracted using the Christiano-Fitzgerald (CF) band-pass filter 2003, which is suited for quarterly economic data as it incorporates business cycle frequency and offers a more theoretically grounded approach compared to other commonly used methods such as the Hodrick-Prescott (HP) filter (Hodrick and Prescott, 1997), as examined in the comparative study by Kavřík (2023).

Let y_t denote the logarithm of real GDP. The CF filter decomposes the series into a trend component y_t^{trend} and a cyclical component y_t^{cycle} such that:

$$y_t = y_t^{\text{trend}} + y_t^{\text{cycle}}. \tag{4}$$

The output gap is then defined as the percentage deviation from the trend, $\tilde{y}_t = 100 \cdot y_t^{\text{cycle}}$ Christiano and Fitzgerald (2003).

Inflation Rate

Inflation is calculated as the year-on-year log difference of the Consumer Price Index (CPI), given by:

$$\pi_t = 100 \cdot [\log(\text{CPI}_t) - \log(\text{CPI}_{t-4})] \,. \tag{5}$$

Both Inflation rate and the resulting output gap are shown in Figure 1.

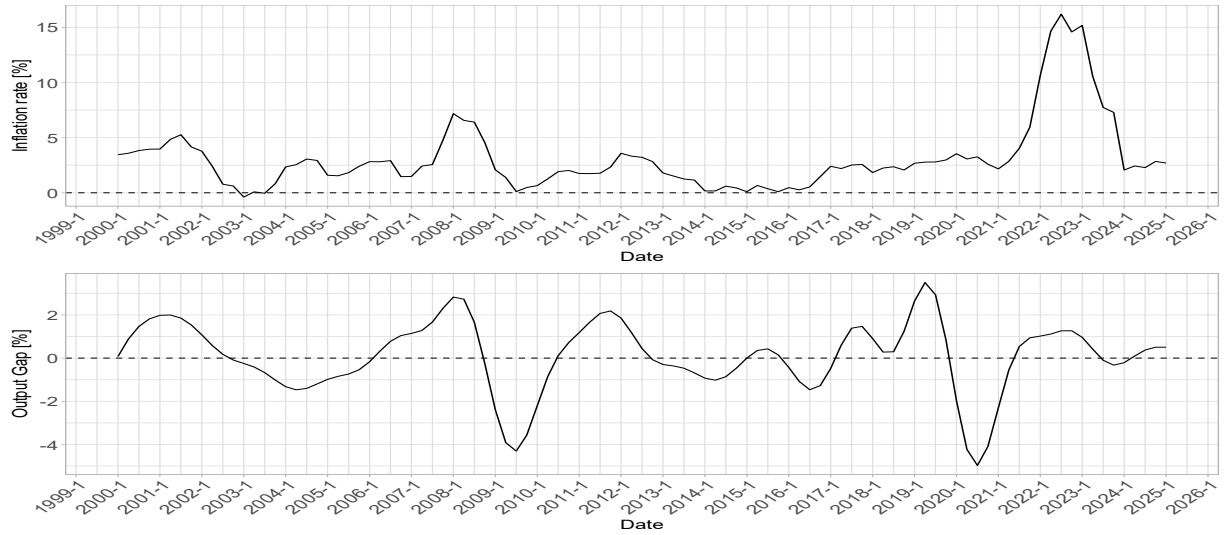


Figure 1 Inflation rate (top) and filtered output gap using CF filter (bottom).

4 Results and Discussion

Prior to modeling, both the inflation and output gap series were tested for the presence of unit roots. The Zivot-Andrews test (2002) was applied to the inflation series to account for a potential structural break in either the intercept or trend. Accounting for a breakpoint in Q4 2021, the test allowed rejection of the null hypothesis of a unit root. However, due to the well-known limitations of unit root tests (Perron, 1989; Zhou, 2013), ARIMA models of order one and zero were both considered in the conditional volatility estimation as a robustness check. Model selection was guided by the Bayesian Information Criterion (BIC).

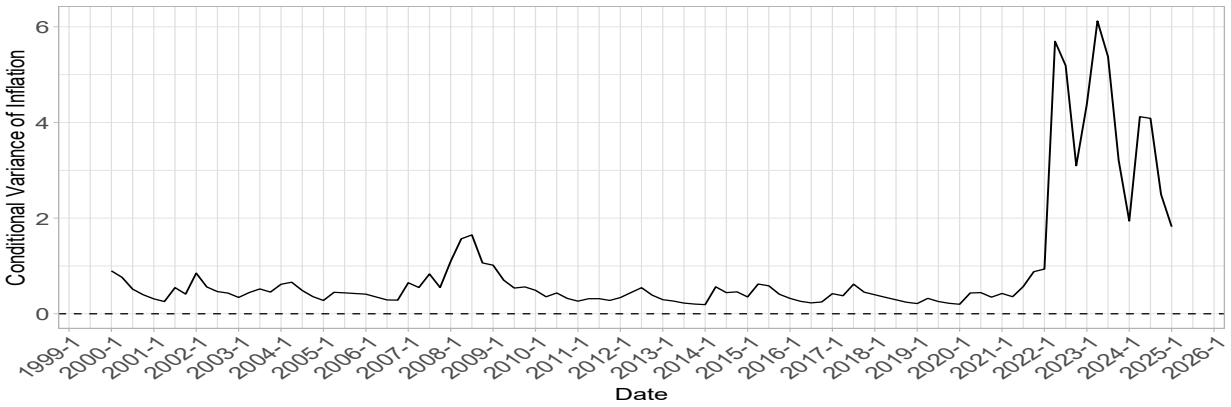


Figure 2 Conditional volatility of Inflation rate from GARCH(1,1) model in the Czech Republic.

Figure 2 shows two prominent spikes in inflation volatility: one in 2008, coinciding with the global financial crisis, and another in early 2022 during the period of high domestic inflation. These high-volatility episodes serve as key periods for evaluating whether the Taylor curve trade-off—between inflation and output-gap variability—held.

For the output gap series, the null hypothesis of a unit root was rejected at the 5% significance level. The corresponding conditional volatility is plotted in Figure 3.

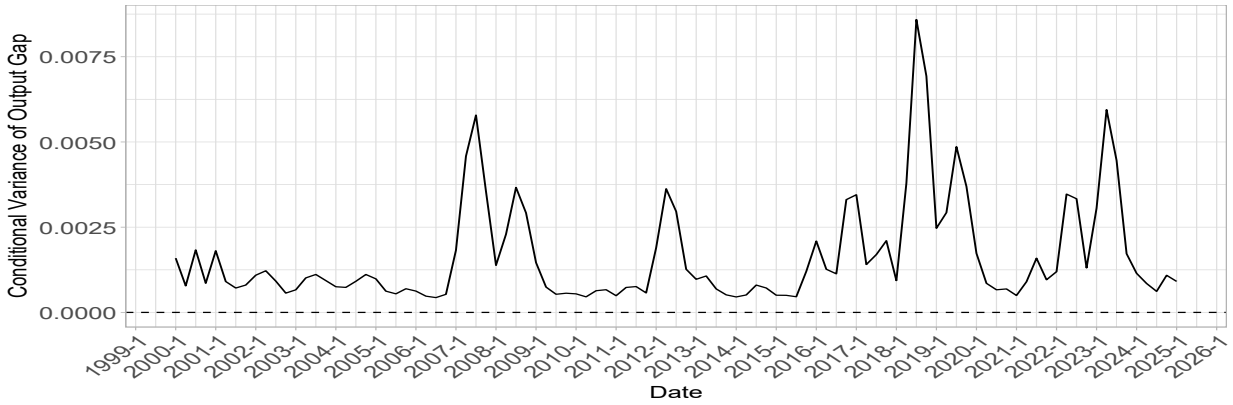


Figure 3 Conditional volatility of CF-filtered output gap in the Czech Republic.

The evolution of the time-varying parameter β_1 from Equation 2 is shown in Figure 4. This parameter reflects the strength and direction of the trade-off. A negative value of $\hat{\beta}_1$ at time t is interpreted as evidence of an efficient monetary policy consistent with the theoretical Taylor curve.

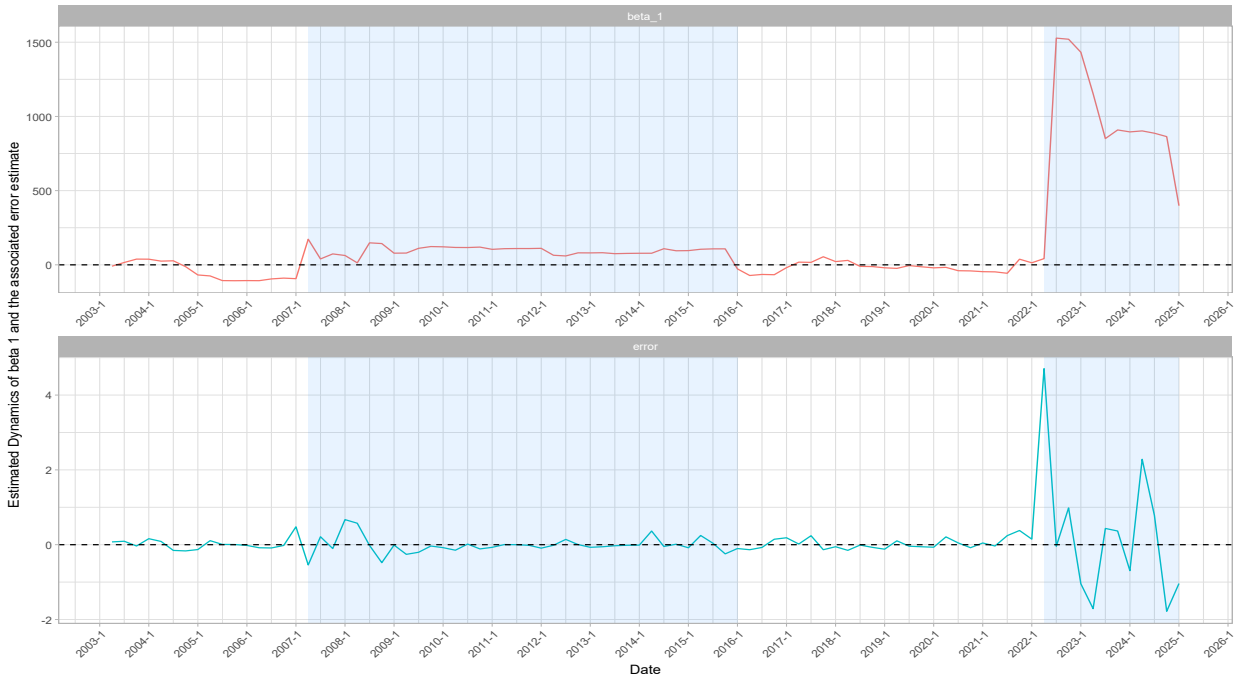


Figure 4 Estimated β_1 from the time-varying parameter model (upper) with highlighted periods of positive values, and estimated error (bottom).

The estimates in Figure 4 identify two distinct episodes during which the Taylor curve trade-off was violated. The first occurs in late 2007, preceding the 2008 financial crisis, and aligns with similar findings in Olson et al (2012) using data from the U.S., although in their case the crisis coincided with the end of the sample. The second violation emerges around early 2022, coinciding with a period of persistently high inflation and policy uncertainty in the Czech Republic. This period is also extensively analyzed in Šestořád and Dvořáková (2025). Notably, $\hat{\beta}_1$ remains positive throughout most of the Czech National Bank's exchange rate commitment regime (2013–2017), discussed in Cecrdlova (2021), which used currency interventions as a form of unconventional monetary policy.

To further assess structural changes in the policy trade-off, a Markov switching regression was estimated, allowing β_1 to alternate between two discrete regimes: one reflecting policy efficiency ($\beta_1 < 0$) and the other inefficiency ($\beta_1 > 0$). Figure 5 displays the smoothed probabilities of being in the inefficient regime. The model detects regime shifts that correspond closely to the same two critical periods identified by the TVP model: the lead-up to the global financial crisis and the recent inflationary surge. Compared to the continuous nature of the TVP model, the Markov approach highlights sharper transitions and reinforces the interpretation of 2007–2008 and post-2021 as periods of

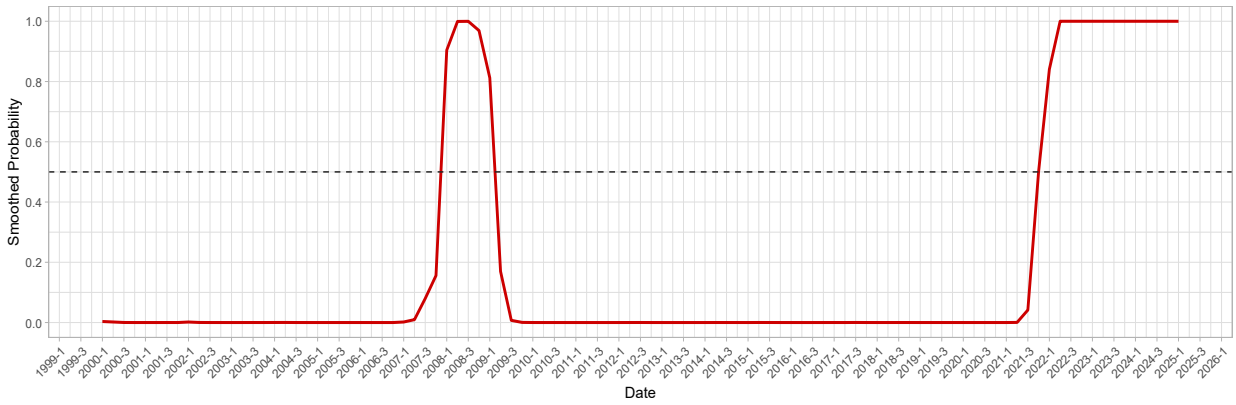


Figure 5 Smoothed probability of $\beta_1 > 0$ (regime 1) from the Markov regime switching model.

heightened policy stress. Interestingly, the regime classification is more persistent in the recent period, suggesting a longer-lasting departure from the optimal policy trade-off, which warrants further investigation. However, the period of currency intervention by the Czech National Bank does not appear to trigger a regime shift in the Markov switching model, suggesting that its impact on the inflation-output gap trade-off may have been more gradual or less pronounced in terms of structural breaks.

5 Conclusion

This paper evaluates the time-varying nature of the Taylor curve relationship between inflation and output gap volatility in the Czech Republic, using two empirical models: a time-varying parameter (TVP) and a Markov switching regression. Conditional volatilities were derived from GARCH models, using Christiano-Fitzgerald filter based output gap and consumer price inflation.

Both models capture key episodes of potential monetary policy inefficiency. The TVP model identifies a smooth departure from the Taylor curve trade-off during the periods surrounding the 2008 financial crisis and the post-2021 inflation surge. In contrast, the Markov switching regression highlights discrete shifts between regimes, emphasizing the abrupt nature of structural breaks during these episodes.

Importantly, while both approaches point to heightened macroeconomic stress in these two periods, the currency intervention period (2013–2017) does not register as a distinct regime shift, suggesting that its influence may have been absorbed more gradually. These findings demonstrate that the interpretation of policy performance can depend significantly on the chosen modeling framework. Future research may explore combining both modeling strategies or applying them to a broader set of economies to further assess the robustness of these conclusions.

Acknowledgements

This research was supported by the Internal Grant Agency of Prague University of Economics and Business under Project F4/39/2025.

References

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. Journal of econometrics 31(3):307–327

Cecrdlova A (2021) Comparison of the approach of the czech national bank and the european central bank to the effects of the global financial crisis. International Journal of Economic Sciences 10(2):18–46

Christiano LJ, Fitzgerald TJ (2003) The band pass filter. international economic review 44(2):435–465

Galí J (2015) Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework and Its Applications - Second Edition. Princeton University Press, google-Books-ID: 5GuYDwAAQBAJ

Goldfeld SM, Quandt RE (1973) A markov model for switching regressions. Journal of econometrics 1(1):3–15

Hodrick RJ, Prescott EC (1997) Postwar US business cycles: an empirical investigation. Journal of Money, credit, and Banking pp 1–16

Kalman RE (1960) A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering 82(1):35–45

Kavřík D (2023) Filtering methods for output gap estimation and the empirical taylor curve: A comparative study. In: Proceedings of the 41st International Conference on Mathematical Methods in Economics. Czech Society for Operations Research, Prague, pp 204–208, https://mme2023.vse.cz/mme_2023_proceedings.pdf, online

Olson E, Enders W, Wohar ME (2012) An Empirical Investigation of the Taylor Curve. Journal of Macroeconomics 34(2):380–390. https://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1297506&lang=cs&site=ehost-live&scope=site

Perron P (1989) The great crash, the oil price shock, and the unit root hypothesis. Econometrica: journal of the Econometric Society pp 1361–1401

Šestořád T, Dvořáková N (2025) Origins of post-covid-19 inflation in central european countries. Economic Systems p 101335

Taylor JB (1979) Estimation and control of a macroeconomic model with rational expectations. Econometrica: Journal of the econometric society pp 1267–1286

Zhou S (2013) Nonlinearity and stationarity of inflation rates: evidence from the euro-zone countries. Applied Economics 45(7):849–856

Zivot E, Andrews DWK (2002) Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. Journal of business & economic statistics 20(1):25–44

Received: 2025-06-05 Last revision: 2025-07-11 Accepted: 2025-07-29

Extension of the WSJF method with the concept of Efficiency for Scrum team

Jan Bartoška[1], Josef Kunhart[2], Jiří Pilný[3], Simona Huguet[4]

Abstract. The paper proposes a modification of the WSJF method using the efficiency principle of the DEA method. WSJF method is used for task prioritization in the Scrum method in agile project teams. Agile project teams use prioritization when selecting tasks to tackle during sprints. Task prioritization (also known as User Stories) is a key attribute for the success of the Scrum method, i.e., for meeting customer needs and expectations in the project. The authors propose an extension of the WSJF calculation using a modified Fibonacci sequence that is used in the Scrum method for estimating user stories. However, the contribution of the paper is a new perspective and a new way of calculating WSJF, where the extension of the method with the efficiency principle provides new information about the nature of tasks for the team and project management. The paper responds to the growing need for quantification of practices in agile project management, especially in teamwork, according to the Scrum method. The paper includes a case study that illustrates the use of the proposed modification for common practice in project management and the Scrum method. In this paper, the authors propose the use of weights for calculating the cost of delay in the WSJF method and a correction for calculating the WSJF score.

Keywords: Project management, Agile approach, Scrum, Project Team, Efficiency, DEA method.

JEL Classification: C44
AMS Classification: 90C15

1 Introduction

Nowadays, agile methods represent the optimal choice for developing software projects in rapidly changing environments. The success of small-scale agile projects and methods such as Scrum led the companies to apply them in larger projects (Uludağ et al., 2021). In the era of agile methodologies, organizations are exploring strategies to scale development across teams, leading to various scaling strategies such as LeSS or SAFe (Verwijs and Russo, 2023). Enterprises move from agile towards scaled agile as they have moved from traditional approaches towards agile in the past. Scaling enables collaborative teams from various functions to keep their agility and be more productive and efficient. As the product's demand rises, so do the features, requirements, and the size of the product backlog (Rashid and Gunjan, 2025). Scaled agile framework (SAFe) is a knowledge base of proven, integrated principles, practices, and competencies for achieving business agility that is built around the seven core competencies of business agility (Scaled Agile, 2024). SAFe is a comprehensive framework that incorporates ideas from agile development and lean production and aims at large-scale organizations (Dingsøyr, Falessi and Power, 2019).

In SAFe, software development effort estimation represents a critical activity. Proper estimation allows the team to efficiently plan development and determine team capacity (Tran, 2020). Cost of delay is a concept that combines understanding of value with how that value dissolves over time, and a tactic that helps prioritize development decisions by calculating the impact of time on value creation (Reinertsen, 2009). Webber (2022) views cost of delay from the financial perspective as the loss of income due to project delays. Many companies face a major challenge with the correct prioritization of multiple critical projects. If one project is prioritized, it usually means that others will be delayed (Riesener, 2023). The same principle applies to user stories in the product backlog. Prioritization of selected user stories leads to delays in non-prioritized user stories. Weighted shortest job first (WSJF) is a general prioritization algorithm. The algorithm is used to prioritize work and maximize output value in agile projects (Scaled Agile, 2024) and to schedule packets in computer networking (Chiang et al., 2022).

[1] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, bartoska@pef.czu.cz.
[2] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, kunhart@pef.czu.cz.
[3] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, pilnyj@pef.czu.cz.
[4] Department of Systems Engineering FEM CULS, Kamýcká 129, Prague, huguet@tf.czu.cz.

In SAFe, WSJF is used to continuously prioritize the backlog. The goal of WSJF is to minify the cost of delay of the evaluated user stories.

The backlog determines the difficulty of the team's work and the future scope of the project solution, where it is desirable to first create tasks with the highest earned value for the lowest effort. In SAFe, the backlog is compiled by the Agile team, which is expected to be autonomous and proactive in scoring of the tasks – the task rating is based on the subjective experiences of team members and determined a point scale in the form of the *Fibonacci sequence* (non-linear). Assigning points in the score of the task is therefore subjective and incomparable between different backlogs or teams. If the WSFJ method is appropriately extended with the DEA method principle, i.e., with the principle of a efficiency (as the ratio of outputs to inputs) and the WSJF score will be in the normalized range (*0-1*), it will be possible to assess the scored tasks in the backlog objectively - with the ability to compare between different backlogs and teams. The paper proposes a modification of the WSJF method using the efficiency principle of the DEA method.

2 Materials and Methods

2.1 Weighted shortest job first

Weighted shortest job first (WSJF) is a mathematical prioritization model used to sequence work for maximum economic benefit and reduced cost of delay. The model's goal is to help decide which user stories to choose next to deliver the best business value and balance between job size and cost of delay in the dynamic environment. In SAFe, WSJF is estimated as the relative cost of delay divided by the relative job duration (Scaled Agile, 2024), as defined in formula 1.

$$s = \frac{c}{j} \tag{1}$$

In formula 1, s represents the calculated WSJF score (index) for a given user story, while the parameters c and j represent relative cost of delay and job size in story points. As stated by Reinertsen (2009), cost of delay can be viewed as a partial derivative of the total expected value with respect to time and combines an understanding of business value with how that value leaks away over time. In other words, what costs incurs, if a product feature (user story) gets completed late or never. Job size indicates how much work effort the team needs to implement a given user story. In Scrum, job size (work effort) of user stories is measured in points. Story points indicate work effort, relative to other items in the product backlog (Schwaber and Sutherland, 2020). The calculated WSJF score indicates the importance of a user story and is subsequently used to prioritize the product backlog. Cost of delay is calculated as a sum of three attributes: business value, time criticality, and risk reduction (or opportunity enablement value), as displayed in formula 2.

$$c = v + t + r \tag{2}$$

In formula 2, c represents the cost of delay, parameter v represents business value, parameter t stands for time criticality, and parameter r reflects risk reduction. Business value includes new features and innovations that push forward the delivered product, time criticality corresponds with task urgency and deadlines, and risk reduction indicates maintenance tasks that reduce technological and business risks and help retain the product in pristine condition. Both formulas 1 and 2 can be combined into a single formula for calculating the final WSJF score, as presented in formula 3.

$$s = \frac{(v + t + r)}{j} \tag{3}$$

In each sprint, the team runs a prioritization session to prioritize user stories in the backlog. The session is held just before the planning event. The team evaluates the impact of every criterion using the *Fibonacci sequence* (1, 2, 3, 5, 8, 13, 21 etc.) or other estimation techniques and discusses the results for each user story. The calculated scores should also be re-evaluated regularly (Scaled Agile, 2024).

2.2 Data Envelopment Analysis

Data envelopment analysis (DEA) is an approach for determining the relative efficiency of a set of homogeneous decision-making units (DMUs) with usually multiple inputs and outputs to improve their performance (Charnes, Cooper and Rhodes, 1978). This method implies a fundamental assumption that differences in performance between units exist and are measurable (Golany and Roll, 1989). The Charnes, Cooper, Rhodes (CCR) model can be formulated (defined) in two ways. One model emphasizes the reduction of inputs, and the other model emphasizes

the increase of outputs. These two formulations give identical results (does not apply to other DEA models such as the BBC model). The choice of formulated methods is made based on the prevailing circumstances. If in some applications the inputs appear rather inflexible, then in this case it would be more appropriate to define the output. The definition of input is selected if the outputs are closely associated with the objectives established by management (Golany and Roll, 1989). According to Charnes et al. (Charnes, Cooper and Rhodes, 1978) the input-oriented CCR model determines individual input and output weights for each unit in order for the unit to maximize its technical efficiency coefficient ϕ_k. The required conditions are that the weights must not be negative and, when using this set of scales for all units, no coefficient of technical efficiency must be less than one. When maximizing the coefficient of technical efficiency, it is possible to determine such weights that are unrealistic in terms of practice. It follows that the coefficient of technical efficiency is an optimistic estimate of actual efficiency. Thus, the unknown variables in the DEA model are the weights v_{ik} assigned to input i and the weights u_{jk} assigned to output j by unit k. Since the weights are determined for each unit individually, it is necessary to construct p units and solve p models for each special unit, although the models are very similar. Each model has $p + 1$ constraint conditions and $m + n$ variables. The mathematical model for unit H has the following formulation:

$$\phi_H = \frac{\sum_{j=1}^{n} u_{jH} y_{jH}}{\sum_{i=1}^{m} v_{iH} x_{iH}} \rightarrow MAX, \tag{4}$$

is subject to

$$\frac{\sum_{j=1}^{n} u_{jH} y_{jk}}{\sum_{i=1}^{m} v_{iH} x_{ik}} \leq 1, k = 1, 2, \dots, p, \tag{5}$$

$$u_{jH} \geq 0, j = 1, 2, \dots, n, \tag{6}$$

$$v_{iH} \geq 0, i = 1, 2, \dots, m. \tag{7}$$

In this model, a stricter restriction of the weight values is often required, namely

$$u_{jH} \geq \varepsilon, j = 1, 2, \dots, n, \tag{8}$$

$$v_{iH} \geq \varepsilon, i = 1, 2, \dots, m. \tag{9}$$

where ε is a very small positive number – conditions ensure that none of the inputs or outputs has zero weight.

3 Results

3.1 Extension of WSJF model

The WSJF model is based on the ratio of the delay costs to the job size (formula 1). The delay costs (c) are calculated as a simple sum of business value (v), time criticality (t) and risk reduction (r) in formula 2. However, we can assume that all three mentioned components have different significance or weights. Using weights for parameters opens the possibility for the project manager or Scrum Master or Product Owner to set the necessary preferences for individual cost components (delay costs). The simple sum can therefore be rewritten as a weighted sum (formula 10). Using a weighted sum for the delay costs (c^*) also leads to better comparability with the value of the job size (j) – now both parameters from the expression take on comparable values from the same range of the modified *Fibonacci sequence:* 1, 2, 3, 5, 8, 13, 21 (Scaled Agile, 2024).

$$c^* = w_v * v + w_t * t + w_r * r \tag{10}$$

$$w_v + w_t + w_r = 1 \tag{11}$$

For the intended extension of the WSJF model, the job size (j) in formula 1 also needs to be adjusted. If we increase the job size (j) value by the modified delay cost (c^*), we obtain a new WSJF model applicable to the DEA method. The assumption of the DEA method is that the sum of all inputs is always less than or equivalent to the sum of all outputs for a unit or task (formula 5). This condition can therefore be ensured by formula 13 and formula 12 for the expression of the WSJF score.

$$s^* = \frac{c^*}{f(j)} \tag{12}$$

$$f(j) = j + c^* \tag{13}$$

The proposed job size *f(j)* is an initial proposal – the function expressing the work effort of the project team could take different forms or calculations (e.g. including different work contours). The possible shape of the function *f(j)* is a topic for further research by the authors of this paper. The proposal for the extension the WSJF model presented here can be further demonstrated in a case study.

3.2 Case study

An example from the daily practice of a Scrum team when planning a Sprint backlog could look like as presented in table 1:

	Business value	**Time criticality**	**Risk reduction**	**Cost of delay**	**Job size**	**WSJF score**
i	v_i	t_i	r_i	c_i	j_i	s_i
User Story 1	2	1	1	4	5	0,80
User Story 2	3	2	1	6	1	6,00
User Story 3	5	3	2	10	3	3,33
User Story 4	8	5	3	16	5	3,20
User Story 5	5	8	3	16	3	5,33
User Story 6	1	3	5	9	13	0,69
User Story 7	1	2	8	11	20	0,55
User Story 8	3	5	3	11	3	3,67
User Story 9	1	3	2	6	5	1,20

Table 1 Case study with original calculation WSJF method

To determine the values of all parameters (*v*, *t*, *r*, *j*), we used a modified *Fibonacci sequence* commonly used in practice, specifically numbers 1, 2, 3, 5, 8, 13, 21. The resulting score values in the original WSJF method are not suitable for use in the DEA method. If the expressions for the extension of the WSJF model (10), (11), (12) and (13), are used in case study, the obtained values are already usable. See the calculation in table 2.

	Business value	**Time criticality**	**Risk reduction**	**Cost of delay**	**Job size**	**mod. Job size**	**WSJF score**
i	v_i	t_i	r_i	c_i^*	j_i	$f(j_i)$	s_i^*
User Story 1	2	1	1	1,2	5	6,2	0,19
User Story 2	**3**	**2**	**1**	**1,7**	**1**	**2,7**	**0,63**
User Story 3	5	3	2	2,9	3	5,9	0,49
User Story 4	8	5	3	4,6	5	9,6	0,48
User Story 5	5	8	3	4,9	3	7,9	0,62
User Story 6	1	3	5	3,6	13	16,6	0,22
User Story 7	1	2	8	4,8	20	24,8	0,19
User Story 8	3	5	3	3,6	3	6,6	0,55
User Story 9	1	3	2	2,1	5	7,1	0,30
weights	*0,2*	*0,3*	*0,5*				

Table 2 Case study with calculation extension WSJF method

Although the DEA method is based on the construction of linear programming models and the use of the simplex algorithm, for the purposes of this paper and for possible project management practice, it is appropriate and acceptable to use a graphical solution of the DEA method. To illustrate the applicability of the extension of WSJF model, a graphical solution for the DEA method will suffice (see figure 1).

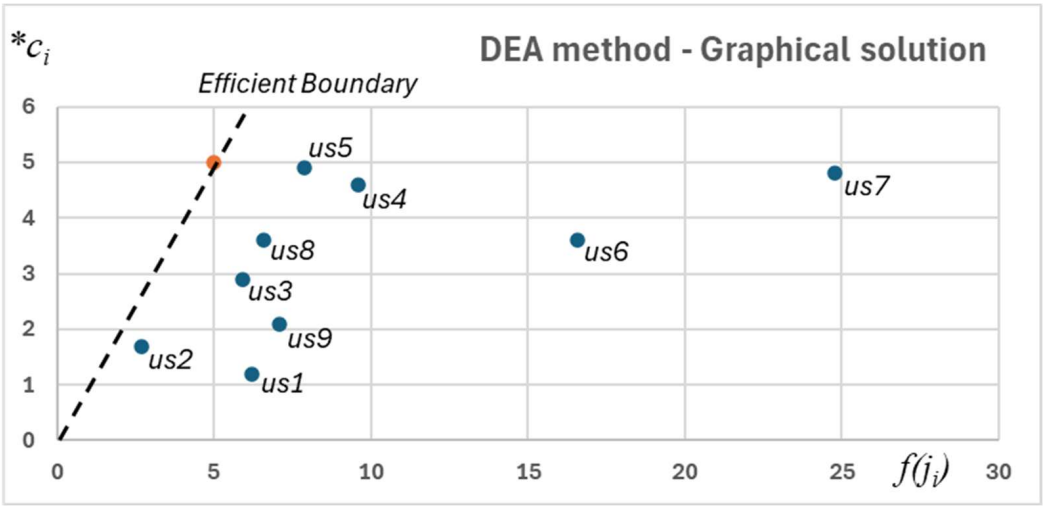


Figure 1 Case study with extension of WSJF model as DEA graphical solution

The dashed line in the graph represents the efficient boundary for the graphical DEA solution. The closest unit (task) to the efficient boundary is User Story 2 (us2 in the graph) - this observation is also confirmed by the highest calculated WSJF score for User Story 2 (see also table 2). The interpretation used here is applicable to Project Management – each activity in a project can be assessed as a task based on the costs incurred to the project outcome (the ratio of inputs to outputs of the project). All other units (tasks) in the case study are less efficient, i.e. they have a lower profitability for their working off (the purpose of the WSJF model is preserved).

4 Conclusions

The paper deals with the possibility of extending the WSJF model in Agile Project Management (Scrum team). The WSJF model uses *Fibonacci sequence* to determine parameter values: business value (v), time criticality (t), risk reduction (r) and job size (j). The use of the WSJF model is based on the ratio of the mentioned parameters, where the purpose is to determine the priority of individual tasks for development (order for progress). The extension of the existing model consists in adapting the calculations to the DEA method, where the assumption of the size of inputs to outputs is maintained. The benefit of the proposed extension is the ability to interpret the planned tasks for the project team using both graphical and mathematical interpretation of the DEA method.

The authors of the paper see a research opportunity in the deeper quantification of methods and approaches in agile project management. Tracking the progress of an agile team or evaluating tasks for an agile team is a key area of current project management practice and theory.

References

Dingsøyr T, Falessi D, Power K (2019). Agile Development at Scale: The Next Frontier. IEEE Software 36:30–38.

Golany B, Roll Y (1989) An Application Procedure for DEA. OMEGA Int. J. of Mgmt Sci. 17(3):237-250. https://doi.org/10.1016/0305-0483(89)90029-7

Charnes A, Cooper WW, 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

Chiang E, Atre N, Sadok H (2022). Robust heuristics: attacks and defenses for job size estimation in WSJF systems. In: Proceedings of the SIGCOMM '22 Poster and Demo Sessions, Sigcomm '22. ACM 52–54. https://doi.org/10.1145/3546037.3546062

Rashid E, Gunjan VK (2025). Achieving Greater Agility Through Scaling Frameworks—An Empirical Analysis. In: Proceedings of the Third International Conference on Cognitive and Intelligent Computing, Volume 1. Springer Nature Singapore, 895–906. https://doi.org/10.1007/978-981-97-9262-7_77

Reinertsen DG (2009) The principles of product development flow. Celeritas Pub. ISBN 978-1935401001.

Riesener M, Kuhn M, Keuper A, Lauf H, Solanki N, Schuh G (2023). Methodology to Determine the Cost of Delay in Projects to Improve Project Prioritization. In: 2023 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE 1–5. https://doi.org/10.1109/ieem58616.2023.10406834

Scaled Agile Inc. (2024) Scaled Agile Framework 6.0. Online: https://scaledagileframework.com/

Schwaber K, Sutherland J (2020) The Scrum Guide. Online: https://scrumguides.org

Tran H (2020). Software Development Effort Estimation Using a Fuzzy Logic-Based System within the Context of the Scaled Agile Framework. IOSR Journal of Computer Engineering (IOSR-JCE). II 22:10–19. https://doi.org/10.9790/0661-2201021019

Uludağ Ö, Putta A, Paasivaara M, Matthes F (2021) Evolution of the Agile Scaling Frameworks. In: Agile Processes in Software Engineering and Extreme Programming. Springer International Publishing 123–139. https://doi.org/10.1007/978-3-030-78098-2_8

Verwijs C, Russo D (2023) Do Agile Scaling Approaches Make A Difference? An Empirical Comparison of Team Effectiveness Across Popular Scaling Approaches. https://doi.org/10.48550/arXiv.2310.06599

Webber R (2022). Unlocking Agile's Missed Potential. Wiley. https://doi.org/10.1002/9781119849117

Received: 2025-04-21 Last revision: 2025-04-21 Accepted: 2025-07-23

News Sentiment Aware Portfolio Optimization Strategy based on Crisis Regime Switching

Qian Gao[1]

Abstract. This study introduces an adaptive portfolio optimization framework that systematically integrates time-varying news sentiment signals with crisis-responsive regime switching, addressing the persistent challenge of regime shifts in financial markets. By leveraging VADER sentiment analysis augmented by news popularity decay weighting, we construct a dynamic sentiment index that captures evolving market narratives from news streams. The core technical innovation lies in a hybrid optimization architecture combining dynamic covariance estimation with sentiment momentum-adjusted return forecasting. At the same time, a self-tuning crisis module activates minimum-volatility allocation when sentiment breaches thresholds of its rolling baseline. The framework demonstrates robust performance through comprehensive backtesting on the Dow Jones component with a 1.46 Sharpe ratio and 2.27 Calmar ratio. These findings illustrate the framework's capacity to enhance return opportunities through sentiment-informed signals while maintaining controlled risk exposure during market dislocations.

Keywords: Crisis regime switching, news sentiment driven, portfolio optimization, strategy comparison

JEL Classification: C45, C58, G11, G17
AMS Classification: 91G70, 68T07, 90C90, 62P05

1 Introduction

The evolution of portfolio optimization methodologies has been significantly shaped by the persistent challenges posed by traditional frameworks in adapting to dynamically shifting financial markets. Modern portfolio theory (MPT) rooted in Markowitz's seminal mean-variance optimization framework, assumes stable return distributions and static correlations (Markowitz, 1952). However, due to its reliance on static historical data, this approach often exhibits rigidity when confronted with abrupt market volatility. Existing methods largely fail to incorporate aggregate market sentiment, which empirical studies have demonstrated to influence investor decision-making. Portfolios constructed with explicit consideration of investor sentiment tend to exhibit greater robustness (Koratamaddi et al., 2021).Assigning importance weights to news items has also proven to be notably effective (Uhl et al., 2015).

A strong yet complex relationship exists between market sentiment and news flow (Mitra et al., 2009). Textual analysis has proven valuable in predicting stock returns, and comparative studies of return patterns across different news types may enhance understanding of how markets process information (Heston and Sinha, 2017). The increasing availability of news and investors' enhanced access to information exert substantial influence on equity price movements, as news rapidly translates into investor sentiment that subsequently drives price action. This process ultimately affects asset prices, volatility, and risk metrics (Allen et al., 2019). Empirical results demonstrate that sentiment-augmented Black-Litterman models yield superior performance (Hung et al., 2024).

This study propose the Enhanced News Sentiment Aware-Crisis Regime Switching Portfolio (ENSA-CRSP) strategy to address these gaps and establish a temporally stratified framework integrating hierarchical sentiment processing and volatility-governed optimization. Its architecture operates through three synchronized temporal layers: news sentiment analysis applies popularity decay weighting to prioritize emergent market narratives, portfolio construction employs expanding-window covariance estimation to capture evolving asset relationships, and regime transitions trigger volatility-conditional capital redistribution. The strategy quantifies sentiment momentum as directionally normalized deviations between instantaneous sentiment intensity and its moving historical baseline, with nonlinear logarithmic scaling enhancing persistent market sentiment signals while suppressing transient noise.

The crisis regime switching employs conditionally-triggered defensive configurations. The system autonomously activates crisis management protocols when real-time sentiment breaches predefined thresholds. The result demonstrates that transitions between "stable" and "risky" market regimes significantly influence asset price

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

co-movements (Guo et al., 2011). These protocols redistribute portfolio weights based on historical volatility rankings, allocating predominant capital to minimum-volatility assets while retaining partial exposure to high-volatility instruments. This dual-objective allocation addresses the critical challenge of balancing risk containment with market participation during extreme conditions.

A strictly causal operationalization governs the architecture, where sentiment indicators undergo systematic temporal lagging to prevent forward-looking contamination. Portfolio rebalancing executes through constrained nonlinear optimization that synchronizes three operational dimensions: rolling-window covariance estimation integrating cumulative market dynamics, decay-weighted sentiment momentum factors reflecting news impact decay patterns, and full-period historical volatility profiling capturing long-term risk signatures. Investor mandates are enforced via position caps and non-negativity constraints, with iterative weight normalization ensuring implementable solutions. Numerical stabilization emerges through covariance matrix conditioning and constrained optimization techniques, maintaining algorithmic stability across diverse market regimes.

This paper is structured as follows. Section 2 introduces the proposed methodology and details the implementation of the baseline models. Section 3 presents a comparative analysis of the empirical results to evaluate the effectiveness of the proposed approach. Section 4 discusses the key findings and concludes the study.

2 Methodology

2.1 Enhanced News Sentiment Aware-Crisis Regime Switching Portfolio

The Enhanced News Sentiment Aware-Crisis Regime Switching Portfolio (ENSA-CRSP) implements dynamic asset allocation by integrating text mining techniques and quantitative modeling. The technical process begins with sentiment analysis of unstructured news texts, utilizing the VADER sentiment analysis framework (Hutto and Gilbert, 2014) to compute a compound sentiment score for each news item. These scores are then aggregated over multiple days using a time-decay weighting scheme to construct a standardized sentiment indicator that captures the short-term momentum effects of market sentiment. Learning-to-rank algorithms are effective in this context and align with the application of investor sentiment toward specific stocks (Song et al., 2017).

The strategy extracts short-term sentiment trends using time-decay weighted aggregation. The sentiment momentum factor is defined as

$$\Delta S_t = S_t - \frac{\sum_{i=1}^{T} w_i \cdot S_{t-i}}{\sum_{i=1}^{T} w_i}, \quad w_i = 0.95^{i-1}, \tag{1}$$

where S_t is the normalized news sentiment score at day t (range $[-1, 1]$), $T = 60$ is the news lookback window, and w_i represents exponential decay weights emphasizing top popularity news.

The sentiment factor is embedded into an enhanced mean-variance optimization model through a nonlinear transformation, wherein a sign-sensitive logarithmic adjustment term is introduced into the objective function to modulate the portfolio's responsiveness to shifts in sentiment dynamically. On the risk modeling front, the strategy employs the sample covariance matrix estimated during the training period to capture the risk-return structure among assets while incorporating a sentiment-threshold-triggered regime-switching mechanism to address extreme market events.Under normal market conditions, portfolio weights are determined by solving

$$\begin{aligned} \max_{\mathbf{w}_t} \quad & \underbrace{\mathbf{w}_t^T \boldsymbol{\mu}_t}_{\text{Expected Return}} + \underbrace{\gamma \cdot \text{sign}(\Delta S_t) \cdot \ln(1 + |\Delta S_t|)}_{\text{Sentiment Momentum Adjustment}}, \\ \text{s.t.} \quad & \mathbf{w}_t \geq 0, \quad \sum_{i=1}^{N} w_{t,i} = 1, \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_t \in \mathbb{R}^N$ is expected return vector (historical mean), $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$ is training sample covariance matrix, $\gamma > 0$ is sentiment momentum intensity, $\text{sign}(\cdot)$ is sign function preserving sentiment directionality

When the sentiment indicator exceeds a predefined critical threshold,

$$S_t < \theta_{\text{crisis}}, \quad \theta_{\text{crisis}} = -0.2, \tag{3}$$

the system abandons conventional optimization logic and shifts to a defensive allocation regime based on identifying historical volatility extremes. In this regime, most capital is reallocated to low-volatility assets to reduce downside risk exposure. In contrast, a limited position in high-volatility assets is retained to hedge against potential liquidity

shocks. This mechanism enhances portfolio robustness under crisis conditions. The strategy resets weights to

$$\mathbf{w}_t = \alpha \cdot \mathbf{w}_{\text{min-var}} + (1 - \alpha) \cdot \mathbf{w}_{\text{max-var}} \tag{4}$$

where $\mathbf{w}_{\text{min-var}}$ is single-asset vector with $w_{\text{min-var},i} = \begin{cases} 1 & \text{if } i = \arg\min(\boldsymbol{\sigma}) \\ 0 & \text{otherwise} \end{cases}$, $\mathbf{w}_{\text{max-var}}$is similarly defined for $\arg\max(\boldsymbol{\sigma})$, $\boldsymbol{\sigma} \in \mathbb{R}^N$ is asset volatility vector (training-sample standard deviation), $\alpha = 0.8$ is defensive allocation ratio.

2.2 Benchmark strategy

The Equally Weighted Portfolio (EWP) is constructed under the assumption of complete diversification, allocating available capital equally across all constituent assets. The Minimum Variance Portfolio (MVP) seeks the risk-optimal allocation by minimizing portfolio variance based on historical asset covariances. Which is expressed as

$$\mathbf{w}^* = \arg\min \quad \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \quad \text{s.t.} \qquad \sum_{i=1}^{n} w_i = 1, \quad w_i \geq 0, \tag{5}$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$ denotes the covariance matrix estimated from historical returns.

The Attention Portfolio (ATTP) strategy employs an end-to-end temporal modeling framework, integrating standardized asset returns and news sentiment indicators into a multi-dimensional feature matrix. The formula as

$$\mathbf{w}_t = \text{Softmax}\Big(\mathbf{W}_o \cdot \text{MultiHead}\big(\text{PositionEncode}(\mathbf{X}_t)\big)\Big), \tag{6}$$

where $\mathbf{w}_t \in \mathbb{R}^N$ is portfolio weight vector (for N assets), $\mathbf{W}_o \in \mathbb{R}^{N \times d_{\text{model}}}$ is output layer weight matrix ($d_{\text{model}} = 128$), MultiHead$(\cdot)$ is multi-head self-attention aggregation operator (4 heads), PositionEncode$(\cdot)$ is finusoidal positional encoding function, $\mathbf{X}_t \in \mathbb{R}^{L \times d}$ is input feature matrix ($d = 36$ features, 30 asset returns + 3 sentiment features + 3 lagged sentiment features). MultiHead$(\cdot)$ is cross-temporal attention with $h = 4$ heads,

$$\text{head}_i = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \ \mathbf{Q}/\mathbf{K}/\mathbf{V} = \mathbf{X}_t \mathbf{W}_i^{Q/K/V}, \tag{7}$$

where $\text{head}_i \in \mathbb{R}^{L \times d_v}$ is output of the i-th attention head ($d_v = 32$), $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{L \times d_k}$is query and key matrices ($d_k = 32$), $\mathbf{V} \in \mathbb{R}^{L \times d_v}$ is value matrix, $\mathbf{W}_i^{Q/K/V} \in \mathbb{R}^{d_{\text{model}} \times d_k / d_v}$ is trainable projection matrices for the i-th head.

In contrast, the News Sentiment Aware Portfolio (NSAP) strategy focuses on linear risk exposure to news sentiment factors. It employs a rolling regression to estimate each asset's sensitivity coefficient to sentiment shocks and constructs expected returns by element-wise multiplication of sentiment values and sensitivity vectors. The formula as

$$r_{i,\tau} = \alpha_i + \beta_i S_\tau + \epsilon_{i,\tau}, \tau \in [t - L, t - 1], \tag{8}$$

where $r_{i,\tau} \in \mathbb{R}$ is return of asset i at time τ, $\alpha_i \in \mathbb{R}$ is intercept term for asset i, $\beta_i \in \mathbb{R}$ is sensitivity coefficient of asset i to news sentiment, $S_\tau \in [-1, 1]$ is standardized news sentiment score at time τ, $\epsilon_{i,\tau} \sim \mathcal{N}(0, \sigma^2)$ is the Gaussian error term, $L = 60$ is rolling regression window length (days).

2.3 Data and basic framework

This study employed a publicly available news dataset (Sun, 2016) with the 30 compont stocks of the Dow Jones Industrial Average on corresponding trading days. The underlying framework follows a rolling window approach, wherein the dataset is split into two segments: 80% for training and 20% for testing. The window length is fixed at 80% of the dataset, and the window rolls forward by 30 trading days at each iteration until all data within the testing set have been evaluated. This overlapping window design ensures temporal continuity while avoiding look-ahead bias. The sentiment threshold is determined based on the average sentiment over the preceding 60 trading days for the crisis regime-switching mechanism.

3 Empirical results

The figure 1 illustrates the temporal dynamics of cumulative returns for five out-of-sample investment strategies. The horizontal axis spans a 19-month market cycle, while the vertical axis displays standardized cumulative returns, ranging from 0.9 to 1.5. In the third quarter of 2015, all strategies experienced a simultaneous drawdown

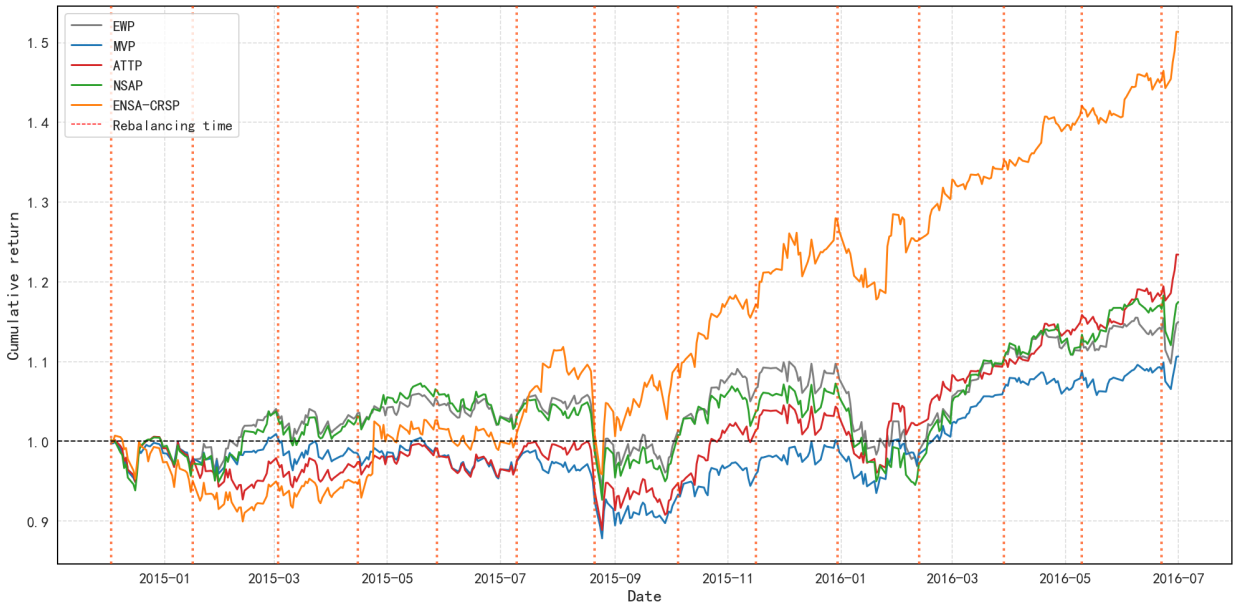


Figure 1 Comparison of out-of-sample cumulative returns of different strategies

toward the 0.9 level, forming a distinct response band to a major market stress event. Both the ENSA-CRSP and ATTP strategies exhibit strong trend-following characteristics. ENSA-CRSP began accelerating upward in October 2015, reaching a peak of around 1.5, while ATTP surpassed the 1.23 level in July 2016. Both strategies achieved slightly higher average monthly return growth during the recovery period than the others. In contrast, the EWP and MVP strategies demonstrated low-volatility trajectories, with cumulative returns fluctuating narrowly between 1.1 and 1.3, reflecting their risk-averse profiles. The NSAP strategy followed an inverse pattern relative to ENSA-CRSP—showing relatively strong early performance but aligning with the others post-event.

This dynamic adjustment mechanism gives rise to a distinctive return distribution profile. As seen in the first half of 2016, the ENSA-CRSP strategy tracked the benchmark during most periods of market stability. However, in response to sudden downturns—such as that of January 2016—it experienced significantly milder drawdowns and rebounded faster. This "slow-fall-fast-rise" pattern results in a right-skewed return curve. It is driven by the strategy's ability to automatically increase exposure to defensive (low-volatility) assets during periods of risk accumulation and swiftly rotate into high-potential assets once market panic subsides. The red dashed lines in the figure, representing rebalancing points, indicate that the transition is not carried out through mechanical operations but is intelligently triggered at rebalancing times based on sentiment score thresholds computed from historical data. This key mechanism enables the strategy to maintain stability across different market cycles.

The figure 2 comprehensively presents the risk characteristics of five distinct investment strategies during the out-of-sample period from January 2015 to July 2016. As illustrated, the ENSA-CRSP strategy exhibited consistently higher volatility levels than alternative approaches throughout most of the observation window, though its volatility profile became comparatively lower post-March 2016. This elevated volatility pattern directly correlates with the strategy's active risk-taking mechanism—its dynamic adjustment of risk asset exposure inherently generates greater short-term fluctuations when capturing market trends. Notably, while the MVP strategy maintained sub-0.01 volatility under normal market conditions, its volatility surged to approximately 0.03 during extreme events, revealing inherent limitations in conventional models' adaptability to tail events.

The drawdown panel further substantiates the symmetry principle between risk and return. All strategies demonstrated similar distribution patterns, with concentrated drawdowns occurring during the late August 2015 market stress event, where the ENSA-CRSP strategy experienced 0.02-0.04 deeper drawdowns than peers within several days—consistent with its higher-risk exposure during volatile periods. Crucially, the strategy displayed superior recovery dynamics, achieving comparatively shallower drawdowns in subsequent months following market troughs. A cross-analysis of both panels indicates that the system promptly increases risk asset allocations when volatility falls below predetermined thresholds to capture rebound opportunities. By actively managing risk exposure, the ENSA-CRSP strategy successfully converts higher volatility and drawdowns into sustained excess return accumulation, thereby validating the efficacy of its risk compensation framework.

The table 1 demonstrates significant disparities in the risk-return profiles across different portfolio strategies. The

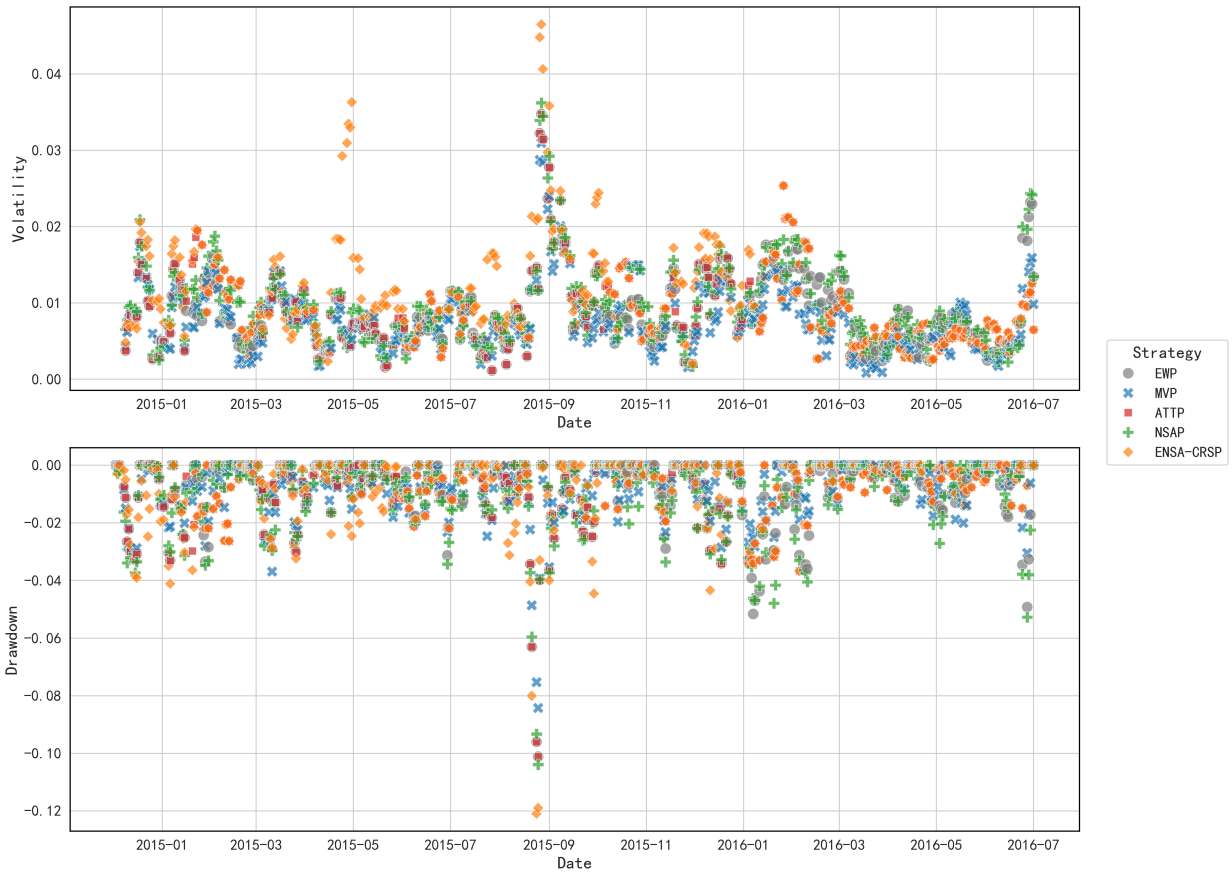


Figure 2 Risk characteristics in out-of-sample

Performance	EWP	MVP	ATTP	NSAP	ENSA-CRSP
Cum Returns	1.150	1.106	1.234	1.175	**1.514**
Annual Return	0.106	0.076	0.157	0.123	**0.324**
Annual Volatility	0.158	**0.135**	0.159	0.169	0.193
Max Drawdown	0.119	0.130	**0.116**	0.136	0.143
Sharpe Ratio	0.638	0.541	0.915	0.688	**1.457**
Calmar Ratio	0.889	0.585	1.350	0.904	**2.268**

Table 1 Comparison of out-of-sample performance of different strategies

ENSA-CRSP strategy demonstrates superior performance with cumulative returns of 1.514 and annualized returns of 0.324, though accompanied by a 42% higher annualized volatility (0.193) compared to the benchmark and a maximum drawdown of 0.143 exceeding that of the news sentiment strategy (0.116). This observation substantiates the strategy's fundamental mechanism of pursuing excess returns through deliberate risk assumption. In contrast, the MVP strategy maintains optimal risk control with the lowest volatility (0.135). However, its annualized return of 0.076 represents merely 23% of the ENSA-CRSP's performance, revealing inherent limitations in return generation within conventional modeling frameworks. Notably, the ATTP strategy exhibits an intriguing equilibrium between its maximum drawdown (0.116) and annualized return (0.157), suggesting unique value in tail risk mitigation while demonstrating incomplete conversion into return advantages.

From the perspective of risk-adjusted returns, the ENSA-CRSP strategy achieves a Sharpe ratio of 1.457 and a Calmar ratio of 2.268, outperforming the second-best strategy by 59% and 68%, respectively. This enhancement stems from its hierarchical optimization framework that asymmetrically manages downside risks. Although the strategy exhibits elevated absolute volatility, its dynamic risk rebalancing mechanism—activated by crisis transition triggers during market stress periods—effectively constrains the frequency and magnitude of extreme losses. Conversely, despite generating cumulative returns of 1.234, the attention strategy manifests structural deficiencies in risk-return optimization with a Sharpe ratio of 0.915, attributable to its singular reliance on market

attention factors. These findings collectively validate the framework's capacity to reconcile growth objectives with tail risk aversion, thereby establishing competitive advantages in complex market environments.

4 Conclusion

This study develops an adaptive portfolio framework that systematically integrates news sentiment dynamics with regime-switching risk controls. The methodology introduces three key innovations: a temporal sentiment index constructed through exponentially decaying VADER polarity scores with position-based weighting, where popularity news impacts decay geometrically at 0.95; a hybrid optimization engine combining covariance matrix estimation with sentiment momentum factors, employing rolling 60-day windows for adaptive parameter calibration; and a circuit-breaker mechanism activating minimum-volatility allocation when sentiment indices breach the -0.2 threshold relative to their 60-day moving average.

The proposed framework demonstrates strong performance in out-of-sample testing on Dow Jones component. The ENSA-CRSP strategy achieves superior risk-adjusted returns with a Sharpe ratio of 1.457 and Calmar ratio of 2.268, significantly outperforming equal-weight (0.638/0.889) and minimum-variance (0.541/0.585) benchmarks. While exhibiting higher annual volatility (19.3% vs 15.8% for EWP), it delivers 32.4% annualized returns with 14.3% maximum drawdown. The ATTP strategy shows competitive drawdown control (11.6% maximum) and superior Sharpe ratio (0.915). Future research will extend the framework positioning it as a scalable foundation for next-generation adaptive portfolio management systems.

Acknowledgements

The research was supported by the Student Grant Competition of the Faculty of Economics, VSB - Technical University of Ostrava, project no. SP2025/043.

References

Allen DE, McAleer M, Singh AK (2019) Daily market news sentiment and stock prices. Applied Economics 51(30):3212–3235. https://doi.org/10.1080/00036846.2018.1564115

Guo F, Chen CR, Huang YS (2011) Markets contagion during financial crisis: A regime-switching approach. International Review of Economics & Finance 20(1):95–109. https://doi.org/10.1016/j.iref.2010.07.009

Heston SL, Sinha NR (2017) News vs. Sentiment: Predicting Stock Returns from News Stories. Financial Analysts Journal 73(3):67–83. https://doi.org/10.2469/faj.v73.n3.3

Hung MC, Hsia PH, Kuang XJ, Lin SK (2024) Intelligent portfolio construction via news sentiment analysis. International Review of Economics & Finance 89:605–617. https://doi.org/10.1016/j.iref.2023.07.103

Hutto C, Gilbert E (2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media 8(1):216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Koratamaddi P, Wadhwani K, Gupta M, Sanjeevi SG (2021) Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation. Engineering Science and Technology, an International Journal 24(4):848–859. https://doi.org/10.1016/j.jestch.2021.01.007

Markowitz H (1952) Portfolio Selection. The Journal of Finance 7(1):77–91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

Mitra L, Mitra G, Dibartolomeo D (2009) Equity portfolio risk estimation using market information and sentiment. Quantitative Finance 9(8):887–895. https://doi.org/10.1080/14697680903448361

Song Q, Liu A, Yang SY (2017) Stock portfolio selection using learning-to-rank algorithms with news sentiment. Neurocomputing 264:20–28. https://doi.org/10.1016/j.neucom.2017.02.097

Sun J (2016) Daily News for Stock Market Prediction. https://www.kaggle.com/datasets/aaron7sun/stocknews

Uhl MW, Pedersen M, Malitius O (2015) What's in the News? Using News Sentiment Momentum for Tactical Asset Allocation. Journal of Portfolio Management 41(2):100–112,8,10.

Received: 2025-06-06 Last revision: 2025-06-06 Accepted: 2025-08-20

Economic Effects of Military Expenditures in Czechia

Petra Ondráčková[1], Jakub Odehnal[2], Jiří Neubauer[3]

Abstract. The changing security situation in Europe brings natural pressure for an increase in military expenditures of European countries. This paper examines the relationship between military expenditures and economic growth in Czechia during the period 1993–2023. The aim of this paper is to test the short-run and long-run equilibrium relationship between military expenditures and economic growth in Czechia using the method of Autoregressive Distributed Lag (ARDL). The authors have used seven economic variables to identify the relationship based on the results of the estimated model. There is examined the link between economic growth and selected economic characteristics, namely military expenditures, government non-defence expenditures, government investment, population, and trade openness represented as a sum of import and export. In the short run, the findings present negative and statistically significant effects mainly of government non-defence expenditures. However, the long-run estimates reveal the existence of a positive effect of defence spending on the economic growth of Czechia likewise trade openness is found to be positive and statistically significant. The existence of a long-run relationship between the analysed variables was confirmed and the stability of the above relationship was also verified. The findings of the study suggest that military expenditures can be one of the factors stimulating economic development, which is consistent with Keynesian Economic theory. These results could have further political implications, as national security concerns need to be carefully balanced against the need for sustainable economic growth, while at the same time fulfilling NATO's commitment.

Keywords: military expenditures, economic growth, ARDL

1 Introduction

Both economy of state and its efficiency belong to the factors influencing levels of governmental expenditures, military expenditures included. Military expenditures are one of the most controversial types of governmental expenditures, often presented as a non-productive effort of governmental spending which should be used rather for providing some additional benefits, for instance to either health system or education. However, the European countries have been presently exposed to a need to bolster military expenditures not only to cover a commitment to NATO, but mainly due to the ongoing Russian-Ukrainian war. The development of the economic environment characterized by gross domestic product (GDP), inflation, budget deficit or the level of indebtedness is one of the so-called economic determinants of military expenditures, which is the subject of several empirical studies examining the influence of defined factors on the size of military expenditures.

Based on the authors Desli and Gkoulgkoutsika (2021), military expenditures could positively influence an economical growth by stimulating an aggregating demand for both goods and services (so called "Military Keynesianism"), or by using investment from the field of military experiments to the private sector. The findings claim that the overall impact of military expenditures on economic growth has been negative from a global perspective over the period 1960–2017.

Waszkiewicz (2018) analysed a relationship between military expenditures and economic growth within the Visegrad group of countries and examined both short and long-standing cause between these mentioned economic variables. To examine these variables, the author used a VAR method and the Granger Causality Test. This study has identified relationship between aggregated military spending and GDP as well as influence of economic growth on four separated areas (material, infrastructure, personnel, and other). An empiric test confirmed the absence of long-standing dependence of military expenditures on economic growth. The study also confirmed that finances

[1] University of Defence, Faculty of Military Leadership, Kounicova 65, 662 01, Brno, Czechia, petra.ondrackova@mo.gov.cz.
[2] University of Defence, Faculty of Military Leadership, Kounicova 65, 662 01, Brno, Czechia.
[3] University of Defence, Faculty of Military Leadership, Kounicova 65, 662 01, Brno, Czechia.

to the military sector contribute in the short term to the economic growth only in Czechia. The results of the study did not show any short-term impact in case of Slovakia, nor Hungary.

The other study is from Topal et al. (2022), focusing on the analysis of 27 NATO countries in different time periods using Granger Causality Test. The empirical results showcased existing both symmetric and asymmetric Granger causality between military expenditures and the economic growth differing for each country.

In 2021, Syed published the results of the analysis focused on the relationship between military expenditures and industrial productivity by using an Autoregressive Distributed Lag method (ARDL) on the dates of India, China, and Pakistan covering a 1990–2018 period. This study has been trying to prove the principles of Military Keynesianism stating that military expenditures are "good stimulus" for economic expansion, and it is one of the tools of the state fiscal politics. The results of the study have proved an asymmetric integration between military expenditures and economic growth in India and China and that military expenditures in these countries have a long-standing impact on the economic. From the short period perspective, the relationship seems to be symmetrical. The overall results undoubtedly present that both China and India benefit from long-standing investment to the armed forces, which is in line with Military Keynesianism stating that with the higher investment to military sector, there is a growing demand for workforce, industrial expansion, source allocation, and investment opportunity.

In 2024, Dimitraki and Emmanouilidis were analysing the economic effects of military expenditures in Spain for a 1954–2021 period by using both, ARDL and Kernel-Based Regularized Least Squares (KRLS) methodologies. They tested the relationship between military expenditures and economic growth from the short and long period perspective. The results are consistent with Military Keynesianism theory describing the stimulus of aggregative demand in the short term. In the long term, the relationship between the variables of interest becomes weaker, albeit still positive.

Based on the analysis of the above selected studies, this paper is focused on the short-run and long-run equilibrium relationship between military expenditures and economic growth in Czechia by using the ARDL method. The authors have used seven economic variables to identify the relationship based on the results of the estimated model, which follows the Dimitraki and Emmanouilidis approach based on the Barro-style model. There is examined the link between economic growth and selected economic characteristics, namely military expenditures, government non-defence expenditures, government investment, population, and trade openness represented as a sum of import and export. The main data were sourced from the World Bank, military expenditures were taken from the SIPRI database. The Barro-style specification controls for variation in the size of central government (military expenditures and other government spending) as an explanatory variable for economic growth. Other control variables are governmental investment, openness to trade and labour. This study uses annual time-series data covering a 30-year period from 1993 to 2023.

2 Data and methods

The authors employed the Autoregressive Distributed Lag (ARDL) model (Pesaran and Shin, 1999) to analyse the relationship between GDP per capita and select economic characteristics. These characteristics included military expenditures as a percentage of GDP, government expenditures as a percentage of GDP, openness as measured by exports and imports, gross capital formation as a percentage of GDP, and population growth.

The authors applied an autoregressive distributed lag model ARDL(p, q_1, q_2, ..., q_k), where p is the number of lags of the dependent variable Y_t, $q_1, q_2, \dots, q_k$ are numbers of lags of explanatory variables X_{it}, $i = 1,2, \dots, k$. The model can be written as

$$Y_t = \alpha + \sum_{i=1}^{p} \gamma_i Y_{t-i} + \sum_{j=1}^{k} \sum_{i=0}^{q_j} \beta_{j,i} X_{j,t-i} + \varepsilon_t, \tag{1}$$

where ε_t is a one-dimensional zero mean error term. The model can be transformed into a long-run representation that demonstrates the dependent variable's response to a change in the explanatory variables over time (Baltagi, 2011; Mills, 2019). The long run estimates are given by (Pesaran and Shin, 1999; Pesaran et al. 2001)

$$\hat{\theta}_j = \frac{\sum_{i=1}^{q_j} \hat{\beta}_{j,i}}{1 - \sum_{i=1}^{p} \hat{\gamma}_i}. \tag{2}$$

The ARDL approach not only provides a dynamic description but also allows for testing of cointegration. The cointegrated system of time series can be estimated as an ARDL model (Pesaran and Shin, 1999) with the advantage that variables in the cointegrating relationship can be either $I(0)$ or $I(1)$ without needing to specify which are $I(0)$ or $I(1)$. For cointegration analysis, we use the form of (1) in differences.

$$\Delta Y_t = \sum_{i=1}^{p-1} \gamma_i^* \Delta Y_{t-i} + \sum_{j=1}^{k} \sum_{i=0}^{q_j-1} \beta_{j,i}^* \Delta X_{j,t-i} - \hat{\phi} EC_{t-1} + \varepsilon_t, \tag{3}$$

where $EC_t = Y_t - \hat{\alpha} - \sum_{j=1}^{k} \hat{\theta}_j X_{j,t}$, and $\hat{\phi} = 1 - \sum_{i=1}^{p} \hat{\gamma}_i$. (Pesaran et al., 2001) proposed a methodology for testing existence of long-run relationship between independent variable and regressors. For so called bounds testing they use the following representation of (3)

$$\Delta Y_t = \sum_{i=1}^{p-1} \gamma_i^* \Delta Y_{t-i} + \sum_{j=1}^{k} \sum_{i=0}^{q_j-1} \beta_{j,i}^* \Delta X_{j,t-i} - \rho Y_{t-1} - \alpha - \sum_{j=1}^{k} \delta_j X_{j,t-1} + \varepsilon_t. \tag{4}$$

The test for existence of long-run relationship is a test of $\rho = 0$ and $\delta_1 = \delta_2 = \cdots = \delta_k = 0$. The distribution of the test statistic based on (4) depends on the fact whether the regressors are $I(0)$ or $I(1)$. (Pesaran et al., 2001) provide critical values for the cases where all regressors are $I(0)$ and the cases where all regressors are $I(1)$. These critical values are used as bound for the more typical cases where the regressors are a mixture of $I(0)$ and $I(1)$.

3 Results

Variable	Abbreviation	Unit
Gross domestic product per capita	GDP_PC	USD, constant price 2021
Military expenditures	MILEX	% of GDP
Government expenditures	GOV_EX	% of GDP
Openness	OPENNES	% of GDP
Gross capital formation	GC_FORMATION	% of GDP
Population growth	POP_GROWTH	%

Table 1 Variables

The following research model is employed to analyse the relationship between GDP per capita and military expenditures as a percentage of GDP, government expenditures as a percentage of GDP, openness as measured by exports and imports, gross capital formation as a percentage of GDP, and population growth: The Akaike Information Criterion (AIC) is employed for the purpose of model selection, with a maximum lag set at 2. The model ARDL(1, 0, 1, 2, 0, 0) was obtained, see Table 1. Calculations were performed in Eviews software.

Variable	Coefficient	Std. error	t-statistic	p-value
GDP_PC(–1)	0.841	0.067	12.470	**0.000**
MILEX	2.858	1.256	2.275	**0.035**
GOV_EX	–0.594	0.340	–1.745	*0.097*
GOV_EX(–1)	0.497	0.400	1.244	0.229
OPENNESS	0.057	0.038	1.511	0.147
OPENNESS(–1)	–0.016	0.043	–0.367	0.718
OPENNESS(–2)	0.061	0.028	2.205	**0.040**
GC_FORMATION	0.226	0.112	2.029	*0.057*
POP_GROWTH	–0.655	0.301	–2.177	**0.042**
C	–14.173	7.424	–1.909	*0.072*

Table 2 ARDL model

The parameter estimates of the ARDL model in Table 2 show that the lagged value of the GDP PC variable is statistically significant at the 0.05 level, as is the case for the MILEX variable (positive coefficient), the OPENNESS variable for lag 2 (positive coefficient) and POPULATION GROWTH (coefficient is negative). The variables GOV_EX and GC_FORMATION are still significant at the 0.10 significance level. The adjusted R^2 of the model has a value of 0.988 indicating a good fit, see Figure 1.

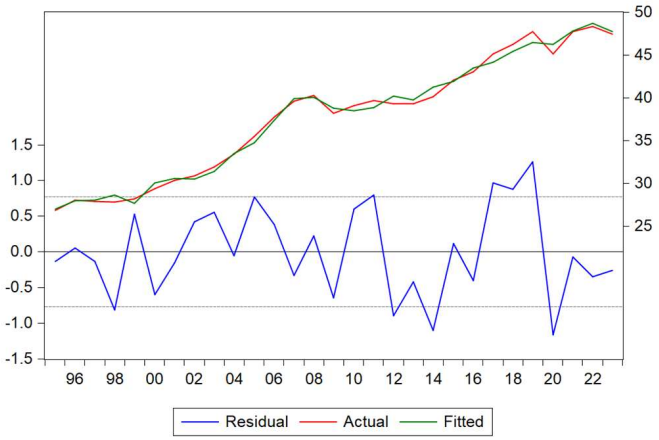


Figure 1 ARDL model

For the purpose of modelling the long-run relationship, we used a model with a restricted constant, the results are shown in Table 3. The existence of a long-run relationship was confirmed using the F-bound test, whose value is 6.16 (the critical value for a significance level of 0.05 is 3.38 for the model). At the 0.05 level of significance, only the variable OPENNES (with a positive coefficient) is statistically significant; at the 0.10 level, a long-run relationship can also be observed for the variable MILEX (positive coefficient).

From a short-run perspective, see Table 4, we can conclude that all variables occurring in the given relationship (3) are statistically significant at the 0.05 level of significance. For the variable d(GOV_EX), the coefficient is negative, but for the variable d(OPENNESS), the coefficient is positive for lag 1, but the overall effect can be considered rather negative.

Variable	Coefficient	Std. error	t-statistic	p-value
MILEX	18.023	9.884	1.824	*0.084*
GOV_EX	–0.609	2.266	–0.269	0.791
OPENNESS	0.648	0.181	3.570	**0.002**
GC_FORMATION	1.427	0.913	1.563	0.135
POP_GROWTH	–4.127	3.005	–1.374	0.186
C	–89.368	45.938	–1.945	*0.067*

Table 3 ARDL model – long run

Variable	Coefficient	Std. error	t-statistic	p-value
D(GOV_EX)	–0.594	0.208	–2.859	**0.010**
D(OPENNESS)	0.057	0.020	2.806	**0.011**
D(OPENNESS(–1))	–0.061	0.021	–2.957	**0.008**
CointEq(–1)*	–0.159	0.021	–7.531	**0.000**

Table 4 ARDL model – short run

4 Conclusion

Military spending in Czechia has evolved and changed over the years, especially in relation to the global situation and political decisions. A more significant increase can be linked to the conclusion of the Wales Summit (2014), where two important commitments were guided to be considered:
"Allies whose current proportion of GDP spent on defence is below this level will halt any decline in defence expenditure, aim to increase defence expenditure in real terms as GDP grows and aim to move towards the 2% guideline within a decade with a view to meeting their NATO Capability Targets and filling NATO's capability shortfalls. Allies who currently spend less than 20% of their annual defence spending on major new equipment, including related Research & Development, will aim, within a decade, to increase their annual investments to 20% or more of total defence expenditures." Over the last decade, a significant increase in the share of investment relative to the total military budget can be observed. This naturally raises the question of whether this share of investment can have an impact on economic growth.
A lot of empirical research on the economic effects of military expenditures have been published. However, an examination of the relationships between military expenditures and economic growth within the Czech economy has received a little attention. By using the ARDL model, the aim of this study is to contribute to the research focused on the Czech case.
In the short run, the findings present negative and statistically significant effects mainly of government non-defence expenditures. However, the long-run estimates reveal the existence of a positive effect of military expenditures on the economic growth of Czechia likewise trade openness is found to be positive and statistically significant. The existence of a long-run relationship between the analysed variables was confirmed and the stability of the above relationship was also verified. The findings of the study suggest that military expenditures can be one of the factors stimulating economic development, which is consistent with Keynesian Economic theory. These results could have further political implications, as national security concerns need to be carefully balanced against the need for sustainable economic growth, while at the same time fulfilling NATO's commitment, as NATO countries declared at the Wales Summit their commitment to reach 2% of GDP for defence spending.

Acknowledgements

The paper was supported by institutional funding aimed at the development of the research organization (Long-Term Organisation Development Goals at the FML).

References

Baltagi, B H (2011) Econometrics, 5th ed. New York: Springer.

Dimitraki O, Menla Ali F (2015)The long-run causal relationship between military expenditure and economic growth in China: revisited. Defence and Peace Economics. 26(3):311-26. https://doi.org/10.10080/10242694.2013.810024

Dimitraki O, Emmanouilidis K (2024) Analysis of the economic effects of defence spending in Spain: a re-examination through Dynamic ARDL simulations and Kernel-Based Regularized Least Squares. Defence and Peace Economics.35(7):908-30. https://doi.org/10.1080/10242694.2023.2245698.

Desli, E., Gkoulgkoutsika, A (2021) Military spending and economic growth: a panel data investigation. Economic Change and Restructuring, 54: 781–806. https://doi.org/10.1007/s10644-020-09267-8.

Emmanouilidis K, Karpetis C. (2021)The effects of military expenditures on economic growth and inflation: evidence from Turkey. Peace Economics, Peace Science and Public Policy. 27(3):369-404. https://doi.org/10.1515/peps-2020-0058.

Mills TC (2019) Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting. London: Academic Press.

NATO, NATO-Official Text: Wales Summit Declaration. Available at: https://www.nato.int/cps/en/natohq/official_texts_112964.html.

NATO, Defence Expenditure of NATO Countries (2014-2024) https://www.nato.int/nato_static_fl2014/assets/pdf/2024/6/pdf/240617-def-exp-2024-en.pdf.

Nikolaidou E(2008) The demand for military expenditure: Evidence from the EU15 (1961-2005). Defence and Peace Economics 19(4):121-128. Available at: https://doi.org/10.1080/10242690802166533.

Odehnal J (2016) Free-Riding Problem in Alliance. Vojenské rozhledy – Czech Military Review, 25(5): 49-60. Available at: https://www.vojenskerozhledy.cz/en/kategorie/cerne-pasazerstvi-v-alianci.

Ondrackova P, Odehnal J, Neubauer J (2024) The Impact of Military Expenditures on the Indebtedness of the Czech Republic and Lithuania. In Proceedings of the Challenges to National Defence in Contemporary Geopolitical Situation, Brno, Czech Republic, 11-13 September 2024.

Pesaran MH, Shin Y, Smith, RJ (2001) Bounds Testing Approaches to the Analysis of Level Relationships. Journal of Applied Econometrics 16: 298–326.https://doi:10.1002/jae.616.

Pesaran, MH, Shin, Y (1999) An Autoregressive Distributed Lag Modelling Approach to Cointegration Analysis. In Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium. Edited by S. Strom. Cambridge: Cambridge University Press.

Syed A A (2021) The Asymmetric Relationship Between Military Expenditure, Economic Growth and Industrial Productivity: An Empirical Analysis of India, China and Pakistan Via the NARDL Approach. Revista Finanzas y Política Económica, 13(1): 77-97. https://doi.org/10.14718/revfinanzpolitecon.v13.n1.2021.4.

Topal MH, Unver M, Türedi S (2022) The Military Expenditures and Economic Growth Nexus: Panel Bootstrap Granger Causality Evidence from NATO Countries. Panoeconomicus, 69(4): 555-578. https://doi.org/10.2298/PAN170914002T.

Waszkiewicz G (2018) Defence Spending and Economic Growth in the Visegrad Countries. Panoeconomicus, 67(4): 539-556. http://dx.doi.org/10.2298/PAN170709029W.

Received: 2025-05-21 Last revision: 2025-06-26 Accepted: 2025-07-21

Heterogeneous Region and its Servicing by Distributed Queuing System

Jaroslav Janáček[1], Marek Kvet[2]

Abstract. Most of the location problems solved on a discrete transportation network use the model, in which both serviced demands and service centers are located at nodes of the transportation network. The nodes represent centers of gravity of the real sources of demands and candidates of service center locations. This simplification enables determination of time or mileage distances as distances between the corresponding points on the real map of the region. On the other side, the simplification may bring severe inaccuracy in the used model, especially in the case when the demands for service emerge randomly in the whole city district represented by the associated network node. Non-negligible size of the district area causes the real distance between demand location and the service center located at the same city to become zero distance and all demands in the city are neglected. If the city generates a considerable number of demands, the simplified objective function of the model is not reliable. This situation is extremely complicated in a heterogeneous region, where the set of dwelling places is a mixture of small and large cities generating different numbers of demands. Within this contribution, we suggest a way to comply with the above-described heterogeneity of the serviced region and publish the influence of it on the suggested service system design.

Keywords: Location problems, emergency services, heterogeneous region, distributed queuing system, optimization

JEL Classification: C44, C61
AMS Classification: 90C05, 90C06, 90C10, 90C27

1 Introduction

This research paper deals with a specific group of location problems aimed at providing service to clients by a distributed queuing system. Solutions to such problems find their application in many areas of common human life without us realizing it. In our case, we focus on emergency services, i.e. the main application field is the health sector, thus building on previous successful research by our department (Czimmermann and Kohání, 2018; Janáček and Kvet, 2022; Jankovič, 2016; Jánošíková, 2007; Kvet, 2014; Matiaško and Kvet, 2017).

When making strategic decisions about public service system design, including the aforementioned emergency services, mathematical models of various location problems are often employed as a supporting tool. These days, we can get the suboptimal solution (or even the optimal one) in a fair amount of time thanks to a number of efficient solving techniques (Arroyo et al., 2010; Avella et al., 2007; Doerner et al., 2005; García et al., 2011; Gendreau and Potvin, 2010; Revelle et al., 2008).

On the other hand, the problem may be extraordinarily challenging if the model structure deviates from the common standards. Another example of such difficulties consists in the character of the real system being optimized, and the associated transportation network may contain certain specifics that must be considered (Brotcorne et al., 2003; Ingolfsson et al., 2008; Karatas and Yakıcıa, 2019).

Most of the location issues addressed within a discrete transportation network employ a paradigm where both serviced requests and service centers are situated at the nodes of the network. The nodes signify the centers of gravity for actual demand sources and potential service center sites. This simplification facilitates the calculation of time or mileage distances as the intervals between corresponding places on the actual map of the region. Conversely, the simplification may introduce significant inaccuracies in the employed model, particularly when service demands arise randomly over the entire city district represented by the corresponding network node.

[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

The substantial size of the district area renders the actual distance between the demand location and the service center within the same city insignificant, resulting in the disregard of all needs in the city. If the city produces a significant volume of emergency calls, the model's simple objective function is unreliable. This scenario is more intricate in a diverse region, where the array of residential areas comprises both small and large cities, resulting in varying levels of demand. In this contribution, we propose a method to address the heterogeneity of the serviced region and examine its impact on the proposed service system design.

The remaining part of this study is structured as follows: The following chapter introduces the stochastic model of the public service system design problem. The third section is aimed at customization of the model to a heterogeneous region. The theoretical part of the paper is followed by a computational study in which we discuss the results of numerical experiments. Finally, the last section concludes with presented research and suggests possible directions for future development.

2 Stochastic Model of the Public Service System Design Problem

The system design problem is defined as a location problem on a network consisting of nodes and links. The nodes represent real cities, towns and villages and this set of serviced network nodes is denoted by the symbol *J*. The links correspond to existing roads, which enable the movement of servicing vehicles from one city to another. Even if each considered dwelling place $j \in J$ has a considerable district area, it is represented by only one point in the map of the serviced region. The point is usually the center of gravity of the district area. Thus, the distance between two cities *i* and *j* connected by a road is computed as the time t_{ij} necessary for vehicle to be moved from one point to another. It follows that diagonal elements of matrix $\{t_{ij}\}$ have zero values. Each node $j \in J$ in the network has its weight b_j (e.g. the number of inhabitants) and it is assumed that the number of randomly emerging demands for service at a given node is proportional to the weight.

Determination of the service system design consists of deployment of *p* servicing vehicles at possible locations, which corresponds with subset *I* of the network nodes. The deployment is given by a series of decisions $y_i \in Z^+$ for $i \in I$. The objective of the location problem is minimization of the average response time of the system to randomly emerging demands for service. The objective function must comply with the case, when the geographically nearest servicing vehicle to the demand location is occupied by some previously arisen demand.

In such a case, the time distance between the demand location and the nearest non-occupied vehicle must be taken into account. It means that the nearest vehicle is available with probability q_1. Each vehicle is considered as a one servicing facility queuing system with no buffer.

If λ is the demand rate per vehicle and μ is the service rate, the Erlang's formula $q_1 = \mu/(\mu+\lambda)$ holds. The probability that the *k*-th nearest vehicle is the first available can be estimated by the following mathematical formula $q_k = q_1(1-q_1)^{k-1}$ for $k = 1, \ldots, r$, where *r* is such integer, that q_r is almost negligible value.

Let us denote by $T_{ij}(n)$ a sequence of *n* elements, which has the same value t_{ij}. Operator $\bigcup_{i \in I} T_{ij}(n_i)$ produces a sequence of $\sum_{i \in I} n_i$ elements, which are contained in the individual $T_{ij}(n_i)$. In the union, some values repeat. $T_{ij}(0)$ is the empty sequence. Furthermore, let us introduce operator $min_k\{T\}$, which gives the *k-th* minimal value from *T*, taking into account repeating values. Then, the objective function of the system design given by decisions y_i, $i \in I$ can be described by (1).

$$f(\mathbf{y}) = \sum_{j \in J} b_j \sum_{k=1}^{r} q_k \min_k \left\{ \bigcup_{i \in I} T_{ij}(y_i) \right\} \tag{1}$$

The service system design problem can be stated as follows.

$$Minimize \quad f(\mathbf{y}) \tag{2}$$

$$Subject\ to \quad \sum_{i \in I} y_i = p \tag{3}$$

$$y_i \in Z^+ \quad for\ i \in I \tag{4}$$

3 Customization of the Model to a Heterogeneous Region

The above model may be used for the system design on the networks, where populations are deployed regularly in the dwelling places represented by the network nodes. As concerns a real region, this assumption is not fulfilled

due to the big variety of dwelling place sizes. Almost each real region consists of large and densely populated centers surrounded by smaller satellites and by numerous villages and hamlets. Usage of the above model with zero values of diagonal entries of time matrix has an important disadvantage. Location of one or more vehicles at a node i causes that possibly big weight b_i is neglected due to zero value of t_{ii}. This natural tendency of minimization process considerably deforms the objective function. That is why the diagonal elements of the matrix of time distances must be adjusted to reflect the average time necessary for moving a vehicle from a center of the city to a demand located in a district area of the city.

Because of the irregular shape of the district area, determination of the average distance represents a serious problem, which can be solved simulation based on Monte Carlo approach applied on each individual dwelling place.

Assuming that the average distance t_{ii} is known for arbitrary $i \in I$, another complication emerges. As the rate of demands is proportional to the node weight, it can be supposed that a vehicle should be located at a node i with bigger weight rather than at a node j with smaller weight. Nevertheless, an irregularity of the district shape may cause that t_{ii} and t_{ij} are almost the same, but b_i is much bigger than b_j and also t_{ii} is much bigger than t_{jj}. In this case, locating a vehicle at node i brings the contribution to the objective function (1) given by formula $q_1 b_i t_{ii} + b_j t_{ij}$ and the alternative location of the vehicle at the node j contributes by $b_i t_{ij} + q_1 b_j t_{jj}$. Considering that $t_{jj} b_j / b_i$ is almost negligible, it can be derived that if $q_1 t_{ii} > (1 - b_j/b_i) t_{ij}$ then the vehicle location at i may bring bigger contribution to the objective function value than vehicle location at j and it follows that the minimization process locates the vehicle at the satellite j.

The third drawback of the above model is that it does not reflect enormous concentration of demand for service in large cities comparing the rate of demands in the surrounding dwelling places. To mitigate influence of the heterogeneity of the modelled region service, we computed the rate of demand per one vehicle as $const \sum_{j \in J} b_j / p$ and added the following set of constraints to the model.

$$y_i \geq f_i = \left\lfloor \frac{b_i p}{\sum_{j \in J} b_j} \right\rfloor \quad for\ i \in I \tag{5}$$

In addition, we fixed the decision variable y_j at zero if the node j is located in radius t_{ii} from a node i, for which $f_i \geq 1$. This way, $\sum_{i \in I} f_i$ is number of vehicles, which locations are fixed in advance and only $p - \sum_{i \in I} f_i$ will be located based on minimization of (1).

The way of rounding (5) yields the lower limit on the numbers of vehicles, which should be assigned to individual nodes to keep balance among the associated large cities and the rest of the network. Nevertheless, the formula (5) does not take into account the nodes, for which the rounded ratio is tightly below an integer value. These cases can be handled by application of the rule (6) with a non-negative parameter α.

$$f_i(\alpha) = \left\lfloor \frac{(1+\alpha) b_i p}{\sum_{j \in J} b_j} \right\rfloor \quad for\ i \in I \tag{6}$$

The rounding rule (6) unjustifiably pulls up the bigger values of the rounded ratios unjustly proportionally to their absolute values. That is why we suggested the following two rules. The rule (7) considers only reminder of division, which is increased depending on α.

$$f_i(\alpha) = \left\lfloor \frac{(1+\alpha) b_i p}{\sum_{j \in J} b_j} - \alpha \left\lfloor \frac{b_i p}{\sum_{j \in J} b_j} \right\rfloor \right\rfloor \quad for\ i \in I \tag{7}$$

The rule (8) increases the rounded ratio only by increment α.

$$f_i(\alpha) = \left\lfloor \alpha + \frac{b_i p}{\sum_{j \in J} b_j} \right\rfloor \quad for\ i \in I \tag{8}$$

Each of the rounding rules is accompanied by a given upper limit of fixed number f_i vehicles assigned to one node of the transportation network.

To keep the number of non-fixed vehicles above a limit given by a percentage β of p, the value of α can be obtained by maximization of α subject to (9).

$$p - \sum_{i \in I} f_i(\alpha) \geq p\beta \tag{9}$$

This paper aims to find the relation between values of parameter β and objective function (1) of the optimal solution of the problem (2)-(5).

4 Numerical Experiments

The numerical experiments with the customized model of emergency medical system servicing higher territorial unit Žilina of the Slovak Republic were performed using two algorithms. The first one performs the adjustments described in Section 3 for given value of parameter β and produces input data for the second algorithm. The second algorithm solves the problem (2)-(5) for adjusted data. The second algorithm uses the optimization software FICO Xpress to yield 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). The specific model has been obtained by adjusting the model (2)-(5) to the radial form with multiple facility locating.

The transportation network of the higher territorial unit Žilina included 315 nodes, which correspond to various dwelling places from large cities to smaller towns and numerous villages and hamlets. The dwelling places produce demand for service at 0.036517 per minute per one inhabitant. The population of the territorial unit consists of 687174 inhabitants. It is considered 36 vehicles to service the territorial unit from the network nodes, location of which should be determined by the second algorithm. Distribution of the population is described in Table 1.

The table describes nine classes including dummy class number 2, which demonstrates the gap between the most populated city and the rest of the dwelling places. The row "Cardinality of the class" contains the number of dwelling places, which belong to the class. The rows "Upper bound of the class" and "Lower bound of the class" specify the class by maximal and minimal size of dwelling place population of the dwelling place included into the class. To demonstrate the maximal population of the dwelling place of the class both population of maximal and the second maximal dwelling place populations are reported in the rows denoted by "The maximal population of node" and "The second maximal population of node". Furthermore, the ratios of dwelling place population and the number of inhabitants per one vehicle were computed for the maximal and the second maximal populated dwelling place of the class in the rows "Ratio of the maximal populated node" and "Ratio of the second maximal populated node".

Class of nodes	9	8	7	6	5	4	3	2	1
Cardinality of the class	299	8	2	1	1	2	1	0	1
Upper bound of the class	6000	11000	16000	21000	26000	31000	51000	76000	81000
Lower bound of the class	0	6000	11000	16000	21000	26000	31000	51000	76000
The maximal population of node	5777	8760	14380	17600	22580	30040	50629	-	80632
The second maximal population of node	5755	7465	11380	-	-	26548	-	-	-
Ratio of the maximal populated node	0.3	0.5	0.8	0.9	1.2	1.6	2.7	-	4.2
Ratio of the second maximal populated node	0.3	0.4	0.6	-	-	1.4	-	-	-

Table 1 Distribution of population in nodes of the higher territorial unit Žilina

The probabilities q_k used in (1) were obtained in the following way. The probability $q_1 = 0.798474$ was computed from real EMS data. The value of the parameter μ has been obtained as 0.0040190396. The values of the associated parameters q_k for $k = 2, \ldots, 5$ follows $q_2 = 0.160913$, $q_3 = 0.0324282$, $q_4 = 0.00653513$ and $q_5 = 0.00164939$.

The following experiments were suggested to find out the dependence of the average response time of the optimized system on parameters used for model customization. The only free parameter in the customized model is parameter β, which represents percentage of the p disposable vehicles assigned to the network nodes representing the most populated cities. The minimal number of assigned vehicles is 9, which corresponds to the definition of f_i by formula (5). This number of vehicles correspond with the value of a quarter of the parameter β. The experiments with each of the three rounding rules were performed for $\beta = 0.25, 0.3, 0.35, \ldots, 0.95$ and 1. The rules are denoted

as *Rule*6, *Rule*7 and *Rule*8 according to the formulae (6), (7) and (8) respectively. The associated results are plotted in Table 2, where the column denoted by *Sumf* contains the number of assigned (or fixed) vehicles and column denoted by *RT* contains the optimized average response time in minutes.

β	Rule 6		Rule 7		Rule 8	
	Sumf	*RT*	*Sumf*	*RT*	*Sumf*	*RT*
0.25	9	6.445	9	6.445	9	6.445
0.3	10	6.445	10	6.445	10	6.445
0.35	12	6.533	12	6.545	12	6.531
0.4	14	6.605	14	6.556	14	6.556
0.45	16	6.680	16	6.579	15	6.556
0.5	18	6.811	18	6.669	18	6.624
0.55	19	6.893	19	6.647	18	6.624
0.6	21	6.920	21	6.725	21	6.660
0.65	23	7.104	23	6.730	23	6.657
0.7	25	7.141	25	6.830	25	6.697
0.75	27	7.234	27	6.914	27	6.698
0.8	28	7.235	27	6.914	27	6.698
0.85	30	7.361	30	6.981	30	6.743
0.9	32	7.371	32	6.984	32	6.857
0.95	33	7.525	34	7.072	34	7.011
1	36	7.903	36	7.261	34	7.011

Table 2 Dependence of optimized response time on number of pre-assigned vehicles

Comparison of the optimized average response times shows that the different rules have approximately the same influence on quality of the system design, if the percentage β of directly assigned vehicles varies from 0.25 to 0.4. In each case the average time slightly increases, but the increase does not exceed 2% of the best value. For the higher percentage, rule 8 outperforms rules 6 and 7. Rounding rule 8 enables us to reach better deployment of the vehicle locations.

The question is, how does the way of customization (rounding) affect the multiple assignment of the vehicles to individual highly populated cities. That is why we use the current design of the present emergency medical system as a benchmark. The average response time of the system computed according to (1) is 6.975. It is about 8 percent worse than the results obtained for $\beta = 0.25$ or $\beta = 0.3$. In the next paragraph, we compare the resulting designs according to the multiple assignments. The current system performs multiple assignments in five cities plotted in Table 3.

City	Population	Number of assigned vehicles				
		Currently	β= 0.25 – 0.3	β=0.3	β=0.4 – 0.45	β=0.5 – 0.95
Čadca	2258	2	1	1	1	1
Žilina	80634	4	4	4	4	4
Martin	50624	2	2	3	3	3
Ružomberok	26548	2	1	1	1	2
Lipt. Mikuláš	30040	2	1	1	2	2

Table 3 Analysis of multiple vehicle assignments

The multiple assignments obtained by solving the model customized by rule 8 are plotted in the columns denoted by the ranges of β. The multiple assignment of the current system design deploys seven vehicles so that they are added to one another vehicle located in all of the plotted cities. This number of multiply assigned vehicles is achieved for optimized instances for $\beta \geq 0.5$. Nevertheless, the average response time of the suggested system design is 6.624. Considering the same number of multiply assigned vehicles, this result is 5.4% better than the average response time of the current system. The improvement of the suggested design is achieved by better deployment of the vehicles, which are located individually at the network nodes.

5 Conclusions

The presented approach proved that the customization of the model of stochastically heterogeneous region may contribute to the more realistic ambulance deployment. The suggested approach can assign a given percentage of disposable ambulances to the more populated communities and optimize the average response time of the system by determining location of the rest of the ambulance fleet. Comparison of the suggested scenarios of ambulance deployment to the current ambulance deployment shows that the average response time of the suggested scenarios considerably outperforms the response time of the current ambulance deployment submit to the condition that the numbers of multiply assigned ambulances are the same.

Acknowledgements

This work was supported by the This work was supported by the research grant VEGA 1/0220/25 “Regional equality in access to public rescue services provided from service centers on the transport network”.

References

Arroyo J E C, dos Santos P M, Soares M S, Santos A G (2010) A Multi-Objective Genetic Algorithm with Path Relinking for the p-Median Problem. In: The 12th Ibero-American Conference on Advances in Artificial Intelligence 2010. Conference Proceedings. pp. 70–79. https://doi.org/10.1007/978-3-642-16952-6_8

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

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

Gendreau M, Potvin J (2010) Handbook of Metaheuristics, Springer Science & Business Media.

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, Kvet M (2022) Repeated refinement approach to Bi-objective p-location problems. In: INES 2022. Conference Proceedings, New York: Institute of Electrical and Electronics Engineers, pp. 41-45. https://doi.org/10.1109/INES56734.2022.9922620

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.

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.

Matiaško K, Kvet M (2017) Medical data management. Informatics 2017. Conference Proceedings, pp. 253-257.

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

Received: 2025-05-29 Last revision: 2025-07-15 Accepted: 2025-07-22

Accuracy and Complexity of the Radial Approach to Ambulance Location in Emergency Medical System

Marek Kvet[1], Jaroslav Janáček[2]

Abstract. The radial approach to large location problems considerably improves the solvability of the models, where a given number of facilities should be deployed in transportation network nodes to minimize average distance between users and their closest facilities. A smart solving algorithm of a location problem is especially important in situations, when an emergency medical system is designed to service a large region with hundreds of nodes of an associated transportation network. An excellent performance of the optimization method based on the radial approach is paid for by worse accuracy. The inaccuracy of the method is caused by replacing a real distance with a stepwise function of distance. The number and size of the steps conflictingly influence the result accuracy and size of the model. This contribution is devoted to quantifying the above relations with the goal to find an acceptable balance between computational time necessary for the problem solution and accuracy of the solution.

Keywords: Location problems, urgent healthcare management, generalized disutility, multiple facility location, radial approach, accuracy sensitivity

JEL Classification: C44, C61
AMS Classification: 90C05, 90C06, 90C10, 90C27

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. It represents an important and irreplaceable part of the provision of urgent care, which belongs to the highest level of differentiated patient care - resuscitation and intensive care. This service is typically offered by private EMS companies, agencies, contingent upon their licensing or authorization to operate a medical service center. This system is a component of the integrated rescue system, which is part of the broader group of public service systems (Brotcorne et al., 2003; Ingolfsson et al., 2008).

Generally, a public service system is characterized by a collection of served entities, namely customers with distinct demands, and by service providers, which can include both private and public organizations, such as corporations and governmental bodies. The clientele of a specific system is predominantly situated within a particular geographical region, while the resources that fulfill their requirements are often centralized in one or more locations, which may be found both inside and beyond the served area (Jánošíková, 2007). All efforts to optimize the rescue system are united by the requirement to achieve the highest possible availability and accessibility of service for all clients subject to all restrictive conditions. Obviously, the closer the nearest available center from which the ambulance vehicle is sent to satisfy the patient's demand is, the better accessibility is achieved (Doerner et al., 2005). No doubt that in emergency situations, the distance or response time plays a crucial role. Searching for the optimal deployment of service centers (finding their best possible locations) is a complex combinatorial problem with an almost innumerable number of feasible solutions. Such problems cannot be solved unless sophisticated software support is available. Thus, a mathematical model is usually formulated to master the combinatorial character of the associated decision-making process. In other words, the associated decision-making problem can be understood as a task to find the best locations of service centers to optimize given objective function, i.e. to achieve the best possible service accessibility for clients. Speaking of emergency services, the accessibility of the service is usually quantified by the average travel time or geographical network distance of the closest ambulance to the emergency location (to the patient), This way, the associated mathematical model of the problem may take the form of well-known weighted *p*-median problem successfully studied by many authors from the viewpoint of efficient exact or approximate solving techniques. But it must be noted that there are plenty of other objectives that may be optimized instead of the mentioned one (Jánošíková et al., 2021; Reuter-Oppermann et al., 2017).

[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

If the mathematical model of the problem is described in the location-allocation form, then the associated solving algorithm based on the branch and bound method usually fails mainly with large instances due to enormous computational requirements and memory demands caused by the necessity of introducing a matrix of binary variables to model the assignment of clients to service centers. To overcome this inconvenience, the radial formulation of the problem can replace the original model (Avella et al., 2007; García et al., 2011; Janáček, 2008). It simplifies solving large instances of many integer-programming problems, because it does not need to model the above-mentioned assignment. Furthermore, it can be easily adjusted to accelerate the computation even more (Janáček and Kvet, 2016; Kvet, 2015). Even if it was originally developed as an approximate approach, it enables us to obtain the exact optimal solution in a short time.

Common mathematical models based on the weighted *p*-median problem may suffer from two weak points that represent a high level of simplification. The first inconvenience consists in stochastic behavior of real systems. The demands for intervention occur randomly at any time and the closest EMS station may be temporarily unavailable because of satisfying previously raised requests. In such a case the current emergency call is covered by the second nearest station or by the one which is currently free and near enough. The mentioned modelling strategy makes the objective function more realistic and enables us to model the stochasticity of real systems at least partially. This idea has been introduced as the concept of so-called *generalized disutility* (Jankovič, 2016; Kvet, 2014).

The second weakness follows the heterogeneity of the served region. The assumption that no more than one service center can be located in any network node is too strict. Looking at real EMS systems, there are several cities in which more stations are operated. This requirement can be fulfilled either by a decomposition of the original network that means one network node is replaced by a small subnetwork (obviously the original weight of the node needs to be somehow distributed among all new nodes) or by permission to locate multiple facilities in one location. If there are some higher populated places, one may request locating a minimal number of ambulance vehicles there to cover all expected demands.

In this contribution, we study the accuracy, sensitivity and complexity of the radial approach to ambulance locating in the EMS system. The theoretical part of the paper contains the mathematical model following the concept of generalized disutility with the possibility of locating more than one facility at one network node. The second half of this study is aimed at presenting and discussing the results of performed numerical experiments making use of the real-world data from the self-governing region of Žilina in Slovakia.

2 Mathematical Formulation of the Problem

This section focuses on the theoretical background of EMS system designing. The attention is paid to mathematical model formulation from the basic expression up to the radial form of the most complex problem. Let us begin with the simple weighted *p*-median problem.

The basic mathematical formulation assumes the existence of a finite set I of candidates for EMS station establishment. The candidate locations are usually chosen from network nodes satisfying specific requirements or they are based on preferences of responsible authorities. The resulting selection of EMS stations must contain exactly p elements from I, where p is a positive integer for which the inequality $p \leq |I|$ holds. Generally, the EMS stations are located to satisfy the demands of served population spread over the elements of a finite set J. Obviously, the sets I and J may equal. A binary variable, $y_i \in \{0, 1\}$, which takes the value of one if a station should be located at $i \in I$ and which equals zero otherwise, will be used to simulate the choice of where to place an EMS station. This way, the vector $\mathbf{y}$ of the binary variables y_i can be used to define any solution to the corresponding problem minimizing given quality criterion $f(\mathbf{y})$. Then, the following expression (1) describes the basic *p*-location model.

$$\min\left\{ f(\mathbf{y}) : y_i \in \{0, 1\}, i \in I, \sum_{i \in I} y_i = p \right\} \tag{1}$$

As far as the concrete form of $f(\mathbf{y})$ in EMS system designing is concerned, it is necessary to add some new symbols for the model input data. Let the nonnegative coefficient b_j denote the expected number of demands for service at $j \in J$. You can interpret it as a weight of the location j, though. Furthermore, let us introduce a nonnegative coefficient d_{ij} for each $i \in I$ and $j \in J$ to denote the distance between the network nodes i and j. The distance is mostly expressed in kilometers, but it can be interpreted also as the time (usually in minutes) which it takes to reach the user location j from the EMS station located at i. It follows that the values of d_{ij} do not have to be integer.

The most used form of $f(\mathbf{y})$ in EMS system designing minimizes the average distance from the service recipients to their closest service providing facility. According to the notations introduced above, the following expression (2) may formally describe this objective function.

$$f(\mathbf{y}) = \frac{\sum_{j \in J} b_j \min\left\{d_{ij} : i \in I,\ y_i = 1\right\}}{\sum_{j \in J} b_j} \tag{2}$$

For easier implementation, it is enough to optimize only the numerator of (2), because the constant in the denominator does not play any role in the optimization process.

Even though the objective function $f(\mathbf{y})$ minimizes the average distance between the demand points and the service providers, its form is not suitable for practical EMS stations deployment, because real emergency systems work more complexly. It follows that maximally r (r is a positive integer) nearest stations can participate in satisfying the emergency call. Let q_k express the probability that the k-th closest EMS station is the nearest available one at the time the demand occurs. To complete the mathematical formulation of the adjusted objective function, let the operator min_k return the k-th smallest element from the list in the operator's parameter. Based on these preliminaries, the former criterion described by (2) can be reformulated into the more general form of (3).

$$f(\mathbf{y}) = \sum_{j \in J} b_j \sum_{k=1}^{r} q_k \min_k \left\{d_{ij} : i \in I,\ y_i = 1\right\} \tag{3}$$

The radial approach is built on the idea of excluding the assignment of served nodes to established EMS stations, and it is based on a specific expression of the distances making use of so-called *radii* or *zones*. To formulate the radial model, let us introduce all new necessary notations and make some little changes in location variables. If we want to admit locating more centers at the same location, the definition scope of y_i for $i \in I$ must change from $\{0, 1\}$ to Z^+. Furthermore, let the symbol m denote the highest relevant distance. If the zones are of the same length equal to 1, then $m = \max\{d_{ij}: i \in I, j \in J\}$. If the zones are smaller, then the value m adequately grows to cover all relevant distances. The radial model uses auxiliary zero-one variables x_{jvk} which are introduced for $j \in J$, $v = 0, ..., m\text{-}1$ and $k = 1, \ldots, r$ to approximate the distance from the user located at j to the k-th nearest service center. The variable x_{jvk} takes the value of 1 if the distance is greater than ve_v and it takes the value of 0 otherwise. The symbol e_v denotes the length of zone v. Unlike the location-allocation approach, in this case it is not necessary to model the assignment of requests to serving centers, which makes the model easier to solve and smaller in scope. For simple implementation we assume that the zones are organized uniformly, i.e. they are of the same size. Then the sum $x_{j0k} + x_{j1k} + \ldots + x_{jm\text{-}1k}$ corresponds to the smallest distance from the user j to the k-th nearest located center. To complete the model, we introduce coefficients a_{ijv} for each $i \in I$, $j \in J$ and $v = 0, \ldots, m\text{-}1$. Their values are computed according to (4).

$$a_{ijv} = \begin{cases} 1 & \text{if } d_{ij} \le ve_v \\ 0 & \text{otherwise} \end{cases} \quad for\ i \in I, j \in J, v = 0, 1, \ldots m-1 \tag{4}$$

The radial model of the problem can be formulated as follows.

$$Minimize \quad \sum_{j \in J} b_j \sum_{k=1}^{r} q_k \sum_{v=0}^{m-1} e_v x_{jvk} \tag{5}$$

$$Subject\ to \quad \sum_{i \in I} y_i = p \tag{6}$$

$$\sum_{k=1}^{r} x_{jvk} + \sum_{i \in I} a_{ijv} y_i \ge r \quad for\ j \in J,\ v = 0, ...,\ m-1 \tag{7}$$

$$y_i \in Z^+ \quad for\ i \in I \tag{8}$$

$$x_{jvk} \in \{0, 1\} \quad for\ j \in J,\ v = 0,\ ...,\ m-1,\ k = 1, ...,\ r \tag{9}$$

The objective function (5) expresses the expected value of traversing distances from users to the nearest available center. Constraint (6) ensures that the number of located facilities equals p. Constraints (7) link-up the location variables y_i to the variables x_{jvk} so that the sum of x_{jvk} is greater than or equal to the value of complement of the number of located service centers in the radius v from the customer located at j to the number r.

The last thing to discuss consists in managing the heterogeneity of served region. As the radial model (5)-(9) allows to locate multiple facilities at one place, a minimal number of ambulance vehicles can be required for each candidate location. This extra condition is usually applied only to larger communities, which cannot be covered

by one station only. The number of fixed vehicles that must be located can be defined in many ways. Within this study we use the following expression (10) with the parameter increment α.

$$y_i \geq fixed_i(\alpha) = \left\lfloor \alpha + \frac{b_i p}{\sum_{j \in J} b_j} \right\rfloor \quad for\ i \in I \tag{10}$$

To keep the number of non-fixed vehicles above a limit given by a percentage β of p, the value of α can be obtained by maximization of α subject to (11).

$$p - \sum_{i \in I} fixed_i(\alpha) \geq p\beta \tag{11}$$

In this contribution, we study the accuracy sensitivity and complexity of the radial approach to ambulance locating in the EMS system. Thus, we deal with the complete radial model as it was presented in this section.

3 Numerical Experiments

This chapter reports on the practical part of the research. We summarize here the results of a small computational study aimed at accuracy and complexity analysis of the presented radial model with multiple facility locating and minimal number of fixed vehicles in the candidate locations. The numerical experiments were performed within the optimization software FICO Xpress 8.3 (64-bit, release 2017). The associated models were solved on a common PC equipped with the 1.80 GHz Intel® Core™ i7 8565U processor and 128 GB RAM. As a studied benchmark, we took the data from real EMS system operated in higher territorial unit of Žilina. The Žilina region is located in the northwestern part of Slovakia. It includes five subareas (Horné Považie, Kysuce, Liptov, Orava, and Turiec), which are divided into eleven districts. In an area of 6809 km^2 (14% of the Slovak territory), there are 315 network nodes representing cities, villages and other small communities. The dwelling places produce demand for service at 0.036517 per minute per one inhabitant. The population of the territorial unit consists of 687174 inhabitants. The EMS system operates 36 fully equipped vehicles to cover the demands for urgent healthcare in emergencies. Currently, the system performs in such a way, that the numerator of (2) takes the value 4793069 and the average response time – the value of the expression (2) equals 6.975 minutes. These values are referential, to which the results of our optimization procedures can be compared to. An individual experiment was divided into two parts. The first phase was focused on identifying the nodes, in which some vehicles must be fixed. The expressions (10) and (11) were applied for the given value of parameter β. Then, the presented radial model was solved to obtain the resulting EMS vehicles deployment. Since our goal was to study the accuracy sensitivity and complexity of the radial approach, we performed 10 runs of the radial model. These runs differed in the zones construction. For simplicity, the zones were distributed uniformly, and their sizes were the same. The length of zones was 0.1, 0.2, …, 0.9 and 1 minutes respectively. Obviously, the zone length affected the number of used intervals for distance approximation, the number of decision variables and the computational demands as well. The most interesting outputs for further analyses are the zone length, parameter m and the computational time in seconds. Since the radial model uses distance approximation, we report both: the resulting value of average response time computed from the expression (5), but also the corresponding real value of (2) evaluated for the resulting vector **y**. The more accurate the distance approximation, the smaller the difference between these two values of average response time. On the other hand, good distance approximation requires many zones and a larger number of decision variables, which means higher memory and computational demands and higher computational time, of course. The summary of obtained results is reported in the following tables.

	Zone length									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
m	410	205	137	103	82	69	59	52	46	41
Computational time [s]	8.050	3.880	2.870	4.430	1.890	1.850	5.190	1.650	1.290	1.230
Avg response time – approx.	6.435	6.490	6.547	6.578	6.637	6.664	6.732	6.781	6.831	6.864
Avg response time – real	6.435	6.436	6.438	6.436	6.436	6.438	6.436	6.440	6.436	6.445

Table 1 Results of numerical experiments for $\beta = 0.25$

	Zone length									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
m	410	205	137	103	82	69	59	52	46	41
Computational time [s]	9.890	4.490	3.390	5.610	2.090	1.840	6.110	1.460	1.220	1.180
Avg response time – approx.	6.435	6.490	6.547	6.578	6.637	6.664	6.732	6.781	6.831	6.864
Avg response time – real	6.435	6.436	6.438	6.436	6.436	6.438	6.436	6.440	6.436	6.445

Table 2 Results of numerical experiments for $\beta = 0.3$

	Zone length									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
m	410	205	137	103	82	69	59	52	46	41
Computational time [s]	8.160	4.540	3.030	3.940	1.870	1.680	4.770	3.360	1.330	0.940
Avg response time – approx.	6.529	6.586	6.642	6.675	6.732	6.758	6.834	6.874	6.926	6.96
Avg response time – real	6.529	6.529	6.533	6.529	6.529	6.533	6.531	6.536	6.532	6.531

Table 3 Results of numerical experiments for $\beta = 0.35$

	Zone length									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
m	410	205	137	103	82	69	59	52	46	41
Computational time [s]	8.590	3.730	2.860	4.780	2.070	1.770	3.650	3.620	1.080	1.120
Avg response time – approx.	6.554	6.611	6.667	6.701	6.757	6.782	6.859	6.897	6.954	6.984
Avg response time – real	6.554	6.554	6.558	6.554	6.554	6.558	6.556	6.561	6.557	6.556

Table 4 Results of numerical experiments for $\beta = 0.4$

	Zone length									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
m	410	205	137	103	82	69	59	52	46	41
Computational time [s]	8.250	4.290	2.930	2.650	2.010	1.740	3.670	3.630	1.210	1.270
Avg response time – approx.	6.554	6.611	6.667	6.701	6.757	6.782	6.86	6.897	6.954	6.984
Avg response time – real	6.554	6.554	6.558	6.554	6.554	6.558	6.565	6.561	6.557	6.556

Table 5 Results of numerical experiments for $\beta = 0.45$

	Zone length									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
m	410	205	137	103	82	69	59	52	46	41
Computational time [s]	7.110	3.330	2.320	1.760	1.490	1.340	1.020	0.960	0.910	0.830
Avg response time – approx.	6.617	6.675	6.72	6.758	6.798	6.836	6.929	6.96	7.011	7.029
Avg response time – real	6.617	6.617	6.62	6.617	6.618	6.625	6.617	6.619	6.623	6.624

Table 6 Results of numerical experiments for $\beta = 0.5$

	Zone length									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
m	410	205	137	103	82	69	59	52	46	41
Computational time [s]	6.740	3.080	2.050	1.660	1.350	1.540	1.730	1.380	1.310	1.240
Avg response time – approx.	6.698	6.758	6.8	6.848	6.879	6.922	7.013	7.049	7.105	7.108
Avg response time – real	6.698	6.698	6.698	6.698	6.698	6.698	6.698	6.698	6.699	6.698

Table 7 Results of numerical experiments for $\beta = 0.75$

4 Conclusions

This study was aimed at specific application of location science in the medical field. The core idea was to develop a mathematical model for EMS station locating in such a way that the model follows the concept of generalized disutility and allows locating multiple facilities at one network node. To avoid introducing assignment variables as in the location-allocation approach, we used so-called radial formulation in our method. The main scientific idea was to study the accuracy and complexity of the radial approach to ambulance location in emergency service systems. Performed numerical experiments have shown that the radial approach represents a suitable modelling technique as it performs fast with acceptable result accuracy. Even if the result accuracy grows with increasing number of zones, the computational time keeps being small.

Acknowledgements

This work was supported by the This work was supported by the research grant VEGA 1/0220/25 “Regional equality in access to public rescue services provided from service centers on the transport network”.

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

Received: 2025-06-08 Last revision: 2025-07-12 Accepted: 2025-08-13

Estimation of Price Elasticity of Demand for Energy in the Czech Republic

Natália Várošová[1], Martin Baťka[2], David Hampel[3]

Abstract. This paper focuses on estimating the price elasticity of energy demand in the Czech Republic, taking into account the significant economic and political changes of recent years. The analysis utilises a large dataset comprising monthly data on electricity consumption, energy prices, meteorological conditions, and socio-economic indicators. The main methodological approach is the application of ordinary least squares (OLS) regression models.

Keywords: elasticity, Czech Republic, electricity, energy consumption, OLS

JEL Classification: C51
AMS Classification: 62P20

1 Introduction

Elasticity measures the responsiveness of one variable to changes in another. In this paper, we focus on how energy consumption levels react to variations in price. Assuming that the quantity of energy consumed (Q) depends on the price (P), we can express the price elasticity of demand as

$$E = \frac{\mathrm{d}Q}{\mathrm{d}P} \frac{P}{Q(P)}. \qquad (1)$$

A relationship often resembles this expression that employs percentage changes in both quantity and price, as illustrated by the following formula:

$$E = \frac{\Delta Q\%}{\Delta P\%}. \qquad (2)$$

Understanding the price elasticity of demand for energy and its estimation has become increasingly crucial, especially in light of the recent rise in energy prices. Since energy costs play a vital role in determining the prices of goods and services across the economy, it is essential to examine this issue in depth. Although several previous studies have attempted to estimate energy price elasticities, a gap remains in the literature regarding the analysis of price fluctuations in the Czech Republic over recent years. In this analysis, a log-log regression model is used, where both electricity consumption and price are expressed in natural logarithms. In this specification, the coefficient on the log of price directly represents the price elasticity of demand. In conducting our research, we have drawn on insights from earlier work in this area, adopting both variables and methodological approaches that have proven helpful in past analyses.

In an early study, Houthakker (1951) recognised that relying solely on economic variables to explain fluctuations in energy consumption, particularly concerning seasonal patterns, has some limitations. Instead, this study incorporated non-economic factors such as temperature and daylight duration, which offered a more accurate explanation of these variations. In our study, we also draw on more recent research, particularly the work Burke and Csereklyei (2016), which employed a panel data model to analyse energy consumption. Their model incorporated variables such as GDP, precipitation, average temperature, land area and population. Their findings revealed that energy-GDP elasticity varies significantly across sectors: residential energy use demonstrated relatively low elasticity in relation to GDP, whereas the industrial, service, and transport industries showed much higher responsiveness. Additionally, the study found that GDP elasticity was small for solid biofuels but comparatively higher for electricity. Importantly, they observed that energy-GDP elasticity tends to increase with higher income levels. A different methodological approach was employed in a study by Sahari (2019) conducted in Finland. The sensitivity of the price was assessed in the study by measuring the elasticity of the total installed capacity of electric heating systems with respect to the electricity distribution price. It was estimated that the elasticity of demand for electric heating technologies was −0.71. Based on this finding, the authors suggested that rising energy prices, such as those driven by increased taxation, could encourage a shift in consumer preferences toward heating technologies

[1] Mendel University in Brno, Department of Statistics and Operation Analysis, Zemědělská 1, 61300 Brno, xvarosov@mendelu.cz
[2] Mendel University in Brno, Department of Statistics and Operation Analysis, Zemědělská 1, 61300 Brno, martin.batka@mendelu.cz
[3]Mendel University in Brno, Department of Statistics and Operation Analysis, Zemědělská 1, 61300 Brno, david.hampel.uso@mendelu.cz

that rely less on electricity purchased from the grid, thereby promoting more energy-efficient or alternative solutions.

Havranek et. al. (2012) identified a tendency toward publication bias in research focused on gasoline price elasticity. Once this bias was adjusted for, their analysis indicated that the typical short-term elasticity of gasoline demand was around −0.09, while the long-term elasticity was closer to −0.31. Earlier studies suggest that the demand for residential energy, including electricity used for charging electric vehicles (EVs), tends to be relatively unresponsive to price changes. For instance, Zhu et al. (2018) estimate the global price elasticity of residential electricity demand to be approximately −0.45. This means that EV charging is considered to be inelastic. Kuang et. al. (2024) assessed the price elasticity of demand for public EV charging to be −0,7581, which indicates that demand in Shenzhen is relatively inelastic in response to electricity prices. Additionally, the results show that negative price changes have a minimal impact on elasticity, whereas positive price changes tend to increase price sensitivity among EV charging users.

The goal of this paper is to estimate and assess the price elasticity of electricity demand in the Czech Republic for the period 2021–2023.

2 Materials and Methods

2.1 Materials

This paper utilises a dataset comprising socio-economic, energy, and meteorological indicators for the Czech Republic. The domestic net electricity consumption, expressed in megawatt-hours (Q), functions as the dependent variable in the model. The independent variables include average realised electricity prices in CZK/MWh (price), meteorological variables, and socioeconomic variables. The data were collected from the Energy Regulatory Office, the Czech Statistical Office, the agri4cast portal, and the Czech Hydrometeorological Office. Data are gathered on a monthly basis. The study focuses on the period from 2021 to 2023 due to data availability. The time series has been divided into two periods due to significant events that occurred during this timeframe, including the retreat from COVID-19 measures and the outbreak of the war in Ukraine.

2.2 Methods

The regression model for estimating the elasticity of demand is given in the general form

$$\ln Q_e = \beta_0 + \beta_1 \ln P + \gamma \mathbf{X} + \varepsilon, \qquad (3)$$

where Q_e denotes the quantity of electricity consumed, β_0 represents a constant, P is the price of electricity, β_1 is a parameter indicating demand elasticity, $\mathbf{X}$ is a vector of control variables, γ signifies parameters associated with control variables, and ε is the random error term. In the vector $\mathbf{X}$, we included several control variables reflecting socioeconomic conditions and weather characteristics, such as precipitation, wind speed, sunshine duration, temperature, and derivatives of temperature. The weather-related variable, which was significant and incorporated into the model, was measured using daily temperature data aggregated to a monthly level to reflect deviations from standard temperature conditions relevant for electricity consumption. Two models for the years 2021–2022 and 2022–2023 were estimated using the OLS method. The Gaussian-Markov assumptions were tested for both models with the appropriate tests listed in Tables 2 and 4. The estimates were analysed for statistical significance using t-tests and F-tests. Particular attention was given to the value of the coefficient of determination (R^2) and other diagnostic tests to verify the validity of the assumptions. The models were examined within the research context to ensure they were not affected by spurious regression. All calculations were performed using Gretl 2025a software, with a significance level of 0.05.

3 Results

Based on data from 2022 to 2023, we developed a model that appears to be statistically significant and explains 94% of the variability, which can be regarded as a reasonably good model. The parameter related to the price has a negative sign, consistent with the theory that as price increases, the quantity consumed decreases. The absolute value of this parameter is relatively low, aligning with the theory that has identified low elasticity values for essential goods, such as electricity. Of the several variables included in the model, only price and the weather variable were found to be statistically significant. We anticipated a positive sign for the weather variable, and this expectation was confirmed. The statistically significant variables, along with their coefficients and p-values, are presented in Table 1.

Variable	Coefficient	p-value
const	8.4607	<0.001
lnP	-0.0288	0.001
weather	0.0004	<0.001

Table 1: OLS estimates of the model based on 2022–2023 data

Classical assumption tests	p-value
Specification LM test (squares)	0.723
Specification LM test (logs)	0.616
RESET test squares and cubes	0.732
RESET test squares only	0.532
RESET test cubes only	0.531
White´s test	0.185
Breusch-Pagan test	0.337
Ljung-Box test	<0.001
Shapiro-Wilk test	0.737
Durbin-Watson test (positive autocorrelation)	0.996
Durbin-Watson test (negative autocorrelation)	0.004

Table 2: Tests of classical assumptions for the model based on 2022–2023 data

The specification tests in Table 2 indicate that the model is correctly specified. None of the tests used rejected the hypothesis that the model is correctly specified. The White's test and the Breusch-Pagan test did not reject the assumption of homoskedasticity. However, based on the Ljung-Box test, we reject the null hypothesis that the residuals are not autocorrelated. This autocorrelation suggests that there is some temporal or ordinal dependence between observations. As the model does not aim to predict but to estimate elasticity or explain relationships between variables, we consider this autocorrelation acceptable. It does not fundamentally alter the qualitative conclusions of our analysis and does not threaten the unbiasedness of the estimated coefficients.

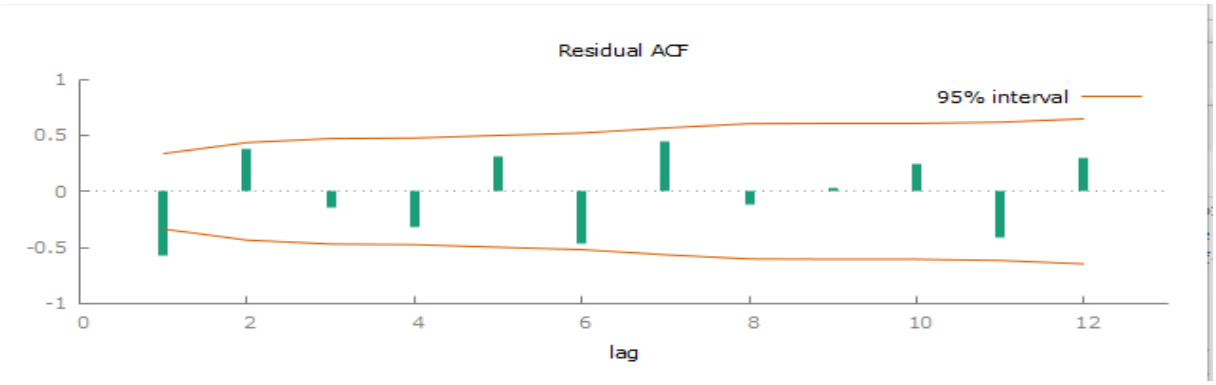


Figure 1: ACF for OLS model with 2022–2023 data

The ACF plot (Figure 1) reveals a significant first-order autocorrelation (lag 1), indicating that the current values are closely related to those in the previous period. This phenomenon indicates the presence of inertia in the data and suggests that the model could be improved by incorporating lagged variables. However, the addition of the lagged variable was not found to be statistically significant, and the autocorrelation was still present, see the ACF plot in Figure 2.



Figure 2: ACF for the OLS model with 2022–2023 data with lagged variable

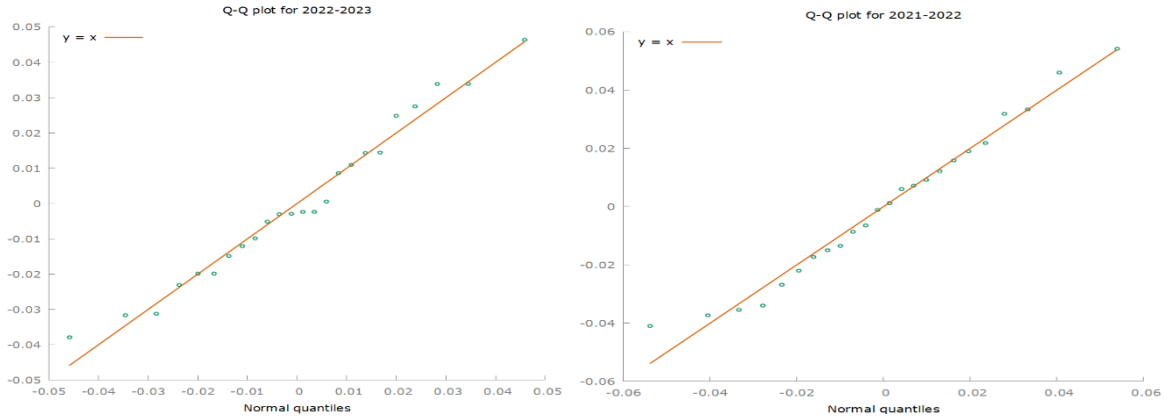


Figure 3: Q-Q plot for residuals of the model for 2022–2023 (left graph) and for 2021–2022 (right graph)

The normality of the residuals was tested using the Shapiro-Wilk test, which did not lead to rejection of the normality assumption. The normal distribution of the residuals is visually supported by the Q-Q plot shown in Figure 3. Regarding the risk of spurious regression, the stationarity of the data was tested by the KPSS test, which did not reject the null hypothesis of stationarity. Thus, spurious regression was ruled out. Overall, these results indicate that the normality assumption for the residuals is satisfied. The quality of the model is also reflected in the plot of actual and fitted values (Figure 4), which appears quite reliable. The graph clearly demonstrates how energy consumption varies seasonally, with a notable decrease in electricity consumption during the summer months. Theil's U coefficient in the model reached a value of 0.355, suggesting that the model has good predictive accuracy. The RMSE value of 0.022 further indicates that the model has a relatively low average prediction error.

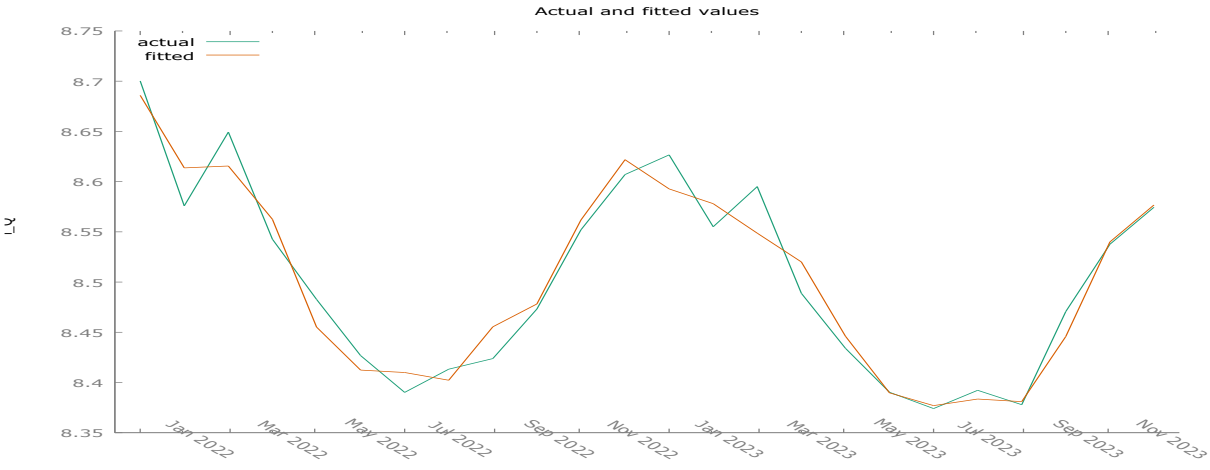


Figure 4: Graph of actual and fitted values for the model with 2022–2023 data

For comparison, the model based on data from 2021–2022, with the same variables, was estimated (Table 3) and shows very similar assumption test values (Table 4). The variables were statistically significant, with coefficient values similar to those in the previous case. The tests for this model yielded very similar results, with only a slight

difference in autocorrelation. In this case, we did not reject the null hypothesis of no autocorrelation. However, the values of RMSE (0.026) and Theil U (0.408) are slightly higher than in the previous case.

Variable	Coefficient	p-value
const	8.4603	<0.001
lnP	-0.0182	0.021
weather	0.0004	<0.001

Table 3: OLS estimates of the model based on 2021–2022 data

Classical assumption tests	p-value
Specification LM (squares)	0.079
Specification LM (logs)	0.536
RESET test squares and cubes	0.652
RESET test squares only	0.889
RESET test cubes only	0.891
White´s test	0.338
Breusch-Pagan test	0.192
Ljung-Box test	0.076
Shapiro-Wilk test	0.737
Durbin-Watson test	0.459

Table 4: Tests of classical assumptions for the model based on 2021–2022 data

4 Discussion

The results of our model confirm the low price elasticity of electricity demand in the Czech Republic, aligning with existing studies that highlight the limited sensitivity of essential energy consumption to price changes. Similarly, Atalla et. al. (2018) emphasise the significant role of meteorological factors and the low price elasticity of electricity consumption worldwide, which corresponds with our finding that seasonal and climatic influences have a significant impact on consumption patterns. A study by Bernstein et al. (2006) highlights regional differences in energy demand elasticity, noting that lower elasticity is typical for essential commodities in developed economies, such as the Czech Republic. This is consistent with our estimated elasticity of -0.0288, indicating nearly inelastic demand, which is comparable to the results in the meta-analysis of residential electricity demand (Zhu et al., 2018). In contrast to the study by Burke et. al. (2016), our results did not reveal the significance of GDP in modelling electricity consumption.

The model was tested for classical assumptions, including model specification, normality, homoskedasticity, and the absence of autocorrelation in residuals. The specification tests showed no significant violations of the basic linear regression assumptions, confirming the proper model structure. However, an important exception is the first-order autocorrelation identified, which was confirmed by the Ljung-Box test and supported by the autocorrelation function. This autocorrelation suggests the inertial nature of electricity consumption, where current values are heavily influenced by previous periods. Unfortunately, adding the lagged variable proved statistically insignificant and did not resolve the issue of autocorrelation. Although this breaches the independence assumption of the residuals, in this exploratory analysis, which is non-predictive, this deviation is acceptable. It does not notably affect the study's qualitative conclusions regarding the estimated elasticity of demand. For creating a complex, predictive model, the Box-Jenkins ARMA model with external inputs may be used; however, this method is not appropriate for estimating the price elasticity of electricity demand.

5 Conclusion

This study aimed to estimate the price elasticity of electricity demand in the Czech Republic, considering meteorological and socio-economic factors that influence electricity consumption. The model was created to determine the elasticities of electricity demand for the entire national economy, using electricity consumption data for the whole Czech Republic. The results confirmed that electricity demand is highly inelastic, with price changes having only a moderate effect on consumption—characteristic of essential goods. The importance of weather as an additional factor was also affirmed: as the temperature-related variable indicating heating needs increases, so does electricity consumption. Although autocorrelation of residuals was observed, the model is still regarded as reliable for explaining the relationships between variables. The findings also support previous empirical studies and highlight the need for further research into the structure of electricity consumption, such as by sectoral disaggregation or by including long-term trends and technological advancements. To improve estimates of demand elasticity, especially for specific groups like businesses and households, high-quality microdata are necessary; however, such data are not readily available at the required level of detail.

Acknowledgements

This work is a part of the TS01010108 project supported by the Technology Agency of the Czech Republic within the Theta 2 Programme.

References

Atalla T, Bigerna S, Bollino CA (2018) Energy demand elasticities and weather worldwide. Econ Polit 35:207–237. https://doi.org/10.1007/s40888-017-0074-2

Bernstein MA, Griffin J (2006) Regional Differences in the Price-Elasticity of Demand for Energy. National Renewable Energy Laboratory (NREL), Golden, CO

Burke PJ, Csereklyei Z (2016) Understanding the energy-GDP elasticity: A sectoral approach. Energy Economics 58:199–210. https://doi.org/10.1016/j.eneco.2016.07.004

Havranek T, Irsova Z, Janda K (2012) Demand for gasoline is more price-inelastic than commonly thought. Energy Economics 34:201–207. https://doi.org/10.1016/j.eneco.2011.09.003

Houthakker HS (1951) Some Calculations on Electricity Consumption in Great Britain. Journal of the Royal Statistical Society Series A (General) 114:359. https://doi.org/10.2307/2980781

Kuang H, Zhang X, Qu H, Zhu R, Li J (2024) Unraveling the effect of electricity price on electric vehicle charging behavior: A case study in Shenzhen, China. Sustainable Cities and Society 115:105836. https://doi.org/10.1016/j.scs.2024.105836

Sahari A (2019) Electricity prices and consumers' long-term technology choices: Evidence from heating investments. European Economic Review 114:19–53. https://doi.org/10.1016/j.euroecorev.2019.02.002

Zhu X, Li L, Zhou K, Zhang X, Yang S (2018) A meta-analysis on the price elasticity and income elasticity of residential electricity demand. Journal of Cleaner Production 201:169–177. https://doi.org/10.1016/j.jclepro.2018.08.027

Received: 2025-06-16 Last revision: 2025-07-18 Accepted: 2025-08-12

Partial Least Squares Structural Equation Modeling in Energy Economics: An Empirical Assessment

Martin Bať ka[1]

Abstract. Structural Equation Modeling (SEM) is a widely used tool in multivariate analysis for examining complex relationships among observed and latent variables. While covariance-based SEM (CB-SEM) is traditionally applied for theory confirmation, Partial Least Squares SEM (PLS-SEM) has gained traction for its predictive capabilities and suitability for exploratory research. This paper presents a case study applying PLS-SEM to the energy sector, specifically to investigate the factors influencing household electricity consumption. Using publicly available monthly data on electricity consumption, prices, and climatic factors, the study constructs a PLS-SEM model to assess the direct and indirect effects of market and environmental drivers on household electricity demand. The results indicate that climatic factors exert a dominant direct influence on electricity consumption, whereas the effect of market conditions, though statistically significant, is considerably weaker. The model explains a substantial proportion of the variance in electricity demand, illustrating the value of PLS-SEM in capturing complex interrelationships among latent constructs within the energy sector. This application demonstrates the potential of PLS-SEM for exploring the nuanced determinants of energy consumption and provides new insights into the modeling of demand drivers in energy economics.

Keywords: energy industry, latent variables, PLS-SEM, structural equation modeling

JEL Classification : C51
AMS Classification : 62P20

1 Introduction

Understanding the drivers of electricity consumption in the residential sector is of critical importance for energy policy, infrastructure planning, and the transition to a more sustainable energy system. Electricity demand in households is shaped by a complex interplay of economic, climatic, and behavioral factors. With the ongoing integration of renewable energy sources, increased electrification of end-uses, and the introduction of dynamic pricing schemes, the ability to reliably model and forecast demand has become an essential component of modern energy management.

Large amounts of research have been dedicated to energy economics, and it usually focused on regular observed variables, but recent events show that even abstract concepts as general mood of society and sense of social security can impact energy consumption. The aim of this study is to determine how viable and effective is Partial Least Squares Structural Equation Modeling (PLS-SEM) in explaining electricity consumption and finding potential latent variables that are effective for this purpose.

Traditional econometric approaches to modeling electricity demand have typically relied on regression-based techniques, often treating observed variables as directly measured and exogenous. However, many of the most important drivers of electricity consumption—such as the broader concepts of "market conditions" or "climate factors"—are inherently latent and multidimensional and can only be indirectly observed through a set of manifest indicators. Structural Equation Modeling (SEM) provides a powerful framework for capturing such complex relationships, allowing for the simultaneous estimation of measurement and structural models, and enabling the explicit modeling of latent constructs.

Among SEM techniques, PLS-SEM has gained popularity in recent years, particularly in fields where theory is less established, data sets are moderate in size, and the focus is on prediction or exploration rather than strict theory testing. PLS-SEM is especially suitable for situations where the data may not conform to multivariate normality or where the underlying relationships may be complex and involve multiple indicators per construct. Its flexibility

[1] Mendel University in Brno, Department of Statistics and Operation Analysis, Zemědělská 1, 61300 Brno, martin.batka@mendelu.cz

and robustness make it a valuable tool for energy economics, where observed data are often noisy and where latent constructs play a significant role in shaping behavior.

The aim of this paper is to apply PLS-SEM to quantify the relative importance of economic and climatic factors in determining household electricity consumption. By constructing a model that integrates both manifest and latent variables, this study provides new insights into how market and climate-related influences jointly shape electricity demand.

An example of using PLS-SEM in energy economics is a recent study by Belaïd (2017). It exemplifies the application of PLS-SEM path modeling in the analysis of residential energy consumption. Drawing on detailed microdata from the French 2012 PHEBUS survey, the authors utilize PLS-SEM to disentangle the direct, indirect, and total effects of household attributes, dwelling characteristics, climate, and behavioral factors on domestic energy demand. The use of PLS-SEM in this context is particularly valuable, as it allows for the simultaneous estimation of complex interrelated relationships among latent and observed variables, including indirect pathways such as the influence of household characteristics on energy use mediated by housing choices. Their findings reveal that while dwelling characteristics have the largest direct impact on energy consumption, the total effect of household-related attributes once indirect pathways are considered approaches to that of dwelling characteristics. This highlights the importance of accounting for both direct and indirect effects when designing effective residential energy policy. The study demonstrates the potential of PLS-SEM to provide nuanced insights into the multifaceted drivers of energy use and offers guidance for integrating housing and energy policy to achieve emissions reduction targets.

Recent research into energy conservation behavior has demonstrated the significant influence of social-psychological factors on household decision-making. Using structural equation modeling, one study by (Kumar et al. 2022) found that identity, social norms, and personal values strongly determine whether households adopt energy-saving measures. These results suggest that policies focused on promoting pro-environmental values and strengthening community engagement can enhance the effectiveness of household energy conservation initiatives.

In the context of technological change, the adoption of new battery electric vehicles has been analyzed across a broad sample of Chinese cities using PLS-SEM. This approach revealed that perceived usefulness and the availability of supportive infrastructure are the most important determinants of consumer willingness to adopt electric vehicles, alongside environmental awareness and policy incentives. The findings highlight the importance of addressing both behavioral motivations and practical barriers in advancing sustainable mobility solutions. (Liu et al. 2024)

Household engagement in energy-saving programs has also been explored through PLS-SEM modeling. Research in this area points to the roles of awareness, perceived barriers, and attitudes toward efficiency as significant predictors of participation in residential energy conservation options. The evidence indicates that targeted communication and efforts to lower participation barriers can meaningfully increase household involvement in energy-saving interventions. (Xu et al. 2021)

Research using structural equation modeling has demonstrated that a combination of household and building characteristics—such as dwelling size, number of occupants, income, and appliance ownership—significantly shapes annual residential energy consumption in the United States. This approach reveals key latent "X-factors" that underlie variations in household energy use by modeling both direct and indirect effects. The findings indicate that physical attributes of the dwelling, especially floor area and building type, exert the strongest direct influence on total energy consumption. At the same time, socio-economic characteristics, including household income and appliance ownership, contribute both directly and indirectly through their effect on housing choices and energy-related behaviors. Notably, the study shows that indirect effects mediated by household composition and equipment can be as substantial as the direct effects of building size. These insights highlight the value of latent variable modeling for capturing the complex interactions that drive residential energy demand and support the design of more effective, targeted energy efficiency policies.(Estiri 2014)

2 Materials and Methods

2.1 Materials

The empirical analysis in this study is based on monthly data covering key variables relevant to household electricity consumption. The dataset includes observed values for average electricity price, average temperature, heating degree days, precipitation, and sunshine hours, alongside total household electricity consumption (Q). These variables were selected to capture both economic and climatic influences on demand. The data were obtained from publicly available national statistical sources and meteorological databases, ensuring consistency and comparability across the observed periods. Specifically, the energy prices data were obtained from Czech Statistical Office

(2025), HDD data were obtained from European Commissions Joint Research Centre (2025) and other weather data were obtained from Czech Hydrometeorological Institute (2025). Prior to modeling, all variables were checked for completeness and standardized to ensure their suitability for use within the PLS-SEM framework. These variables were chosen based on the fact that hey ae frequently used in similar research such as previously mentioned studies by Belaïd (2017) and Estiri (2014).

2.2 PLS-SEM principle and estimation

In this work, the PLS-SEM method was used to analyze the relationships between factors influencing electricity consumption. PLS-SEM enables the simultaneous estimation of both the measurement and structural models, even when the data are not normally distributed or the sample size is relatively small.

The mathematical foundation of PLS-SEM consists of two key components: the measurement (outer) model and the structural (inner) model. The following **figure 1** presents the conceptual model structure. The rectangles represent observed variables and circles represent unobserved or latent variables.(Hair et al. 2017)

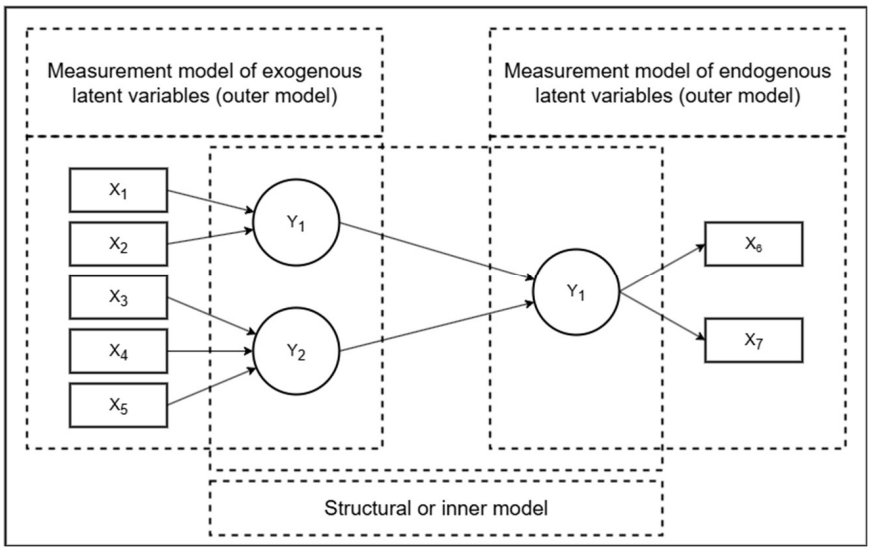


Figure 1 Conceptual model of PLS-SEM

The measurement model describes the relationship between latent variables and their manifest indicators. Its general form is given by equation 1.

$$\boldsymbol{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{1}$$

where the vector $\boldsymbol{x}$ represents the manifest indicators, $\boldsymbol{\Lambda}_x$ is the matrix of factor loadings, $\boldsymbol{\xi}$ are the latent variables, and $\boldsymbol{\varepsilon}$ is the vector of measurement errors.

The structural model formalizes the relationships between latent variables and is typically expressed as in equation 2.

$$\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \tag{2}$$

where $\boldsymbol{\eta}$ is the vector of endogenous latent variables, $\boldsymbol{\xi}$ is the vector of exogenous latent variables, $\boldsymbol{B}$ is the matrix of path coefficients between endogenous latent variables, $\boldsymbol{\Gamma}$ is the matrix of path coefficients from exogenous to endogenous variables, and $\boldsymbol{\zeta}$ is the vector of residual errors.

The PLS-SEM path modeling algorithm is an iterative, component-based approach for estimating complex relationships among observed and latent variables. The algorithm proceeds in three main stages:(Hair et al. 2017)

1. **Outer Model Estimation (Measurement Model):**
   Latent variables are estimated as weighted linear composites of their observed indicators. Initial weights are assigned, and these are iteratively updated to maximize the explained variance of the latent variables by their manifest variables.
2. **Inner Model Estimation (Structural Model):**
   Using the estimated latent variable scores, the algorithm estimates path coefficients representing the relationships between latent constructs. This typically involves ordinary least squares regression applied to the structural (inner) model equations.
3. **Iterative Updates and Convergence:**
   The algorithm alternates between updating the weights (measurement model) and estimating the path coefficients (structural model) until the changes in the estimated weights and path coefficients are sufficiently small, indicating convergence.

At convergence, the model provides estimates for both the outer model loadings/weights and the inner model path coefficients. The final output allows for assessment of the significance and strength of both measurement and structural relationships in the model. All model results are computed in SmartPLS software.

2.3 PLS-SEM model verification

The quality of the measurement model is assessed based on the values of the outer loadings and the Average Variance Extracted (AVE), given by equation 3.

$$AVE = \frac{1}{n}\sum_{i=1}^{n}\lambda_i^2 \tag{3}$$

where λ_i is the standardized loading of the i-th indicator and n is their number.

Discriminant validity is evaluated using the Fornell-Larcker criterion, according to which the square root of AVE for each construct must be greater than all correlations between this construct and other latent variables in the model.
The structural model is further evaluated using the coefficient of determination R^2 and the effect size f^2.

Overall model fit in PLS-SEM is expressed using several fit indices. The most used is SRMR (Standardized Root Mean Square Residual), which measures the average difference between observed and model-estimated correlations. Additionally, d_ULS and d_G are used. d_ULS (Unweighted Least Squares discrepancy) expresses the total difference between the empirical and model-estimated covariance (or correlation) matrices. It is calculated as the sum of squared differences between individual elements of both matrices, treating all differences equally. A lower value of d_ULS indicates a better model fit.

d_G (geodesic distance) offers an alternative way to assess fit, using a logarithmic transformation of matrix elements. This makes the index more sensitive to deviations in smaller values, thus capturing more subtle discrepancies between model and actual data. Lower d_G values again indicate a better fit.

3 Results and Model Evaluation

Figure 2 presents the conceptual structure of the estimated PLS-SEM model. The model consists of three latent variables: **Market conditions**, **Climate factors**, and **Demand drivers**. The latent variable *Market conditions* is represented by a single observed indicator, the electricity price. The latent variable *Climate factors* is defined by four manifest variables: average temperature (avg_t), heating degree days (hdd), precipitation (rain), and sunshine hours (sun_h). The latent variable *Demand drivers* (interpreted as household electricity consumption) is measured by a single observed variable Q. The variable of GDP was removed from the final model due to its low impact on overall model accuracy and low relusts in several fit indicators.

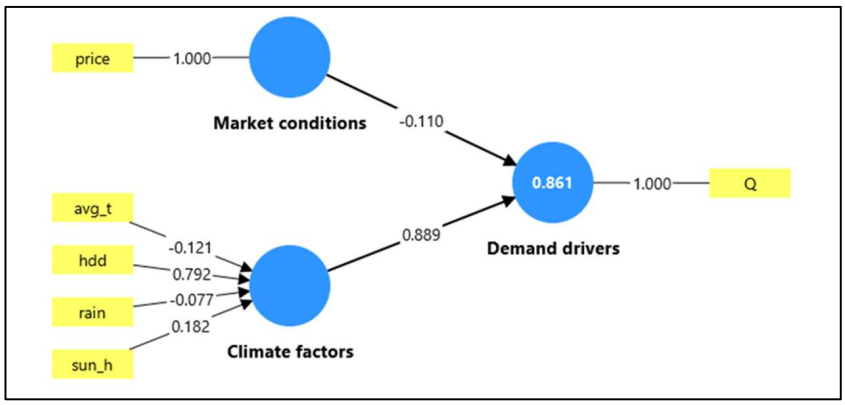


Figure 2 Representation of final model

The main result of the model shows a very high degree of explained variability in household electricity consumption. The coefficient of determination R^2 for the latent variable Demand drivers was 0.861, meaning the model explains 86.1% of the variability of this variable. This result can be considered exceptionally high, especially given the data range and the number of indicators involved.

A detailed look at the individual paths in the model structure reveals that climatic factors have a dominant influence on electricity demand. The standardized path coefficient from the Climate factors construct to Demand drivers reaches a value of 0.889, indicating a very strong positive relationship. This result confirms that seasonality,

temperature extremes, and climatic conditions in general play a key role in determining household electricity consumption.

In contrast, the path from the Market conditions construct (represented here by the average electricity price) to Demand drivers has a value of −0.110. This effect is negative (suggesting that consumption may slightly decrease as prices rise), but its magnitude is substantially smaller compared to climatic factors. Thus, in the short term, the relatively low price elasticity of electricity demand in households is confirmed.

An important indicator of the significance of individual variables is also the effect size (f^2). For climatic factors, the calculated value is $f^2 = 5.168$, which according to recommended interpretive scales represents an extremely strong effect. In contrast, for market conditions, f^2 is only 0.079, corresponding to a small to medium effect, but still statistically relevant.

The measurement model demonstrated high indicator quality and validity of the latent variables. The AVE (Average Variance Extracted) for the Climate factors construct was 0.537, and for both Demand drivers and Market conditions it was 1.000 (in these cases, the constructs had only one indicator, which is typical for AVE). All outer loadings were above or very close to the threshold of 0.5; only the indicators precipitation (−0.447) and sunshine (0.472) were weaker, but given the AVE and other validation criteria, these values are still acceptable.

Discriminant validity of the model was confirmed using the Fornell-Larcker criterion: the square roots of AVE for all latent variables were higher than their correlations with the other latent constructs. This proved that the model truly measures different concepts and there is no inappropriate overlap of meaning between the latent variables.

The global fit of the model was assessed using standard indices (SRMR, NFI, d_ULS, d_G, chi-square), all of which indicated a very good fit of the model to the data. SRMR reached a value of 0.026, and NFI was 0.991, which is well above the recommended threshold for a good fit.

Overall, it can be concluded that the presented model is robust, well specified, and validated, with clearly interpretable relationships and a strong ability to explain electricity consumption depending on key external factors. The results confirm that short-term electricity demand is fundamentally influenced by climate, while the role of prices remains marginal within the given data.

4 Discussion

The present study evaluated the viability and effectiveness of PLS-SEM in modelling household electricity consumption, with particular attention to the identification of latent variables that best capture the complex drivers of demand. The empirical results demonstrate that PLS-SEM is a powerful and versatile approach for disentangling the interrelated effects of climate and market factors on energy use, especially when the theoretical structure is only partially established and the available data may not fully meet the assumptions required by traditional methods.

Interpretation of the final model highlights the dominant role of climatic factors in shaping household electricity consumption. The path coefficient from the “Climate factors” construct to demand is both statistically strong and practically significant, supporting the prevailing consensus in the literature that temperature, heating degree days, and related variables are among the primary determinants of residential energy use. This result is consistent with previous findings that highlight the sensitivity of household electricity demand to climatic variability and the potential impact of climate change on energy systems.

Market conditions, represented in this analysis by electricity price, also show a statistically significant but substantially weaker effect on demand. This observation echoes the low price elasticity typically found in the residential sector, where electricity is an essential commodity and short-term behavioural adjustments are limited. The dominance of climatic factors over economic ones in the model reinforces the view that policy interventions aimed at demand-side management should place a strong emphasis on addressing weather-related demand peaks, through measures such as improved building insulation or the adoption of adaptive technologies.

The high explanatory power of the model, with an R^2 value of 0.861 for demand, further attests to the suitability of the selected constructs and the robustness of the PLS-SEM methodology in this context. Diagnostics for the measurement and structural models confirm that the latent constructs are well measured and the model specification is appropriate for capturing the complex relationships among variables.

Nevertheless, certain limitations should be acknowledged. The “Market conditions” construct was modelled using only a single observed indicator, which may restrict the breadth of the economic factors considered. The model does not explicitly account for behavioural or socio-demographic drivers, which have been shown in other research

to influence household energy consumption. Additionally, the study is limited to one geographic context and a specific time period, which may affect the generalizability of the findings.

Overall, these results provide important insights for policymakers and researchers. They emphasize the need to integrate climate-responsive strategies into energy planning and to recognize the relatively modest short-term impact of price-based interventions in the residential sector.

5 Conclusion

This study confirms the effectiveness of PLS-SEM in explaining household electricity consumption and identifying key latent variables relevant to energy demand. The results show that climatic factors exert the strongest direct effect on residential electricity use, while the influence of market conditions, though statistically significant, is comparatively weak. The model demonstrates high explanatory power and robustness, supporting the use of PLS-SEM for this type of economic analysis. These findings underline the importance of prioritizing climate-oriented measures in energy policy and suggest that economic instruments alone may be insufficient to achieve significant reductions in household electricity consumption. Future research should seek to broaden the scope of latent variables, including behavioral and socio-demographic factors, and test the generalizability of the results in other contexts and over longer time horizons.

Acknowledgements

This work is a part of the IGA25-PEF-DP-008 project supported by the Internal Grant Agency of the Faculty of Business and Economics (FBE) at Mendel University in Brno.

References

Belaïd F (2017) Untangling the complexity of the direct and indirect determinants of the residential energy consumption in France: Quantitative analysis using a structural equation modeling approach. Energy Policy 110:246–256. https://doi.org/10.1016/j.enpol.2017.08.027

Czech Hydrometeorological Institute (2025) Index of meteorology [dataset]. Czech Hydrometeorological Institute, Prague. https://opendata.chmi.cz/meteorology/climate/historical_csv/data/monthly/

Czech Statistical Office (2025) Electricity prices (CENY-ELEK) [dataset]. Public Database (VDB), Czech Statistical Office, Prague. https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=vystup-objekt&katalog=31779&pvo=CENY-ELEK (accessed 17 July 2025)

Estiri H (2014) Building and household X-factors and energy consumption at the residential sector: A structural equation analysis of the effects of household and building characteristics on the annual energy consumption of US residential buildings. Energy Economics 43:178–184. https://doi.org/10.1016/j.eneco.2014.02.013

European Commission – Joint Research Centre (2025) Monthly Heating and Cooling Degree Days in the European Union [dataset]. Agri4Cast Resources Portal, Ispra (Italy). https://agri4cast.jrc.ec.europa.eu/dataportal/

Hair JF, Hult GTM, Ringle CM, Sarstedt M (2017) A primer on partial least squares structural equation modeling (PLS-SEM), Second edition. SAGE, Los Angeles London New Delhi Singapore Washington DC Melbourne

Kumar P, Caggiano H, Cuite C, Andrews C, Felder F, Shwom R, Floress K, Ahmed S, Schelly Ch (2022) Behaving or not? Explaining energy conservation via identity, values, and awareness in U.S. suburban homes. Energy Research & Social Science 92:102805. https://doi.org/10.1016/j.erss.2022.102805

Liu J, Zhang N, Yang Y (2024) Acceptance model of new energy vehicles based on PLS-SEM model. Heliyon 10:e30350. https://doi.org/10.1016/j.heliyon.2024.e30350

Xu Q, Hwang B-G, Lu Y (2021) Exploring the Influencing Paths of Behavior-driven Household Energy-saving Intervention – Household Energy Saving Option (HESO). Sustainable Cities and Society 71:102951. https://doi.org/10.1016/j.scs.2021.102951

Received: 2025-06-08 Last revision: 2025-07-29 Accepted: 2025-07-29

A Vector-Based Weight Adjustment Algorithm for Portfolio Optimization Boosting Genetic Algorithm Efficiency

Jakub Neugebauer[1]

Abstract. This paper presents a vector-based weight adjustment algorithm developed to enhance the efficiency of metaheuristics in portfolio optimization. The method transforms any given weight vector into a new vector with all values fitted within a desired interval. It is primarily applicable to complex portfolio optimization problems where metaheuristics are favored over exact methods due to computational challenges. In practice, investors often restrict portfolio weights to a certain range, regardless of the chosen optimization model. Typically, such constraints are enforced by introducing penalty terms in the objective function. However, the proposed algorithm overcomes this approach by directly recalculating weights at each iteration of the underlying metaheuristic. Proposed algorithm functionality is tested on a genetic algorithm proposed for second order stochastic dominance constraints model. Results are tested on S&P 500 index selected as a benchmark and its components as a stocks from which is the portfolio constructed.

Keywords: portfolio optimization, second-order stochastic dominance, genetic algorithm, weight adjustment, mixed-integer linear programming

JEL Classification: C44, G11
AMS Classification: 90C11, 90C59

1 Introduction

Stochastic dominance provides a powerful framework for decision-making under uncertainty by comparing entire distributions rather than relying on summary statistics like mean or variance. First introduced in economics by Hadar and Russell 1969, it has since been adopted in finance to model preferences that align with risk-averse behavior. In contrast to risk measures such as variance or Value-at-Risk (VaR), stochastic dominance captures more comprehensive distributional information. The second-order stochastic dominance (SSD) condition, in particular, ensures that a portfolio is preferred by all risk-averse investors.

A major challenge in portfolio optimization under SSD constraints arises when real-world considerations, such as minimum and maximum asset allocations and maximum number of assets to invest in are added. These conditions add further constraints and binary variables into the optimization model, leading to a mixed-integer linear programming formulation (MILP). While MILP can find exact solutions, its complexity grows rapidly with the number of assets and historical return observations, making it computationally infeasible in large-scale or high-frequency applications.

To address these limitations, this paper investigates both exact and heuristic approaches to solving SSD-constrained portfolio problems. In particular, weight adjustment algorithm is proposed which modifies candidate solutions to satisfy cardinality, weight bounds, and constraints on sum of weights being 1, enabling their use in population-based metaheuristics like genetic algorithms (GA).

The structure of the paper is as follows. Section 2 formalizes the SSD framework and its formulation as an optimization problem. Section 3 introduces the weight adjustment algorithm and its integration into a GA-based solver. Section 4 presents the fitness function used for evaluating candidate portfolios under SSD and feasibility constraints. Section 5 contains numerical results comparing the proposed method with exact MILP and unadjusted GA approaches. Section 6 concludes the paper and suggests directions for future research.

2 Second Order Stochastic Dominance

Let X and Y be random variables with cumulative distribution functions $F_X(x)$ and $F_Y(x)$, respectively. Define $D^1_X(x) = F_X(x)$ and, for $s \geq 2$:

$$D^s_X(x) = \int_{-\infty}^{x} D^{s-1}_X(t)\, dt.$$

[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

Then, X dominates Y in the weak stochastic sense of order s if $D_X^s(x) \leq D_Y^s(x)$ for all x.

In portfolio optimization, SSD can be incorporated using a linear programming model known as LPSSD, originally proposed by Dentcheva and Ruszczyński 2006. Let n be the number of assets in portfolio and T the number of historical observations. Given empirical returns r_{it} and benchmark b_t, the LPSSD problem is:

$$\max \sum_{t=1}^{T} \sum_{i=1}^{n} r_{it} w_i p_t$$

subject to:

$$d_{tk} \geq b_k - \sum_{i=1}^{n} r_{it} w_i \qquad t, k = 1, \cdots, T, \tag{1}$$

$$\sum_{t=1}^{T} d_{tk} p_t \leq v_k \qquad k = 1, \cdots, T, \tag{2}$$

$$\sum_{i=1}^{n} w_i = 1, \tag{3}$$

$$d_{tk}, w_i \geq 0 \qquad i = 1, \cdots, n \quad t, k = 1, \cdots, T, \tag{4}$$

where $v_k = E((b_k - B)^+$ is the expected shartfall of a benchmark B. Assuming equal probabilities $p_t = 1/T$, the LPSSD constraints can be reformulated using a Kolmogorov-Smirnov-like statistic McFadden 1989 as:

$$KS_j = \sum_{t=1}^{T} \frac{1}{(s-1)!} \left[\max \left\{ 0, b_j - \sum_{i=1}^{n} r_{it} w_i \right\} - \max \left\{ 0, b_j - b_t \right\} \right] \leq 0, \quad \forall j. \tag{5}$$

This nonlinear reformulation enables vectorized implementation suitable for population-based optimization like genetic algorithms.

To enforce practical investment rules, we impose the following constraints:

$$w_i \geq w_{\min} z_i, \qquad \forall i \tag{6}$$

$$w_i \leq w_{\max} z_i, \qquad \forall i \tag{7}$$

$$\sum_{i=1}^{n} z_i \leq k, \tag{8}$$

$$z_i \in \{0, 1\}, \qquad \forall i \tag{9}$$

where $w_{\min}$ and $w_{\max}$ set asset-level bounds and k restricts the portfolio cardinality. These conditions jointly enforce that $z_i = 1$ if and only if $w_i > 0$.

3 Genetic and Weight Adjustment Algorithms

In this section, a weight adjustment algorithm designed to enforce practical portfolio constraints in portfolio optimization models is introduced. This procedure ensures that generated portfolios satisfy bounds on individual asset weights, maintain a specified maximum number of active positions (cardinality constraint), and remain normalized (i.e., their weights sum to 1). As illustrated in **Algorithm 1**, the `adjust_weights` routine modifies candidate solutions to meet these requirements through a series of redistribution steps applied to excess and deficit weights.

The adjustment algorithm operates column-wise on candidate weight vectors. For each portfolio, the procedure first ensures that the number of non-zero weights is sufficient to allow normalization under the given upper bound $w_{\max}$. If not, it randomly activates additional weights. Then, any weight exceeding $w_{\max}$ is clipped, and the excess is stored. To enforce the cardinality constraint, only the largest weights are retained, while smaller ones are zeroed out, further accumulating excess weight. This accumulated excess is then redistributed among weights that lie strictly between zero and $w_{\max}$, proportionally to their remaining capacity. A similar process is applied to enforce the lower bound $w_{\min}$: weights below this threshold are raised, and the resulting deficit is redistributed

among non-bound weights. Because each adjustment conserves the total weight budget—either by clipping and redistributing or reallocating—the final sum of each portfolio column remains exactly 1.

To demonstrate its application in a stochastic dominance setting, the adjustment method is integrated into a Genetic Algorithm (GA) framework presented in **Algorithm 2**. Both algorithms are represented in Table 1. The GA serves as a flexible population-based optimization tool that generates diverse candidate weight vectors, which are then filtered and corrected using the adjustment procedure. The genetic algorithm is one of the most widely adopted metaheuristics, belonging to the family of evolutionary algorithms (EA), originally introduced by Holland 1975 and further developed in Holland 1992. Although the GA itself is not the primary contribution of this work, it provides a natural and effective environment to evaluate the feasibility-enforcing behavior of the proposed adjustment algorithm under empirical return data and structural portfolio constraints. The parameters of the GA were selected by grid search as follows: *patience* = 50, $p_{\text{mut}} = 0.5$, $\mu = 0.25$, $\sigma_m = 0.25$, $p_{\text{inj}} = 0.5$, and *popsize* = 5.

Algorithm 1 `adjust_weights` Procedure

$W_{\text{adj}} \leftarrow W$
$m \leftarrow \lceil 1/w_{\max} \rceil$
for each column $j = 1$ to $popsize$ **do**
  if non-zeros in $j < m$ **then**
    Randomly activate zeros with values in $[w_{\min}, w_{\max}]$
  end if
  Normalize column j so $\sum w_{ij} = 1$
end for
Initialize $excess, deficit \in \mathbb{R}^{popsize}$ as zero vectors
for each column $j = 1$ to $popsize$ **do**
  for each $i = 1$ to n **do**
    if $w_{ij} > w_{\max}$ **then**
      $excess[j] \leftarrow excess[j] + (w_{ij} - w_{\max})$
      $w_{ij} \leftarrow w_{\max}$
    end if
  end for
  Enforce cardinality: keep top weights, zero the rest, update $excess[j]$
  Redistribute $excess[j]$ over $0 < w_{ij} < w_{\max}$
  for each i **do**
    if $0 < w_{ij} < w_{\min}$ **then**
      $deficit[j] \leftarrow deficit[j] + (w_{\min} - w_{ij})$
      $w_{ij} \leftarrow w_{\min}$
    end if
  end for
  Redistribute $deficit[j]$ over $w_{\min} < w_{ij} < w_{\max}$
end for
return W_{adj}

Algorithm 2 Genetic Algorithm for SD

Initialize population $\mathbf{W}$
if adj_weights is **True then**
  $\mathbf{W} \leftarrow$ `adjust_weights`$(\mathbf{W})$
end if
$patience \leftarrow 0, best_fitness \leftarrow -\infty$
$fitness \leftarrow$ `evaluate`$(\mathbf{W})$
while $patience < max_patience$ **do**
  Generate children $\mathbf{W}'$ via crossover (tournament, size $0.5 \cdot popsize$)
  for each child w_i in $\mathbf{W}'$ **do**
    Mutation: with p_{mut}:
      For $w_{ij} > 0$, mutate with μ: $w_{ij} \leftarrow w_{ij} + \mathcal{N}(0, \sigma_m)$
    Injection: with p_{inj}:
      Replace random j with $v \sim \mathcal{U}(w_{\min}, w_{\max})$
      Normalize w_i
  end for
  if adj_weights is **True then**
    $\mathbf{W}' \leftarrow$ `adjust_weights`$(\mathbf{W}')$
  end if
  $\mathbf{W}_{\text{best}} \leftarrow$ `top_half`$(\mathbf{W})$
  $\mathbf{W} \leftarrow \mathbf{W}_{\text{best}} \cup \mathbf{W}'$
  $fitness \leftarrow$ `evaluate`$(\mathbf{W})$
  $current_best \leftarrow \max(fitness)$
  if $current_best > best_fitness$ **then**
    $best_fitness \leftarrow current_best, patience \leftarrow 0$
  else
    $patience \leftarrow patience + 1$
  end if
end while
return best-performing w

Table 1 *adjust_weights* and Genetic Algorithm algorithms

The fitness function balances return maximization with constraint satisfaction by penalizing violations of stochastic dominance and portfolio feasibility. For a portfolio with weight vector $\mathbf{w}$, its fitness is based on the expected return $\mathbb{E}[\mathbf{rw}] = \sum_{t=1}^{T} \sum_{i=1}^{n} r_{it} w_i p_t$ and a penalty for dominance violation, measured by scaling the maximal positive deviation in (5). When weight adjustment is applied (`adj_weights = True`), the structural constraints are already satisfied, so the fitness is defined as:

$$\text{fitness}(\mathbf{w}) = \mathbb{E}[\mathbf{rw}] - (\max\{0, \max(KS)\} \cdot \text{scale})^2 ,$$

where $\max(KS)$ denotes the maximum violation of the stochastic dominance constraint over all benchmark points b_j.

If the adjustment is not applied (`adj_weights = False`), additional penalties are included for violating portfolio constraints:

$$\text{fitness}(\mathbf{w}) = \mathbb{E}[\mathbf{rw}] - (\max\{0, \max(KS) - \text{cv}\} \cdot \text{scale} + \text{violations}(\mathbf{w}))^2 ,$$

where the violation term is defined as:

$$\text{violations}(\mathbf{w}) = \sum_{i=1}^{n} \mathbf{1}\{w_i > w_{\max}\} + \sum_{i=1}^{n} \mathbf{1}\{0 < w_i < w_{\min}\} + \max\left\{ \sum_{i=1}^{n} \mathbf{1}\{w_i > 0\} - k,\ 0 \right\} .$$

This formulation ensures that only feasible portfolios which respect second-order stochastic dominance achieve high fitness values, while infeasible or dominated portfolios are penalized accordingly. The parameter `scale` is set as scale = $\frac{1}{\left(\frac{1}{T}\cdot 0.001\right)}$, which amplifies the penalty such that a deviation equivalent to a return change of 0.001 is weighted comparably to the average return.

Quadratic penalty is used for the stability and smooth convergence of the genetic algorithm. It strongly penalizes large violations while maintaining gradient-like behavior around constraint boundaries, helping the population to gradually correct infeasible solutions rather than immediately discarding them.

All operations in the case study are implemented using matrix-based computations in the TensorFlow framework, which enables efficient evaluation over large populations and leverages GPU acceleration.

4 Results Comparison

The results of multiple testing sets are collected for different optimization techniques discussed in previous sections: mixed-integer linear programming based on stochastic dominance ($MILPSSD$), genetic algorithm with weight adjustment (GA_{adj}), and genetic algorithm without adjustment (GA_{noadj}). For each method, the following performance statistics are collected for comparative analysis:

- **Mean(%)**: average percentage of the $MILPSSD$-optimal return achieved by the sub-optimal solution.
- **Std(%)**: standard deviation of that sub-optimal performance.
- **SSDs**: average scaled stochastic-dominance deviation (higher values indicate stronger constraint violation).
- **Low(%)**: total shortfall below the minimum weight bound, expressed in percentage.
- **High(%)**: total excess above the maximum weight bound, in percentage.
- **Sum(%)**: deviation of the portfolio weight sum from unity.
- **Card**: number of assets exceeding the allowed cardinality.
- **T Mean(s)**: average run time per optimization, in seconds.
- **T Std(s)**: standard deviation of the run time.

For each of the five evaluation sets and each considered investment history length h, the $MILPSSD$ model was solved once using the `HiGHS` solver[1], and its execution time was recorded. To assess convergence behavior, both GA_{adj} and GA_{noadj} were run 30 times for each combination of set and history length. The key convergence indicators are **Mean(%)** and **Std(%)**, which reflect how closely and consistently the GA solutions approximate the optimal return.

The return and benchmark data used in this study are based on daily historical prices of S&P 500 index components and the index itself, collected from `Yahoo Finance`[2]. The data span the full decade from January 1, 2015 to January 1, 2025. To capture varying market conditions, five representative test sets were selected uniformly across this time interval, each corresponding to a different market regime or structural behavior.

The results in Table 2 confirm the dominant performance of the weight-adjusted genetic algorithm (GA_{adj}) in reliably converging towards the optimal solutions produced by $MILPSSD$. Across all data sets and history lengths, GA_{adj} consistently achieves mean performance values close to 100% of the MILP optimum, often with extremely low standard deviations, demonstrating both accuracy and stability of the optimization.

Minor feasibility deviations can occasionally appear—for example, a small scaled stochastic-dominance deviation (SSDs = 0.21) in set 1 with $h = 250$—which indicate rare cases of constraint relaxation or slow convergence under difficult conditions. Interestingly, even the $MILP$ solver provides rounding violations in certain cases, such as low weight bound violations (Low(%) = 2.00) in set 3. These observations confirm the practical difficulty of strict feasibility even in commercial solvers.

In contrast, the genetic algorithm without weight adjustment (GA_{noadj}) shows significantly less stable behavior. Although it sometimes achieves near-optimal mean performance, its standard deviation is often much higher and its convergence far less reliable. The lack of an internal feasibility enforcement mechanism causes increased sensitivity to initialization, mutation effects, and constraint violations, ultimately limiting its robustness across repeated runs.

From a computational perspective, the advantage of the proposed approach becomes even more apparent. The run time of GA_{adj} remains nearly constant with respect to the history length h, showing clear scalability benefits. In contrast, $MILPSSD$ exhibits exponential growth in computation time with increasing h, reaching an extreme

[1] `https://highs.dev/`
[2] `https://finance.yahoo.com`

Set	h	Method	O Mean(%)	O Std(%)	SSDs	Low(%)	High(%)	Sum(%)	Card	T Mean(s)	T Std(s)
1		$LPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00
	20	GA_{adj}	100.00	0.00	0.00	0.00	0.00	0.00	0.00	1.74	0.49
		GA_{noadj}	90.19	24.30	0.00	0.00	0.00	0.00	0.00	3.46	1.26
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	14.21	0.00
	60	GA_{adj}	99.23	0.46	0.00	0.00	0.00	0.00	0.00	4.41	1.34
		GA_{noadj}	97.27	8.15	0.00	0.00	0.00	0.00	0.00	4.14	1.08
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	51.56	0.00
	125	GA_{adj}	99.89	0.20	0.00	0.00	0.00	0.00	0.00	4.26	0.62
		GA_{noadj}	94.81	17.76	0.00	0.00	0.00	0.00	0.00	4.80	1.75
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	601.89	0.00
	250	GA_{adj}	98.65	0.41	0.21	0.00	0.00	0.00	0.00	4.09	1.84
		GA_{noadj}	97.90	0.94	0.06	0.00	0.00	0.00	0.00	4.06	1.33
2		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00
	20	GA_{adj}	99.99	0.06	0.00	0.00	0.00	0.00	0.00	1.53	0.04
		GA_{noadj}	88.74	27.92	0.00	0.00	0.00	0.00	0.00	3.39	0.96
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	5.07	0.00
	60	GA_{adj}	99.66	0.30	0.00	0.00	0.00	0.00	0.00	3.28	1.10
		GA_{noadj}	95.22	16.48	0.00	0.00	0.00	0.00	0.00	3.65	0.96
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	94.24	0.00
	125	GA_{adj}	99.52	0.16	0.01	0.00	0.00	0.00	0.00	4.10	1.18
		GA_{noadj}	99.46	0.24	0.01	0.00	0.00	0.00	0.00	5.13	2.08
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	19594.31	0.00
	250	GA_{adj}	99.33	0.39	0.01	0.00	0.00	0.00	0.00	6.25	2.25
		GA_{noadj}	97.05	9.72	0.00	0.00	0.00	0.00	0.00	5.62	2.06
3		$MILPSSD$	100.00	0.00	0.00	2.00	0.00	0.00	0.00	0.70	0.00
	20	GA_{adj}	99.70	0.17	0.00	0.00	0.00	0.00	0.00	3.51	0.36
		GA_{noadj}	93.30	15.77	0.00	0.00	0.00	0.00	0.00	3.78	1.30
		$MILPSSD$	100.00	0.00	0.00	2.00	0.00	0.00	0.00	5.34	0.00
	60	GA_{adj}	99.40	0.30	0.00	0.00	0.00	0.00	0.00	3.96	1.31
		GA_{noadj}	95.77	12.55	0.00	0.00	0.00	0.00	0.00	3.79	1.72
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	9.39	0.00
	125	GA_{adj}	99.71	0.54	0.00	0.00	0.00	0.00	0.00	3.75	0.77
		GA_{noadj}	98.75	1.11	0.00	0.00	0.00	0.00	0.00	4.31	0.46
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	643.72	0.00
	250	GA_{adj}	99.29	0.33	0.02	0.00	0.00	0.00	0.00	3.50	0.73
		GA_{noadj}	99.10	0.72	0.00	0.00	0.00	0.00	0.00	4.30	0.97
4		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00
	20	GA_{adj}	99.90	0.00	0.00	0.00	0.00	0.00	0.00	1.62	0.06
		GA_{noadj}	92.99	21.04	0.00	0.00	0.00	0.00	0.00	4.06	1.00
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	8.71	0.00
	60	GA_{adj}	99.68	0.12	0.00	0.00	0.00	0.00	0.00	3.71	0.94
		GA_{noadj}	97.14	10.03	0.00	0.00	0.00	0.00	0.00	4.35	1.04
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	52.49	0.00
	125	GA_{adj}	99.34	0.51	0.00	0.00	0.00	0.00	0.00	3.84	1.08
		GA_{noadj}	93.16	19.86	0.00	0.00	0.00	0.00	0.00	3.79	1.33
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	104.28	0.00
	250	GA_{adj}	99.22	0.60	0.01	0.00	0.00	0.00	0.00	4.99	1.72
		GA_{noadj}	94.48	11.71	0.03	0.00	0.00	0.00	0.00	5.63	2.14
5		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.52	0.00
	20	GA_{adj}	99.42	0.29	0.00	0.00	0.00	0.00	0.00	3.61	0.85
		GA_{noadj}	97.06	7.61	0.00	0.00	0.00	0.00	0.00	4.15	0.96
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	12.80	0.00
	60	GA_{adj}	99.42	0.57	0.00	0.00	0.00	0.00	0.00	5.42	1.37
		GA_{noadj}	98.85	0.70	0.00	0.00	0.00	0.00	0.00	5.91	1.70
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	63.95	0.00
	125	GA_{adj}	98.68	0.42	0.00	0.00	0.00	0.00	0.00	3.09	0.58
		GA_{noadj}	94.63	13.13	0.00	0.00	0.00	0.00	0.00	3.79	1.14
		$MILPSSD$	100.00	0.00	0.00	0.00	0.00	0.00	0.00	1169.48	0.00
	250	GA_{adj}	99.73	0.14	0.01	0.00	0.00	0.00	0.00	5.29	2.24
		GA_{noadj}	99.62	0.26	0.05	0.00	0.00	0.00	0.00	5.21	1.48

Table 2 Effectiveness of GA with and without adjusting weights compared to MILP optimum.

case of 19,594.31 seconds (i.e., over 5.4 hours) in set 2. Such results highlight the impracticality of exact MILP formulations for large-scale or high-frequency investment scenarios, while GA_{adj} maintains both feasibility and speed, making it highly suitable for practical portfolio optimization under stochastic dominance constraints.

5 Conclusion

This paper proposed a weight adjustment algorithm to enforce structural portfolio constraints within optimization frameworks based on second-order stochastic dominance (SSD). While exact solutions using mixed-integer linear programming (MILPSSD) provide benchmark optimality, they suffer from severe scalability limitations as the problem size increases. In contrast, population-based metaheuristics like genetic algorithms (GA) offer flexibility and computational efficiency but struggle with constraint satisfaction in the absence of a dedicated feasibility mechanism.

The introduced adjustment algorithm guarantees satisfaction of cardinality, weight bounds, and normalization by redistributing excesses and deficits in asset weights, all while preserving the total budget. When integrated into the GA framework, the method significantly enhances convergence stability, reduces constraint violations, and maintains near-optimal performance across various test scenarios.

Extensive empirical analysis confirms that GA with weight adjustment (GA_{adj}) reliably approximates MILPSSD solutions even in high-dimensional settings, while unadjusted variants (GA_{noadj}) suffer from greater variability. Moreover, the proposed method maintains low computational complexity and benefits from vectorized matrix operations, making it well-suited for large-scale practical applications.

Future research could explore the integration of the adjustment mechanism into other metaheuristics, investigate adaptive adjustment schemes, and apply the method in more dynamic market environments with rolling horizon optimization or transaction cost modeling.

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.

References

Dentcheva D, Ruszczyński A (2006) Portfolio optimization with stochastic dominance constraints. Journal of Banking Finance 30(2):433–451. https://doi.org/https://doi.org/10.1016/j.jbankfin.2005.04.024, https://www.sciencedirect.com/science/article/pii/S0378426605000725, risk Management and Optimization in Finance

Hadar J, Russell WR (1969) Rules for ordering uncertain prospects. The American Economic Review 59(1):25–34. http://www.jstor.org/stable/1811090

Holland JH (1975) Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence. University of Michigan Press

Holland JH (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press

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

Received: 2025-06-08 Last revision: 2025-06-08 Accepted: 2025-06-26

Convex Enclosures for Interval Linear Programming

Elif Garajová[1], Miroslav Rada[2]

Abstract. Interval programming provides a mathematical framework for representing uncertainty in optimization models. In an interval programming model, it is assumed that only lower and upper bounds on the exact data are known in advance, and the real values can be independently perturbed within these bounds.
Even in the linear case, interval programming often leads to nonconvex solution sets, such as the set of all feasible or optimal solutions. Such nonconvex sets are traditionally approximated through interval-box enclosures, which provide a simple outer approximation but often suffer from significant overestimation. A more accurate alternative involves representing the nonconvex set as a union of interval boxes, which can improve precision, but is often computationally demanding and can lead to overly complex representations, especially in higher dimensions.
To balance accuracy and efficiency, we propose to investigate the use of convex polyhedral approximations beyond simple interval boxes. In particular, convexification techniques from nonlinear programming, such as McCormick relaxations, can be adapted to construct more sophisticated convex enclosures. This research direction aims to provide more efficient yet sufficiently accurate representations of the nonconvex sets, supporting better algorithms for solving key problems in interval optimization.

Keywords: linear programming, interval data, convex enclosure

JEL Classification: C44, C61
AMS Classification: 90C31, 90C59

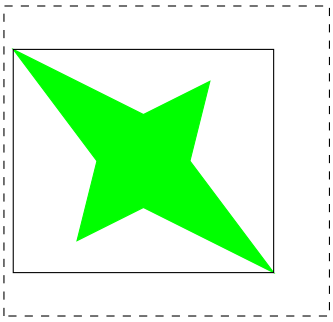
1 Introduction

Interval linear programming is an extension of classical linear programming, which allows to incorporate uncertainty and imprecision inherently present in practical problems of operations research. In this framework, some or all of the data are given as intervals rather than exact numbers, allowing each uncertain parameter to vary independently within its defined range. The main tasks commonly studied in interval programming are analogous to those of classical linear programming: computing the optimal values and finding solutions that are feasible or optimal across all admissible values within the specified intervals with respect to a given solution concept (see the surveys by Rohn (2006) and Hladík (2025) for an overview of results). Interval programming models have found a range of applications in operations research, including uncertain models for transportation problems (Garajová and Rada, 2023) or portfolio optimization (Chaiyakan and Thipwiwatpotjana, 2021).

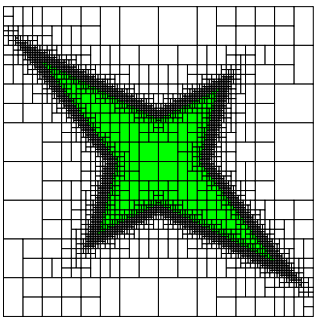
One of the challenges in solving interval linear programs lies in the nonconvexity of the considered solution sets, which stems from the fact that the uncertainty is modeled through independent interval variations in the data. This significantly complicates both the analysis of the problems and computation of the solutions, as many classical optimization techniques rely on convexity to guarantee computational tractability and optimality of the obtained solution. In particular, the united set of all feasible solutions can be nonconvex and even disconnected, if the solutions are not restricted to a single orthant (Oettli and Prager, 1964). Moreover, even if the feasible set is convex, the united set of all optimal solutions can still be nonconvex (Garajová and Hladík, 2019).

Since computations over the exact feasible or optimal set of an interval program may be too time-consuming, methods for approximating the sets are also of interest. A traditional approach to approximating nonconvex sets in interval analysis and interval optimization is to construct an interval envelope, i.e. an axis-aligned box that fully contains the considered set (Skalna, 2018; Mishmast Nehi et al, 2020). However, even computing the tightest possible interval envelope of the feasible set (known as the interval hull; see also Figure 1a) leads to an NP-hard problem, in general (Rohn and Kreinovich, 1995). While efficient methods for interval-envelope approximations are available in the literature, the restriction to axis-aligned boxes can lead to significant overestimation of the original nonconvex set. Nevertheless, interval envelopes remain attractive due to their simplicity: they are easy to compute, store, and handle, as their representation requires only the lower and upper bounds in each coordinate. As such, they can also serve as initial enclosures further refined by more sophisticated approximation methods.

[1] Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Econometrics, nám. W. Churchilla 4, Prague, Czech Republic, elif.garajova@vse.cz
[2] Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Econometrics & Faculty of Finance and Accounting, Department of Financial Accounting and Auditing, nám. W. Churchilla 4, Prague, Czech Republic, miroslav.rada@vse.cz



(a) An approximation by the interval hull (solid line) and by an interval envelope (dashed line).



(b) A paving approximation (feasible boxes in green, undetermined boxes in gray and infeasible boxes in white).

Figure 1 Commonly used approximations of the weakly feasible set of an interval linear program.

An alternative to using a single interval box for approximating nonconvex sets is to generate a collection of boxes whose union encloses the set of interest. This is typically achieved through a branch-and-prune framework, which recursively subdivides the search space into non-overlapping interval boxes categorized into three types: those entirely contained within the set, those that do not intersect the set, and undetermined boxes (see Figure 1b). This approach was originally developed for set inversion via interval analysis (Jaulin et al, 2001), and has since been extended through the introduction of new pruning tests tailored to the characterization of feasible sets in parametric interval linear systems (Rada et al, 2023). The main advantage of this approach is that it can yield a tighter approximation that captures the shape of the nonconvex region more accurately. However, the branch-and-prune algorithm can be computationally intensive and may produce a very large number of boxes, especially for high-dimensional problems or sets with complex geometry.

We aim to strike a balance between the two aforementioned approaches by using more general convex polyhedra to obtain approximations that are computationally efficient and tighter than using axis-aligned boxes, while still maintaining a manageable level of complexity in the representation. So far, this approach has been less explored in the literature on interval programming. Namely, Hladík (2012) designed a contractor for approximating the optimal solution set using a linear approximation for the absolute-value description of the set. In this paper, we propose to apply McCormick envelopes (McCormick, 1976), which yield over- and underestimators for bilinear terms. To make use of these envelopes in the context of interval systems, we model the interval parameters as bounded variables, thereby enabling the formulation of a convex McCormick relaxation for the set of feasible or optimal solutions. A similar technique was also previously employed by Mohammadi and Gentili (2021) to derive approximations for the outcome range problem in interval programming.

2 Interval Linear Programming and Nonconvex Sets

Let us first introduce the notation and formalize the terminology used in the paper. We use the symbol $\mathbb{IR}$ to denote the set of closed real intervals; bold letters are used to denote interval objects. Given two real matrices $\underline{A}, \overline{A} \in \mathbb{R}^{m\times n}$ satisfying $\underline{A} \le \overline{A}$, we define an *interval matrix* $\mathbf{A} = [\underline{A}, \overline{A}] \in \mathbb{IR}^{m\times n}$ as the set

$$\mathbf{A} = \{A \in \mathbb{R}^{m\times n} : \underline{A} \le A \le \overline{A}\}, \tag{1}$$

where all inequalities are understood componentwise. Alternatively, an interval matrix can be given in terms of its *center* $A_c = \frac{1}{2}(\overline{A} + \underline{A})$ and *radius* $A_\Delta = \frac{1}{2}(\overline{A} - \underline{A})$ as $\mathbf{A} = [A_c - A_\Delta, A_c + A_\Delta]$.

As a special case of an interval matrix, we define an *interval vector* $\mathbf{a} = [\underline{a}, \overline{a}] \in \mathbb{IR}^n$ as the set $\mathbf{a} = \{a \in \mathbb{R}^n : \underline{a} \le a \le \overline{a}\}$. From a geometric perspective, an interval vector is also often referred to as an *interval box*.

For given interval constraint matrices $\mathbf{A} \in \mathbb{IR}^{m_1\times n}$ and $\mathbf{C} \in \mathbb{IR}^{m_2\times n}$, interval right-hand-side vectors $\mathbf{b} \in \mathbb{IR}^{m_1}$ and $\mathbf{d} \in \mathbb{IR}^{m_2}$ and an interval objective vector $\mathbf{c} \in \mathbb{IR}^n$, we define an *interval linear program*

$$\text{maximize } \mathbf{c}^T x \quad \text{subject to } \mathbf{A}x \le \mathbf{b},\ \mathbf{C}x = \mathbf{d}, \tag{2}$$

as the set of all linear programs in the same form with $A \in \mathbf{A}$, $C \in \mathbf{C}$, $b \in \mathbf{b}$, $c \in \mathbf{c}$, $d \in \mathbf{d}$. A particular linear program with fixed coefficients that is contained in this set is called a *scenario* of interval program (2).

Note that the standard transformations between different formulations are not always applicable for interval linear programs, and as a result, their properties may differ significantly depending on the formulation (see Garajová et al (2019) for details).

Various solution concepts have been proposed for interval programs, with weak feasibility and weak optimality among the most commonly used. A given vector $x \in \mathbb{R}^n$ is said to be a *weakly feasible solution*, if it is feasible for at least one scenario of the interval program. Similarly, a vector $x \in \mathbb{R}^n$ is a *weakly optimal solution*, if it is optimal for at least one scenario of the interval program.

A characterization of the set of all weakly feasible solutions is known for both systems of interval linear equations and inequalities. Theorem 1 states the description of the weakly feasible set via a system of absolute-value inequalities, which forms one of the seminal results derived in interval linear algebra and optimization.

Theorem 1 (Oettli and Prager, 1964; Gerlach, 1981)**.**

$$x \in \mathbb{R}^n \text{ is a weakly feasible solution of } \mathbf{A}x = \mathbf{b} \;\Leftrightarrow\; |A_c x - b_c| \leq A_\Delta |x| + b_\Delta$$
$$x \in \mathbb{R}^n \text{ is a weakly feasible solution of } \mathbf{A}x \leq \mathbf{b} \;\Leftrightarrow\; A_c x - A_\Delta |x| \leq \overline{b}$$

Note that if all variables are restricted to a single orthant, the characterization simplifies to a linear system, making the feasible set a convex polyhedron. However, in the general case, the feasible set is a union of convex polyhedra, one corresponding to each orthant of the search space.

It directly follows that the set of all weakly optimal solutions can also be nonconvex. Notably, this nonconvexity may arise even when the weakly feasible set itself is convex (see also (Garajová and Hladík, 2019) for examples and details). In this paper, we focus on deriving an approximation of the weakly feasible set for interval linear systems. However, the same approach can also be extended to approximate the weakly optimal set of an interval program, leveraging the fact that a relaxation of the optimal set can be obtained as the feasible set of an interval system through linear programming duality.

3 McCormick Relaxations for Interval Linear Programming

To derive a convex enclosure for the weakly feasible (and weakly optimal) solution set of an interval linear program, we utilize the McCormick envelopes (McCormick, 1976). These envelopes provide a standard technique for convexifying bilinear terms in challenging nonlinear optimization problems.

Assume a bilinear function $f(x, y) = xy$ is given with bounded variables $\underline{x} \leq x \leq \overline{x}$ and $\underline{y} \leq y \leq \overline{y}$. In the McCormick relaxation, the bilinear product is replaced with a new variable z_{xy} that is constrained by four additional inequalities representing the underestimators and overestimators (see also Figure 2):

$$\begin{aligned} z_{xy} &\geq \underline{x}y + x\underline{y} - \underline{x}\underline{y}, & z_{xy} &\geq \overline{x}y + x\overline{y} - \overline{x}\overline{y}, \\ z_{xy} &\leq \overline{x}y + x\underline{y} - \overline{x}\underline{y}, & z_{xy} &\leq x\overline{y} + \underline{x}y - \underline{x}\overline{y}. \end{aligned} \tag{3}$$

Note that the tightness of the McCormick relaxation depends on the variable bounds: the narrower the bounds, the tighter the approximation. Consequently, improving bound estimates or partitioning the search space into smaller regions can significantly enhance the quality of the relaxation.

Consider an interval linear system that defines the weakly feasible set of an interval program. For simplicity, we focus on systems constrained by inequalities in the form $\mathbf{A}x \leq \mathbf{b}$; however, systems with mixed constraints can be treated analogously. We also assume that the weakly feasible set is bounded, which is typically the case in practical optimization problems. The bounds $\underline{x}_j, \overline{x}_j$ on the variables x_j can be obtained using existing methods for computing interval envelopes for the weakly feasible solution set of interval linear systems (Skalna, 2018).

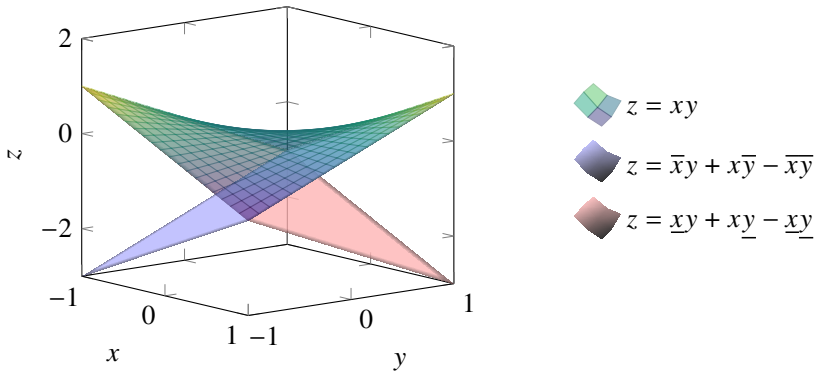


Figure 2 Visualization of McCormick underestimators for the bilinear function $f(x, y) = xy$ with $x, y \in [-1, 1]$.

We can observe that when describing the weakly feasible solutions, rather than distinguishing between the decision variables and the interval parameters, we can also treat both as variables and then obtain the feasible set as the projection on the original decision variables. This approach leads to creating bilinear terms in the constraint expressions. Namely, we obtain the nonlinear system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &\le b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &\le b_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &\le b_m, \end{aligned} \tag{4}$$

where the variables corresponding to the interval parameters are bounded to their respective intervals:

$$\begin{aligned} \underline{A}_{11} \le a_{11} \le \overline{A}_{11}, \ \ldots, \ \underline{A}_{mn} \le a_{mn} \le \overline{A}_{mn}, \\ \underline{b}_1 \le b_1 \le \overline{b}_1, \ \ldots, \ \underline{b}_m \le b_m \le \overline{b}_m. \end{aligned} \tag{5}$$

Now, we can apply the McCormick relaxation to system (6). First, we substitute new variables z_{ij} for the bilinear terms $a_{ij}x_j$ to obtain the system of linear inequalities

$$\begin{aligned} z_{11} + z_{12} + \cdots + z_{1n} &\le b_1, \\ z_{21} + z_{22} + \cdots + z_{2n} &\le b_2, \\ &\vdots \\ z_{m1} + z_{m2} + \cdots + z_{mn} &\le b_m. \end{aligned} \tag{6}$$

Then, we add the McCormick inequalities bounding the variables z_{ij}:

$$\begin{aligned} z_{ij} &\ge \underline{A}_{ij}x_j + a_{ij}\underline{x}_j - \underline{A}_{ij}\underline{x}_j, \quad & z_{ij} &\ge \overline{A}_{ij}x_j + a_{ij}\overline{x}_j - \overline{A}_{ij}\overline{x}_j, \\ z_{ij} &\le \overline{A}_{ij}x_j + a_{ij}\underline{x}_j - \overline{A}_{ij}\underline{x}_j, \quad & z_{ij} &\le \underline{A}_{ij}x_j + a_{ij}\overline{x}_j - \underline{A}_{ij}\overline{x}_j. \end{aligned} \tag{7}$$

Using this technique, we obtain a convex (linear) relaxation of the weakly feasible solution set of an interval linear system $\mathbf{A}x \le \mathbf{b}$. For a system with m constraints and n variables, the relaxation introduces a polynomial number of variables and constraints with respect to the size of the original interval system: there are $2mn$ additional variables and $4mn$ additional inequalities. This approach is particularly effective when only a subset of the coefficients is affected by uncertainty, as it allows for a reduction in the number of additional variables and constraints needed.

Example 1. Consider the interval linear system

$$\begin{aligned} 3x_1 + [-3,\ \ 1]x_2 &\le 2, \\ [\ \ 0,\ \ 4]x_1 + \qquad x_2 &\le 2, \\ [-3,-1]x_1 + \qquad x_2 &\le 5, \\ -2x_1 + [-4,-2]x_2 &\le 4. \end{aligned} \tag{8}$$

Furthermore, assume that an initial interval envelope $\mathbf{x} = ([-5,5],[-5,5])$ enclosing the weakly feasible solution set is given. The exact weakly feasible set of system (10), which can be characterized by the absolute-value inequalities stated in Theorem 1, is depicted in Figure 3. The figure also shows the interval hull of the weakly feasible set, which is represented by the interval box $\mathbf{x}^H = ([-4, \frac{8}{3}], [-4, \frac{22}{5}])$.

The McCormick envelopes for the feasible set are derived for each of the bilinear terms associated with the four interval coefficients. For example, for the interval coefficient $\mathbf{a}_{12} = [-3, 1]$ in the first inequality, we linearize the product $\mathbf{a}_{12}x_2$ using the set of constraints

$$\begin{aligned} z_{12} &\ge -3x_2 - 5a_{12} - 15, \quad & z_{12} &\ge x_2 + 5a_{12} - 5, \\ z_{12} &\le x_2 - 5a_{12} + 5, \quad & z_{12} &\le 5a_{12} - 3x_2 + 15, \end{aligned} \tag{9}$$

with the original interval bounds $-3 \le a_{12} \le 1$. Then, we can replace the first constraint of system (10) by the linear constraint $3x_1 + z_{12} \le 2$.

Let us now solve the interval linear program

$$\min x_1 + x_2 \ \text{ subject to (10)} \tag{10}$$

using different approximations of the feasible set.

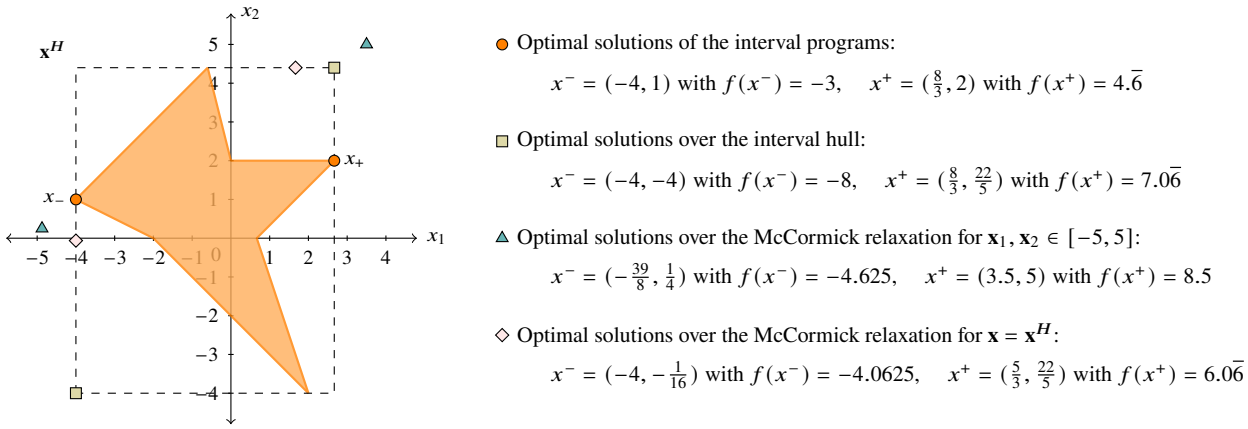


Figure 3 The weakly feasible set of interval linear system (10), its interval hull and approximations of the optimal value of interval linear programs (12) and (13).

If we use the interval hull as an approximation for the feasible set, we obtain the solution $(-4, -4)$ with the objective value -8. On the other hand, optimizing over the McCormick relaxation yields the objective value $-\frac{37}{8} = -4.625$ attained at $(-\frac{39}{8}, \frac{1}{4})$. Here, we obtain a tighter approximation of the exact optimal value of the interval linear program, which is the value -3 corresponding to the weakly feasible solution $(-4, 1)$.

Consider now the maximization version of program (12), i.e. the interval program

$$\text{maximize } x_1 + x_2 \text{ subject to (10).} \tag{11}$$

When optimizing over the interval hull, we obtain the solution $(\frac{8}{3}, \frac{22}{5})$ with the objective value $\frac{106}{15} = 7.0\overline{6}$. The McCormick relaxation leads to the objective value 8.5 attained at $(3.5, 5)$. In this case, the value obtained by using the interval hull gives a better approximation of the optimal value computed on the exact weakly feasible set, which is $\frac{14}{3} = 4.\overline{6}$ at the solution $(\frac{8}{3}, 2)$.

However, note that if we start with the initial bounds on the variables x_1 and x_2 determined by the interval hull $\mathbf{x}^H$ instead of the intervals $[-5, 5]$, the McCormick envelopes are able to improve the approximation of the optimal objective value to $-\frac{65}{16} = -4.0625$ at $(-4, -\frac{1}{16})$ for the minimization program, and the value $\frac{91}{15} = 6.0\overline{6}$ at $(\frac{5}{3}, \frac{22}{5})$ for the maximization program.

This example demonstrates that McCormick envelopes can offer a better approximation of the nonconvex sets arising in interval programming. Nevertheless, this advantage does not hold universally. In certain instances, the interval hull (or a tight interval envelope) may yield better values, depending on the objective function, the geometry of the feasible region and tightness of the variable bounds. Consequently, combining multiple techniques, such as intersecting a simple and fast-to-compute interval envelope with the McCormick relaxation, can lead to improved approximations.

4 Conclusion

In this work, we addressed the challenge of approximating the nonconvex feasible and optimal solution sets arising in uncertain linear programs with interval data. Our aim was to highlight convex enclosures that achieve a balance between tightness of approximation and complexity of the description. Since the sets of interest can be characterized by systems with bilinear constraints, we proposed utilizing McCormick envelopes to construct convex outer approximations. This approach enables capturing at least some of the structure induced by interval uncertainty while preserving tractability of handling and storing the representation of the sets. The technique can be utilized to improve simple interval-box approximations that provide coarse but computationally efficient enclosures, which can significantly aid in solving the main problems in interval programming.

The study of convex enclosures for the nonconvex sets encountered in interval programming provides substantial potential for further research. Promising research directions include improving McCormick envelope tightness through partitioning strategies, which break the domain into smaller regions to reduce relaxation gaps. Additionally, another promising line of research is the exploration of alternative convex polyhedra for enclosures, such as zonotopes, which naturally arise in affine interval arithmetic and have an efficient representation that could prove useful for interval computations.

Acknowledgements

The authors were supported by the Czech Science Foundation under grant 23-07270S.

References

Chaiyakan S, Thipwiwatpotjana P (2021) Bounds on mean absolute deviation portfolios under interval-valued expected future asset returns. Computational Management Science 18(2):195–212. https://doi.org/10.1007/s10287-021-00392-x

Garajová E, Hladík M (2019) On the optimal solution set in interval linear programming. Computational Optimization and Applications 72(1):269–292. https://doi.org/10.1007/s10589-018-0029-8

Garajová E, Rada M (2023) Interval transportation problem: Feasibility, optimality and the worst optimal value. Central European Journal of Operations Research 31(3):769–790. https://doi.org/10.1007/s10100-023-00841-9

Garajová E, Hladík M, Rada M (2019) Interval linear programming under transformations: Optimal solutions and optimal value range. Central European Journal of Operations Research 27(3):601–614. https://doi.org/10.1007/s10100-018-0580-5

Gerlach W (1981) Zur lösung linearer ungleichungssysteme bei störimg der rechten seite und der koeffizientenmatrix. Mathematische Operationsforschung und Statistik Series Optimization 12(1):41–43. https://doi.org/10.1080/02331938108842705

Hladík M (2012) An interval linear programming contractor. In: Ramík J, Stavárek D (eds) Proceedings of the 30th International Conference on Mathematical Methods in Economics (MME 2012), Part I. Silesian University in Opava, School of Business Administration in Karviná, Karviná, Czech Republic, pp 284–289, https://csov.vse.cz/eng/publications

Hladík M (2025) Interval Linear Programming and Extensions, Springer Optimization and Its Applications, vol 208. Springer Nature Switzerland, Cham, https://doi.org/10.1007/978-3-031-85096-7

Jaulin L, Kieffer M, Didrit O, et al (2001) Interval Analysis. In: Jaulin L, Kieffer M, Didrit O, et al (eds) Applied Interval Analysis: With Examples in Parameter and State Estimation, Robust Control and Robotics. Springer, London, p 11–43, https://doi.org/10.1007/978-1-4471-0249-6_2

McCormick GP (1976) Computability of global solutions to factorable nonconvex programs: Part I — Convex underestimating problems. Mathematical Programming 10(1):147–175. https://doi.org/10.1007/BF01580665

Mishmast Nehi H, Ashayerinasab HA, Allahdadi M (2020) Solving methods for interval linear programming problem: A review and an improved method. Operational Research 20(3):1205–1229. https://doi.org/10.1007/s12351-018-0383-4

Mohammadi M, Gentili M (2021) The outcome range problem in interval linear programming. Computers & Operations Research 129:105160. https://doi.org/10.1016/j.cor.2020.105160

Oettli W, Prager W (1964) Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides. Numerische Mathematik 6(1):405–409. https://doi.org/10.1007/BF01386090

Rada M, Garajová E, Horáček J, et al (2023) New pruning tests for the branch-and-prune framework for interval parametric linear systems. Soft Computing 27(18):12897–12912. https://doi.org/10.1007/s00500-022-06971-7

Rohn J (2006) Interval linear programming. In: Fiedler M, Nedoma J, Ramík J, et al (eds) Linear Optimization Problems with Inexact Data. Springer US, Boston, MA, p 79–100, https://doi.org/10.1007/0-387-32698-7_3

Rohn J, Kreinovich V (1995) Computing Exact Componentwise Bounds on Solutions of Lineary Systems with Interval Data is NP-Hard. SIAM Journal on Matrix Analysis and Applications 16(2):415–420. https://doi.org/10.1137/S0895479893251198

Skalna I (2018) Linear Systems. In: Skalna I (ed) Parametric Interval Algebraic Systems. Studies in Computational Intelligence, Springer, Cham, p 85–98, https://doi.org/10.1007/978-3-319-75187-0_4

Received: 2025-06-08 Last revision: 2025-07-03 Accepted: 2025-07-07

On Discrete Dynamical System of New Keynesian Model

Pavel Pražák [1], Jiří Křenek [2]

Abstract. The paper analyses a simplified version of the New Keynesian macroeconomic model. The model comprises three fundamental equations: the IS curve, the forward-looking Phillips curve, and the Taylor rule. The focus of this study is the dynamic properties of the system, which can be expressed in the form of a linear difference equation of the second order or a linear system of difference equations. An analytical derivation of the matrix of the given discrete dynamical system is undertaken and the characteristic polynomial and eigenvalues of the system are investigated. The results obtained demonstrate that the central bank response rate to inflation exerts a crucial influence on the stability of the equilibrium. It has been observed that for higher values of the parameter, the roots of the characteristic polynomial may be complex, resulting in different dynamic properties of the model. The paper provides an overview of the conditions for possible solutions.

Keywords: New Keynesian model, Inflation, Interest rate, Difference equation, Stability, Eigenvalue analysis

JEL Classification: C60, E50
AMS Classification: 39A60

1 Introduction

The New Keynesian model is an important model of contemporary macroeconomic analysis, particularly in monetary economics, (Miao, 2014; Galí, 2015). Building upon traditional Keynesian ideas, it addresses criticisms raised by the New Classical school, particularly with regard to microeconomic foundations and rational expectations. A key feature of the model is its incorporation of nominal rigidities, such as sticky prices and wages. These rigidities help explain the short-term ineffectiveness of monetary policy. These rigidities are typically modelled using firm-staggered price-setting behaviour, often represented by the Calvo pricing mechanism. Another central element of the New Keynesian model is the intertemporal optimisation of households and firms. Households choose consumption and labour supply to maximise utility over time, while firms maximise profits subject to constraints relating to pricing frictions. Expectations play a crucial role in the model - economic agents form forward-looking expectations based on the information available to them. Consequently, the model often takes the form of a dynamic stochastic general equilibrium (DSGE) system.

This paper deals with a standard three-equation New Keynesian model with augmented rational expectations ($\mathbb{E}_t x_{t+1} = x_{t+1}$, where x_t is an arbitrary state variable), for more details (Miao, 2014; Novales et al, 2009). The model includes three core equations: the dynamic IS curve, which describes intertemporal consumption decisions; the New Keynesian Phillips curve, which captures inflation dynamics; and a monetary policy rule, often represented by a Taylor rule (Woodford, 2003; Sorger, 2015; Miao, 2014). The Taylor rule dictates how a central bank adjusts nominal interest rates in response to deviations in inflation and output.

The model provides a framework for evaluating the stability and uniqueness of economic equilibria, which are crucial for effective policy implementation. The present text contributes to the study on the dynamic properties of a linearised New Keynesian system. Specifically, it investigates the mathematical behaviour of the model expressed as a second-order linear difference equation or a system of difference equations. By analytically deriving the matrix representation of the discrete dynamical system, the study examines its characteristic polynomial and eigenvalues. Understanding these dynamics is critical for both theoretical and practical reasons. From a theoretical perspective, the study improves our understanding of the structure of the model and possible outcomes. Practically, it informs policymakers about the consequences of aggressive or conservative interest rate responses to inflation.

The motivation for the paper stems from the need to better understand how specific parameters influence macroeconomic stability within the New Keynesian framework. Although much of the literature has focused on simulation-based approaches or empirical estimations, this study provides an analytical perspective. Clarifies the underlying

[1] University of Hradec Králové, Faculty of Informatics and Management, Department of Informatics and Quantitative Methods, Rokitanského 62, Hradec Králové, Czech Republic, pavel.prazak@uhk.cz
[2] University of Hradec Králové, Faculty of Informatics and Management, Department of Informatics and Quantitative Methods, Rokitanského 62, Hradec Králové, Czech Republic, ji.krene@gmail.com

mathematical mechanisms that govern the evolution of the system and helps to identify thresholds beyond which the implications of the policy may change.

2 Model

The standard New Keynesian model comprises three equations: the New Keynesian IS curve (or consumption Euler equation), the New Keynesian Phillips curve, and the monetary policy rule (or Taylor rule); more details can be found in (Woodford, 2003; Galí, 2015) and (Galí, 2011). The deterministic version of the model is a forward-looking macroeconomic model that explains the dynamics of short-run output and inflation in an economy without uncertainty. It is important to note that this model is only applicable to very short time periods.

First, the consumption equation is considered. This is a description of how households choose to consume over time.

$$y_t = \mathbb{E}_t y_{t+1} - \frac{1}{\sigma}\left(i_t - \mathbb{E}_t \pi_{t+1} - \bar{r}\right), \tag{1}$$

where y_t, i_t and π_{t+1} denote output gap, the nominal interest rate in period t and the rate of inflation in the next time period $t+1$. Moreover, $\bar{r}$ is a natural real interest rate and σ, $\sigma > 0$, denotes intertemporal elasticity of substitution and $\mathbb{E}_t$ is operator of expectation in period t. The key insight of equation (1) is that when real interest increases (that is, $i_t - \pi_{t+1}$ increases), the householder saves more, leading to a decrease in consumption and output. The output gap y_t in period t is considered as

$$y_t = \log Y_t - \log \overline{Y}_t,$$

where Y_t is the actual real output and $\overline{Y}_t$ is natural (or potential) output. The New Keynesian model is built by log-linearised nonlinear equilibrium conditions around a steady state, which makes the output gap naturally expressed in log deviations. When the deviation is small, the log difference is approximately the percentage deviation, i.e.

$$y_t = \log Y_t - \log \overline{Y}_t = \log \frac{Y_t}{\overline{Y}_t} \approx \frac{Y_t}{\overline{Y}_t} - 1 = \frac{Y_t - \overline{Y}_t}{\overline{Y}_t},$$

for more details (Pražák, 2024; Pražák, 2017). Similarly, we can deal with inflation. If P_t denotes the price level in period t, then

$$\pi_{t+1} = \log P_{t+1} - \log P_t = \log \frac{P_{t+1}}{P_t} \approx \frac{P_{t+1}}{P_t} - 1 = \frac{P_{t+1} - P_t}{P_t}.$$

The forward looking Phillips curve is the second relationship to be introduced

$$\pi_t = \beta \mathbb{E}_t \pi_{t+1} + \kappa y_t + u_t, \tag{2}$$

where β, $0 < \beta < 1$, and κ, $\kappa > 0$, are the discount factor and the sensitivity of inflation to the output gap, respectively. The variable u_t denotes an exogeneous cost shock in the period t. The key insight of equation (2) is that inflation increases when firms expect future inflation or when demand (output gap) is high.

The last equation included in the model is the monetary policy rule or the Taylor rule given by

$$i_t = \bar{r} + \phi_\pi \pi_t + \phi_y y_t \tag{3}$$

where feedback coefficients ϕ_y, $\phi_y > 0$, and ϕ_π, $\phi_\pi > 0$, measure how strongly the central bank responds to changes in the rate of inflation and the output gap in order to stabilise inflation - in addition, it is possible to consider an aggressive politics of central bank $\phi_\pi > 1$.

3 Results and Discussion

The given model (1)-(3) is a dynamic model with rational expectations, which usually requires a forward-looking solution. This type of solution is explained briefly in the Appendix below. There are two ways in which the model can be transformed. These reductions are described in this section.

3.1 Matrix Form of the Model

The interest rate i_t from (3) can be substituted into the IS curve (1), and the assumption of rational expectation can be used. It means that the system (1)-(3) can be then reduced to the following system of two first-order difference

equations

$$\frac{1}{\sigma}\pi_{t+1} + y_{t+1} = \frac{\phi_\pi}{\sigma}\pi_t + \left(1 + \frac{\phi_y}{\sigma}\right) y_t,$$
$$\beta\pi_{t+1} = \pi_t - \kappa y_t - u_t.$$

We will rewrite this system in matrix form rather than as a second-order equation. Let

$$X_t = \begin{pmatrix} \pi_t \\ y_t \end{pmatrix}$$

and

$$A = \begin{pmatrix} 1/\sigma & 1 \\ \beta & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \phi_\pi/\sigma & 1 + \phi_y/\sigma \\ 1 & -\kappa \end{pmatrix}, \quad Z_t = \begin{pmatrix} 0 \\ -u_t \end{pmatrix}$$

then the above given system can be shortly written as

$$AX_{t+1} = BX_t + Z_t.$$

Because

$$A^{-1} = 1/\beta \begin{pmatrix} 0 & 1 \\ \beta & -1/\sigma \end{pmatrix} \tag{4}$$

the system (1)-(3) can be finally written as

$$X_{t+1} = MX_t + U_t \tag{5}$$

where

$$M = A^{-1}B = 1/\beta \begin{pmatrix} 1 & -\kappa \\ (\beta\phi_\pi - 1)/\sigma & (\beta\phi_y + \beta\sigma + \kappa)/\sigma \end{pmatrix}$$

and

$$U_t = A^{-1}Z_t = \begin{pmatrix} -u_t/\beta \\ u_t/(\beta\sigma) \end{pmatrix}.$$

The behaviour of the solution to the system (5) depends on the eigenvalues of the matrix M, (Zhang, 2006; de la Fuente, 2000; Kelley and Peterson, 2001) and the eigenvalues depend on the values of the parameters. The characteristic polynomial $P(\lambda)$, (de la Fuente, 2000), of matrix M with eigenvalue λ is

$$P(\lambda) = \det(\lambda I - M) = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - \mathrm{tr}(M)\lambda + \det(M),$$

where I is the unit matrix of order two, λ_1, λ_2 are the corresponding eigenvalues and

$$\mathrm{tr}(M) = \lambda_1 + \lambda_2 = \frac{1}{\beta}\left(1 + \frac{\kappa}{\sigma}\right) + 1 + \frac{\phi_y}{\sigma},$$

$$\det(M) = \lambda_1 \cdot \lambda_2 = \frac{1}{\beta}\left(1 + \frac{\phi_y}{\sigma} + \kappa\frac{\phi_\pi}{\sigma}\right).$$

The Viète relations for quadratic equations were also used in the latter case. Since $1/\beta > 1$, it can be observed that

$$\mathrm{tr}(M) > 2 \text{ and } \det(M) > 1.$$

Due to these properties, it is not possible to unambiguously decide what sign the discriminant

$$D = \mathrm{tr}(M)^2 - 4\det(M) = \left(\frac{1}{\beta} + \frac{\kappa}{\beta\sigma} + 1 + \frac{\phi_y}{\sigma}\right)^2 - 4 \cdot \frac{\phi_y + \sigma + \kappa\phi_\pi}{\beta\sigma}$$

of the characteristic equation has. It is assumed that $\phi_\pi > 1$, and thus it can be observed that

$$P(1) = 1 - \mathrm{tr}(M) + \det(M) = \frac{\phi_y}{\sigma}\left(\frac{1}{\beta} - 1\right) + \frac{\kappa}{\beta\sigma}(1 - \phi_\pi) > 0. \tag{6}$$

As was shown, the roots of the characteristic polynomial $P(\lambda)$ can be complex, depending on the specific values of the parameters and the discriminant D. Therefore, it is necessary to distinguish three cases. The situation in which $D = 0$ is highly unlikely, so we will omit it.

- If $D < 0$, which means that $\phi_\pi > 1$ is high enough, then the characteristic equation $P(\lambda) = 0$ has two complex conjugate roots $\lambda_{1,2} = a \pm ib$. It means that $\det(M) = \lambda_1 \cdot \lambda_2 = a^2 + b^2 > 1$. Thus $|\lambda_1| = |\lambda_2| > 1$ and the roots of the characteristic equation are located outside the unit circle.
- If $D > 0$, which means that $\phi_\pi > 1$ is small enough, then both eigenvalues are real. Using the observation (6) both roots are greater than or smaller than one, see also Fig. 1. Since $\mathrm{tr}(M) = \lambda_1 + \lambda_2 > 2$, they cannot be smaller than one.

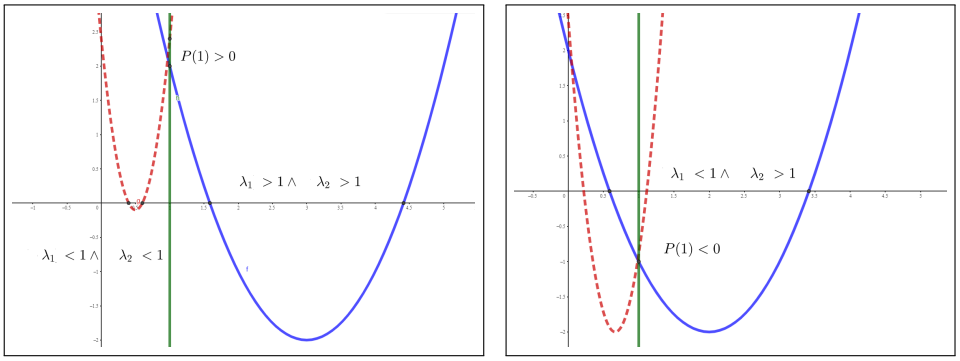


Figure 1 Characteristic polynomial $P(\lambda)$. The left panel shows the situation when $P(1) > 0$. On the right panel, there is an illustration of the situation when $P(1) < 0$.

We therefore reach the same conclusion in both cases - the absolute value of eigenvalues is greater than one. The forward solution to the system (5) can be unbounded, which means that backward solution to this system is better to use, (Kelley and Peterson, 2001; Elaydi, 2005). Consider that $M = P\Lambda P^{-1}$ is Jordan decomposition of matrix M, where Λ is the Jordan matrix of M with two different eigenvalues λ_1 and λ_2 and P is the matrix with columns given by eigenvectors of corresponding eigenvalues. Now, the backward solution to system (5) can be written as follows

$$X_t = P\Lambda^t P^{-1} c - \sum_{j=0}^{\infty} P\Lambda^{-(j+1)} P^{-1} U_{t+j},$$

where the vector c can be given by a boundary value. If $c \neq 0$ then $P\Lambda^t P^{-1} c$ represents the explosive part of the solution. To avoid this situation, it is necessary to put $c = 0$. Now, the solution to (5) can be written in the form

$$X_t = \sum_{j=0}^{\infty} P \begin{pmatrix} \lambda_1^{-(j+1)} & 0 \\ 0 & \lambda_2^{-(j+1)} \end{pmatrix} P^{-1} \begin{pmatrix} u_{t+j}/\beta \\ -u_{t+j}/(\beta\sigma) \end{pmatrix}, \quad \in \mathbb{N}, \tag{7}$$

that represents the impact of all expected future shocks.

3.2 Interest Rate as Exogenous Variable

In this section, we will assume that the nominal interest rate is exogenous, so that the relation (3) can be considered redundant (for example, the base interest rate in the Czech Republic remained at the same level from June 2022 to December 2023). The assumption of rational expectations is also used here. The system (1), (2) can be written as

$$\frac{1}{\sigma}\pi_{t+1} + y_{t+1} = y_t + \frac{r_t}{\sigma},$$
$$\beta\pi_{t+1} = \pi_t - \kappa y_t - u_t,$$

where $r_t = i_t - \overline{r}$. Let

$$X_t = \begin{pmatrix} \pi_t \\ y_t \end{pmatrix}$$

and

$$A = \begin{pmatrix} 1/\sigma & 1 \\ \beta & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 \\ 1 & -\kappa \end{pmatrix}, \quad Z_t = \begin{pmatrix} r_t/\sigma \\ -u_t \end{pmatrix}$$

then the above given system can be shortly written as

$$AX_{t+1} = CX_t + Z_t.$$

Because A^{-1} is given by (4) the system (1)-(3) can be finally written as

$$X_{t+1} = NX_t + V_t \tag{8}$$

where

$$N = A^{-1}C = 1/\beta \begin{pmatrix} 1 & -\kappa \\ -1/\sigma & (\beta\sigma + \kappa)/\sigma \end{pmatrix}$$

and

$$V_t = A^{-1}Z_t = \begin{pmatrix} -u_t/\beta \\ r_t/\sigma + u_t/(\beta\sigma) \end{pmatrix}.$$

The behavior of the solution to the system (8) depends on the eigenvalues of the matrix N. The characteristic polynomial $P(\lambda)$ of matrix N with eigenvalue λ is

$$P(\lambda) = \det(\lambda I - N) = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - \mathrm{tr}(N)\lambda + \det(N),$$

where I is the unit matrix of order two, λ_1, λ_2 are corresponding eigenvalues and

$$\mathrm{tr}(N) = \lambda_1 + \lambda_2 = \frac{1}{\beta}\left(1 + \frac{\beta\sigma + \kappa}{\sigma}\right) = 1 + \frac{1}{\beta} + \frac{\kappa}{\beta\sigma},$$

$$\det(N) = \lambda_1 \cdot \lambda_2 = \frac{1}{\beta}.$$

Since $1/\beta > 1$, it can be observed that

$$\mathrm{tr}(N) > 2 \text{ and } \det(N) > 1.$$

Since $\det(N) = \lambda_1\lambda_2 > 1 > 0$, both eigenvalues have the same sign. The discriminant of the characteristic equation is

$$D = \mathrm{tr}(N)^2 - 4\det(N) = \left(1 - \frac{1}{\beta}\right)^2 + \frac{\kappa}{\beta\sigma}\left(2 + \frac{2}{\beta} + \frac{\kappa}{\beta\sigma}\right) > 0.$$

It means that both characteristic roots are real and different. Moreover, it can be observed that

$$P(1) = 1 - \mathrm{tr}(N) + \det(N) = -\frac{\kappa}{\beta\sigma} < 0, \tag{9}$$

which means that one eigenvalue is less than one and the second one is greater than one and both eigenvalues are positive, see Fig. 1. To summarize, $0 < \lambda_1 < 1 < \lambda_2$, which means that the equilibrium point is an unstable saddle. The present model contains two non-predetermined variables - inflation π_t and output gap y_t and only one eigenvalue with modulus greater than one was found. Using Blanchard-Kahn condition, (Miao, 2014), it means that the equilibrium of the model is indeterminate and it is not possible to find a unique solution, (Galí, 2011).

4 Conclusion

This paper examined the deterministic New Keynesian model with rational expectations under two scenarios: one where the nominal interest rate is set endogenously by a Taylor-type rule, and another where it is exogenously fixed. In both cases, the model was rewritten as a system of first-order difference equations in matrix form, allowing for a detailed analysis of its dynamic properties.

The results show that equilibrium stability crucially depends on the model's parameters, particularly the central bank's response to inflation. When the Taylor principle is met, the eigenvalues lie outside the unit circle, ensuring a unique forward-looking solution. However, when the interest rate is exogenous, the model exhibits saddle-path dynamics with an indeterminate equilibrium, violating the Blanchard-Kahn condition. These findings underscore the essential role of monetary policy in maintaining determinacy and stability in forward-looking models.

Appendix

Consider the system of first-order linear difference equations (5), where $X_t \in \mathbb{R}^2$ is a state vector, $M \in \mathbb{R}^{2\times 2}$ is a constant coefficient matrix, and $U_t \in \mathbb{R}^2$ is an exogenous input variable. We are interested in finding a forward-looking solution, i.e. expressing X_t as a function of future values of $U_{t+j},\ j \in \mathbb{N}$. This type of solution is typical if the eigenvalues of M lie outside the unit circle, the system is then unstable forward in time, and a stable solution must be derived by solving backward.

Assuming that M is invertible (i.e., $\det M \neq 0$), we can write the system recursively as:

$$X_t = M^{-1}X_{t+1} - M^{-1}U_t \tag{10}$$

Iterating this equation forward

$$X_t = M^{-1}X_{t+1} - M^{-1}b_t = M^{-2}X_{t+2} - M^{-2}U_{t+1} - M^{-1}U_t = M^{-3}X_{t+3} - M^{-3}U_{t+2} - M^{-2}U_{t+1} - M^{-1}U_t$$
$$= \cdots = M^{-T}X_{t+T} - \sum_{j=0}^{T-1} M^{-(j+1)}U_{t+j}$$

Assuming the transversality condition holds

$$\lim_{T\to\infty} M^{-T}x_{t+T} = 0,$$

we obtain the closed-form of forward-looking solution

$$X_t = -\sum_{j=0}^{\infty} M^{-(j+1)}U_{t+j} \tag{11}$$

The current state X_t depends on the entire future path of the exogenous input U_t.

Acknowledgements

The support of the 2025 Specific Research Project of the Faculty of Informatics and Management of the University of Hradec Králové is kindly acknowledged.

References

Elaydi S (2005) An Introduction to Difference Equations. Springer Science and Business Media, Inc., New York

de la Fuente A (2000) Mathematical Methods and Models for Economics. Cambridge University Press, Cambridge

Galí J (2011) Are central banks' projections meaningful? Journal of Monetary Economics 58:537–550

Galí J (2015) Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework. Princeton University Press, Princeton, New Jersey

Kelley WG, Peterson AC (2001) Difference Equations, An Introduction with Applications. Academic Press, New York

Miao J (2014) Economic Dynamics in Discrete Time. The MIT Press, Cambridge, Massachusetts

Novales A, Fernándes E, Ruiz J (2009) Economic Growth, Theory and Numerical Solution Method. Springer-Verlag, Berlin Heidelberg

Pražák P (2017) Dynamic Linear IS-LM Model. In: Hradec Economic Days, vol 7(1). University of Hradec Kralove, Hradec Kralove, pp 715–722

Pražák P (2024) Discrete-time Dynamic Exchange Rate Overshooting Model. In: The 42nd International Conference on Mathematical Methods in Economics. Conference Proceedings. Jan Evangelista Purkyne University, Usti nad Labem, pp 250–255

Sorger G (2015) Dynamic Economic Analysis, Deterministic Models in Discrete Time. Cambridge University Press, Cambridge

Woodford M (2003) Interest and Prices: Foundations of a Theory of Monetary Policy. Princeton University Press, Princeton, New Jersey

Zhang WB (2006) Discrete Dynamical Systems, Bifurcations and Chaos in Economics. Elsevier B.V., Amsterdam

Received: 2025-06-09 Last revision: 2025-07-22 Accepted: 2025-07-29

Portfolio selecting under vague preferences

Adam Borovička[1]

Abstract. The capital market is a dynamic, volatile space where a diverse range of investment opportunities meet a mix of different strategic approaches, emotional outcomes, temperaments. In such a varied environment, decision-making is potentially extremely complex. Not only the less experienced investor can be naturally hesitant, or shy about expressing the preferences or requirements for investment. Preferences about the shape of the portfolio, or its characteristics, are then expressed tentatively, with a certain amount of uncertainty, tolerance or reserve. Recent mathematical optimization models supporting investment decision-making do not allow preferences to be expressed in an imprecise form. To eliminate this deficiency, their integration into model structures via so-called one-sided triangular fuzzy numbers is proposed. The fuzzy model is convertible into a strict form due to the piecewise linear preference. The methodology is designed in an interactive form of STEM method so that an investor in an already difficult situation can form preferences gradually. The step-by-step optimization process allows to achieve a portfolio composition reflecting the investor's preferences, which enhances its credibility. The applicative power of the proposed methodological approach is demonstrated in the selecting an investment portfolio of sustainable mutual funds.

Keywords: mutual fund, portfolio, preference, vague

JEL Classification: C44, C61, G11
AMS Classification: 90B50, 90C29

1 Introduction

The capital market is a place of diversity, which is shaped by the breadth of investment instruments on offer, the variety of characters, investment approaches, and the mix of facts, unexpected events, moods and feelings. In such an environment, the investor sometimes naturally flounders, unsure of next steps. His preferences are thus uncertain, vague.

Portfolio theory is significantly extended by multi-objective optimization methods. The very pillar of mean-variance theory is a two-objective model, thus incorporating arguably the two most important criteria in investment decision making. However, these factors are not the only ones that shape investor preferences. Moreover, preferences, especially of the less experienced investor, naturally evolve. It is therefore more than appropriate to use multi-objective methods with continuous information that allow preferences to be adjusted interactively.

Despite this aspect, interactive methods are neglected. Often a mean-variance model modified by a measure of risk is used. In addition to semivariance (Bollestero, 1995), absolute negative variance (Konno and Yamazaki, 1991) or VaR (Lwin et al., 2017) are offered. The mix of criteria is expanded, e.g., turnover (Chen and Huang, 2009), liquidity (Liu and Zhang, 2013), transaction costs (Yoshimoto, 1996) or ESG (Fereydooni et al., 2024). Also, the concept of goal programming is widely used, e.g., by Greco et al. (2016). Interactive multi-objective programming methods started to be developed especially in the 1970s. Among the first were Geoffrion-Dayer-Feinberg, Zionts-Wallenius, or the STEM method (Benayoun et al., 1971). Many early development methods are not very user-friendly. The most common obstacle is the required hard-to-access information from the decision maker, especially in the form of strict value trade-offs, as ultimately demonstrated by the comprehensive work of Miettinen (1998). A more friendly application face is set by methods that use the principle of implicit substitution of values of monitored functions. The pioneer of these approaches is the STEM method. More recent interactive methods can be found in Zapounidis and Pardalos (2010) or in the comprehensive work of Xin et al. (2018).

Although the interactive procedure allows for the gradual formation of preferences, it does not reflect the limited possibilities of the investor in expressing them. Preferences can thus be vague and indeterminate. The acceptable deterioration of the values of the selected functions is then only approximate, with a certain degree of tolerance. Thus, the theoretical goal is to integrate vague preferences into the decision-making process. The user-friendly

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

STEM method is a suitable candidate for this extension. Reflection of vague, approximate preferences is proposed through so-called one-sided triangular fuzzy number inspired by Gupta and Bhattacharjee (2010). This proposed fuzzy extension requires partial modifications, in particular the standardization of the objective function values in light of the intended optimization process.

The practical applicability of the proposed concept is demonstrated through the process of making an investment portfolio from sustainable mutual funds, which are increasingly popular nowadays. The specifics of strict and vague preferences are thus empirically confronted, which shapes, hence fulfills the empirical goal of this paper.

2 Modified STEM with vague preferences

The STEM algorithm is presented in minimalistic detail. Subsequent partial modifications are presented with due focus on the integration of vague preferences into decision making.

2.1 STEM algorithm

Based on the original paper of Benayoun et al. (1971), the algorithm of the STEM method can be briefly described in the following few steps.

Step 1: Let J_{min}, or J_{max}, define the set containing indices of minimizing, or maximizing objective functions. $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ represents the vector of n variables. The weights of the objectives, denoted as w_j, $j \in J_{min} \cup J_{max}$, for the j-th function, are specifically calculated (e.g. using the scoring method).

Step 2: In order to find an initial solution, a model minimizing the maximum weighted deviation from the ideal value is defined, which can be linearized into the following form

$$\begin{aligned} &\min \quad DEV \\ &w_j\left(f_j(\mathbf{x}) - f_j^I\right) \le DEV \qquad j \in J_{min} \\ &w_j\left(f_j^I - f_j(\mathbf{x})\right) \le DEV \qquad j \in J_{max} \\ &\mathbf{x} \in X \end{aligned}, \tag{1}$$

where $f_j^I, j \in J_{min} \cup J_{max,}$ represents the ideal (optimal) value of the j-th (standardized) objective function on the set X containing all the attendant constraints of the problem. The solution of model (1), denoted $\mathbf{x^1}$, represents the initial solution (decision). If the values of all the objective functions are acceptable, a compromise solution is found. However, if the values of only some of the functions are unacceptable, another solution can be found just in the direction of the interactively forming preference (see Step 3).

Step 3: Let J_{min}^D, or J_{max}^D, define the set of indices of minimizing, or maximizing objective functions with the acceptable deterioration of their current value $f_j(\mathbf{x^1})$. The indices of the other minimizing or maximizing functions are arranged in the set J_{min}^E, or J_{max}^E. Then the following model is solved

$$\begin{aligned} &\min \quad DEV \\ &w_j\left(f_j(\mathbf{x}) - f_j^I\right) \le DEV \qquad j \in J_{min}^E \\ &w_j\left(f_j^I - f_j(\mathbf{x})\right) \le DEV \qquad j \in J_{max}^E \\ &f_j(\mathbf{x}) \le f_j(\mathbf{x^1}) \qquad j \in J_{min}^E \\ &f_j(\mathbf{x}) \ge f_j(\mathbf{x^1}) \qquad j \in J_{max}^E \\ &f_j(\mathbf{x}) \le f_j(\mathbf{x^1}) + \Delta f_j(\mathbf{x^1}) \qquad j \in J_{min}^D \\ &f_j(\mathbf{x}) \ge f_j(\mathbf{x^1}) - \Delta f_j(\mathbf{x^1}) \qquad j \in J_{max}^D \\ &\mathbf{x} \in X \end{aligned}, \tag{2}$$

where $\Delta f_j(\mathbf{x^1}),\ j \in J_{min}^D \cup J_{max}^D,$ reflects the acceptable deterioration of the value of the j-th objective function. The weights of these functions are reset. The weights of other objectives remain unchanged. The second solution is represented by a vector $\mathbf{x^2}$ with the values of the objective functions $f_j(\mathbf{x^2}),\ j \in J_{min} \cup J_{max}$. Other solutions are

evaluated according to gradually forming preferences. The interactive process ends with the acceptance of the (current) solution.

2.2 Algorithm modifications in the light of vague preferences

It can be relatively easy to project a preference into the selection of objective functions with an unacceptable value. Determining functions with potentially deteriorating values is a more complex process. Not to mention its concrete quantification. An important guide in forming a preference may be the knowledge of the basal and ideal value of the relevant characteristic, which under initial conditions delineate the interval of its movement. Yet the decision maker may not be able to quantify strictly what deterioration would be acceptable. In the end, he often does not want to make such a statement, not only because of a lack of erudition, but also because of emotional motives, intuition, and, consequently, in the spirit of a certain natural reserve.

This vague preference can be expressed by the following fuzzy constraint based on the partial formulation (2)

$$\begin{aligned} f_j(\mathbf{x}) &\tilde{\leq} f_j(\mathbf{x^1}) + \Delta f_j(\mathbf{x^1}) \qquad j \in J^D_{min} \\ f_j(\mathbf{x}) &\tilde{\geq} f_j(\mathbf{x^1}) - \Delta f_j(\mathbf{x^1}) \qquad j \in J^D_{max} \end{aligned}, \tag{3}$$

$\Delta f_j(\mathbf{x^1}),\ j \in J^D_{min} \cup J^D_{max},$ reflects the approximate acceptable deterioration of the value of the *j*-th objective function. However, this approximation is naturally bounded, through the tolerance $\Delta' f_j(\mathbf{x^1}),\ j \in J^D_{min} \cup J^D_{max}$. The deterioration of the value of the *j*-th selected objective function in the interval $\left\langle \Delta f_j(\mathbf{x^1}) + \varepsilon, \Delta f_j(\mathbf{x^1}) + \Delta' f_j(\mathbf{x^1}) \right\rangle,$ where $\varepsilon = 10^{-5},$ is acceptable, but with a lower degree of willingness. The degree of reluctance increases linearly as the value of the minimizing, or maximizing function increases or decreases, towards the extreme position of the difference $\Delta f_j(\mathbf{x^1}) + \Delta' f_j(\mathbf{x^1})$.

Vague preference can be incorporated, in the spirit of the inverse approach of Bellman's concept, through a so-called one-sided triangular fuzzy number that just reflects the natural linearity in human indecision. Potential change $\Delta f^1_j(\mathbf{x}) = f_j(\mathbf{x}) - f_j(\mathbf{x^1}),\ j \in J^D_{min},$ or $\Delta f^1_j(\mathbf{x}) = f_j(\mathbf{x^1}) - f_j(\mathbf{x}),\ j \in J^D_{max},$ of the *j*-th minimizing, or maximizing objective function is described through a so-called left-side triangular fuzzy number $\tilde{O}_{\Delta f^1_j(\mathbf{x})}$ with the following membership function

$$\mu_{\tilde{O}_{\Delta f^1_j(\mathbf{x})}}\left[f_j(\mathbf{x}) \right] = \begin{cases} \frac{\Delta f^1_j(\mathbf{x}) - \Delta f_j(\mathbf{x^1})}{\Delta' f_j(\mathbf{x^1})} & \Delta f_j(\mathbf{x^1}) \leq \Delta f^1_j(\mathbf{x}) \leq \Delta f_j(\mathbf{x^1}) + \Delta' f_j(\mathbf{x^1}) \\ 0 & \text{otherwise} \end{cases} \qquad j \in J^D_{min} \cup J^D_{max}. \tag{4}$$

The vague preference about the possible change, or deterioration of the value of the objective function can be visualized as follows (Figure 1).

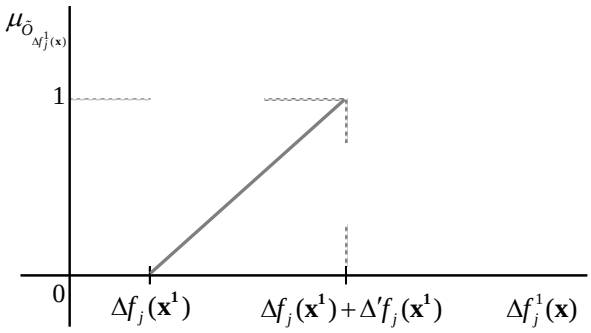


Figure 1 Vague preference as a so-called left-side triangular fuzzy number, Source: own processing

The vague preference can be integrated to the model (2) through the following constraint(s)

$$w_j \frac{\Delta f^1_j(\mathbf{x}) - \Delta f_j(\mathbf{x^1})}{\Delta' f_j(\mathbf{x^1})} \leq DEV \qquad j \in J^D_{min} \cup J^D_{max}. \tag{5}$$

These conditions (5) logically replace the conditions expressing strict preferences in the model (2). Thus, the following conditions are eliminated.

$$f_j(\mathbf{x}) \le f_j(\mathbf{x}^1) + \Delta f_j(\mathbf{x}^1) \qquad j \in J_{min}^D, \qquad f_j(\mathbf{x}) \ge f_j(\mathbf{x}^1) - \Delta f_j(\mathbf{x}^1) \qquad j \in J_{max}^D. \tag{6}$$

However, the adjusted model is now non-compact because it combines relative and absolute deviations. In addition, it exhibits an important limitation of the applicability of the concepts of model (1) and (2), not yet mentioned, namely the assumption of comparability of the values of the observed (objective) functions. If this is not met, the value of the "deviance" objective function of the models is misleading. In order to overcome this shortcoming, a standardization procedure (7) using both ideal and basal values is proposed. Then the following holds

$$\frac{f_j(\mathbf{x}) - f_j^I}{f_j^B - f_j^I} \in \langle 0,1 \rangle \quad j \in J_{min}, \qquad \frac{f_j^I - f_j(\mathbf{x})}{f_j^I - f_j^B} \in \langle 0,1 \rangle \quad j \in J_{max}, \tag{7}$$

where f_j^B, $j \in J_{min} \cup J_{max}$, is the basal value determined with respect to the ideal values of the other objectives. In the spirit of the proposed concepts (4) and (5), a value close to 0, or 1 indicates proximity to the ideal, or basal value. After standardization of the objectives and integration of vague preferences, model (1) and model (2) can be transformed into the following form

$$\begin{aligned} &\min \quad DEV \\ &w_j \frac{f_j(\mathbf{x}) - f_j^I}{f_j^B - f_j^I} \le DEV \qquad j \in J_{min} \\ &w_j \frac{f_j^I - f_j(\mathbf{x})}{f_j^I - f_j^B} \le DEV \qquad j \in J_{max} \\ &\mathbf{x} \in X \end{aligned} \tag{8},$$

$$\begin{aligned} &\min \quad DEV \\ &w_j \frac{f_j(\mathbf{x}) - f_j^I}{f_j^B - f_j^I} \le DEV \qquad j \in J_{min}^E \\ &w_j \frac{f_j^I - f_j(\mathbf{x})}{f_j^I - f_j^B} \le DEV \qquad j \in J_{max}^E \\ &f_j(\mathbf{x}) \le f_j(\mathbf{x}^1) \qquad j \in J_{min}^E \\ &f_j(\mathbf{x}) \ge f_j(\mathbf{x}^1) \qquad j \in J_{max}^E \\ &w_j \frac{\Delta f_j^1(\mathbf{x}) - \Delta f_j(\mathbf{x}^1)}{\Delta' f_j(\mathbf{x}^1)} \le DEV \qquad j \in J_{min}^D \cup J_{max}^D \\ &\mathbf{x} \in X \end{aligned} \tag{9}.$$

According to STEM algorithm, the process of the interactive procedure remains unchanged. It should be noted that the calculation of the weights will be adapted to the needs of the decision-making procedure when appropriate.

3 Portfolio selection of ESG funds under vague preferences

First, a sustainable investment strategy is specified. Suitable open-ended mutual funds with their typical characteristics are then presented. An investment portfolio is then constructed based on the precise and vague preferences expressed by the investor. The results are analysed.

3.1 Investment strategy, ESG funds and their characteristics

The empirical study seeks to demonstrate the effectiveness of integrating the expression of vague preference into the investment decision-making process. To approximate the investment reality, a balanced investment strategy will be chosen. This most common approach reflects rather risk averse behavior. It is usually a longer-term strategy in order to be financially secure in retirement or to provide financial support for events in early adulthood. The investor does not have much capital market experience. There is little overview of the range of investment instruments available. However, the investor is interested in broadening the investment horizons. Access to investment instruments should be easy, preferably through his "home" bank. Portfolio management is planned to be rather passive. In addition to the usual characteristics, the sustainability of the investment is also monitored.

The chosen investment strategy can be conveniently implemented through open-ended ESG[2] mutual funds. To maintain the typicality of the investor, ESG mutual funds offered by Česká spořitelna, which has over 4.5 million clients in the Czech Republic, are selected. There are ten established open-ended mutual funds in the sustainable fund investment category. These are two bond, five equity and three mixed funds. Return and risk are undoubtedly the most important investment characteristics. For mutual funds, fees can play a decisive role. While the ESG

[2] Environmental (E), social (S), and governance (G) is an acronym for an investment principle that prioritizes environmental, social, and corporate governance issues.

mutual fund group includes investments subject to ESG benchmarks, the degree to which they meet the ESG classification can vary.

The return is calculated as the average annualized over the historical 10-year period from January 2015 to December 2024. The period was judiciously chosen given the need to capture long-term trends. This period reflects both times of great fluctuation during the viral pandemic as well as calmer periods including, but not limited to, 2015/2016. Risk is measured by a standardized SRRI (Synthetic Risk and Reward Indicator) based on a statistical analysis of historical price trends (European Securities and Markets Authority, 2010). Costs include all current fees paid on a continuous basis (licensing, management, audit, etc.). The entry fee is not included because of its one-time nature and minimal variability across the selected funds. The ESG factor is measured through an internal ESGenius score (Erste Asset Management, 2025).

Risk and costs are downloaded from the Investment center (2025). Prices, on which the returns are calculated, are available in the same place. The current ESGenius score is provided in the fund's annual report, which is also available on the web of Investment center (2025). All the necessary data is compiled in Table 1.

Mutual fund	**Return** [%]	**Risk** [1-7]	**Cost** [%]	**ESG** [0-100]
Erste Bond Europe-High Yield	1.63	3	1.17	57
Erste Bond Euro Corporate	-0.14	3	0.71	62
Erste Stock Biotec	2.70	5	2.07	56
Erste Stock Global	9.77	4	2.07	61
Erste Responsible Stock Global	8.66	4	1.79	66
Erste Responsible Stock America	12.31	5	2.04	64
Erste WWF Stock Environment	7.86	5	1.77	83
ESG Mix 10	1.37	2	1.13	74
ESG Mix 30	1.82	2	1.52	71
ESG Mix 50	2.13	3	2.26	69

Table 1 ESG funds' data; Source: Investment center (2025), annual reports, own calculations

3.2 Preference portfolio forming

Naturally, the evaluation criteria are of varying importance to the investor. Return is the purpose of investing. Its importance must therefore be high. However, the balanced investor is cautious. He therefore also attaches high importance to risk. Costs are not so important to the investment. However, the sustainability factor plays a significant role alongside return and risk. The weights are determined by the user-friendly scoring method. Points from $\langle 1,10 \rangle$ are assigned to each criterion according to the preference specified above. Then the weights (score) are as follows: return – 0.31 (8), risk – 0.31 (8), cost – 0.12 (3), ESG – 0.27 (7). Although the investor does not plan to actively manage the portfolio, he wants to have a basic overview of the investment. It is therefore not desirable to have too many assets in the portfolio. Moreover, the share of one asset should not be dominant as well as negligible. Based on this diversification aspect, the lower, or upper limit for the fund's share is set at the level of 15%, or 45%.

To determine the initial portfolio, the following model will be formulated and subsequently solved.

$$
\begin{aligned}
&\min \quad DEV \\
&0.31\frac{\mathbf{s}^{\mathbf{T}}\mathbf{x}-2.15}{4.45-2.15} \le DEV, \quad 0.12\frac{\mathbf{c}^{\mathbf{T}}\mathbf{x}-0.98}{2.03-0.98} \le DEV \\
&0.31\frac{10.74-\mathbf{r}^{\mathbf{T}}\mathbf{x}}{10.74-0.79} \le DEV, \quad 0.27\frac{77.6-\mathbf{u}^{\mathbf{T}}\mathbf{x}}{77.6-61.8} \le DEV \\
&\mathbf{e}^{\mathbf{T}}\mathbf{x}=1 \\
&0.15\mathbf{y} \le \mathbf{x} \le 0.45\mathbf{y} \\
&\mathbf{x} \ge \mathbf{0} \\
&\mathbf{y} = \{0,1\}
\end{aligned}
\qquad (10)
$$

where $\mathbf{x}=(x_1,x_2,...,x_{10})^T$ is a vector of the variables expressing the share of the mutual funds (in the order listed in Table 1), $\mathbf{y}=(y_1,y_2,...,y_{12})^T$ is then a vector of binary variables involved in the limitation for the fund share,

$\mathbf{e}^{\mathbf{T}} = (1,1,...,1)$. The ten-component vectors **r**, **s**, **c** and **u** contain successively the levels of return, risk, cost and ESG factor for the funds considered. Then the functions $\mathbf{r}^{\mathbf{T}}\mathbf{x}$, $\mathbf{s}^{\mathbf{T}}\mathbf{x}$, $\mathbf{c}^{\mathbf{T}}\mathbf{x}$ and $\mathbf{u}^{\mathbf{T}}\mathbf{x}$ represent return, risk, cost and ESG factor of the portfolio. The ideal and basal values are determined through a single-objective optimization process declared by the STEM algorithm. The partial objective extremes are found on the set $X = \left\{\mathbf{e}^{\mathbf{T}}\mathbf{x} = 1,\ 0.15\mathbf{y} \le \mathbf{x} \le 0.45\mathbf{y}, \mathbf{x} \ge \mathbf{0},\ \mathbf{y} = \{0,1\}\right\}$. All models are solved in Lingo optimization software.

The portfolio has the following composition: 41.36% Erste Responsible Stock America, 29,73% ESG Mix 10 and 28.91% ESG Mix 30. Its characteristics are at the following levels: return – 6.02%, risk – 3.24, cost – 1.68%, ESG - 69. The high share of Erste Responsible Stock America is not surprising. The fund has the highest return by far. However, its higher riskiness must be offset by low-risk assets such as the mixed funds ESG Mix 10 and ESG Mix 30. The absence of other equity funds is mainly due to their high riskiness, which is not sufficiently compensated by a higher level of return and/or ESG factor. Bond funds have virtually no chance of participating in the portfolio due to the unfortunate combination of low return and ESG factor.

Imagine a situation where our more risk averse investor wants to reduce the level of risk. Furthermore, he would like to maintain the current level of ESG factor. To this end, the investor is willing to give up a part of return and hence costs. Let us distinguish two cases. In the first case, the investor is decisive, clearly accepting a deterioration of 0.3 or 0.22 percentage points in returns and costs respectively. In the second case, the investor is uncertain about his preferences. He is convinced that he is willing to give up 0.2 and 0.12 percentage points of return and cost respectively. However, he is hesitant to accept a possibly larger deterioration in the levels of these criteria. Thus, he specifies a tolerance zone of 0.1 percentage points for both characteristics. Compared to the first case, the degree of willingness to change decreases within this zone. For strictly and vaguely expressed preferences, the revised model (8) and (9) will be applied in the spirit of the processes proposed above. The revised portfolios and their characteristics can be shown in Table 2.

Mutual fund	**Share** *Strict*	**Share** *Vague*	**Criteria**	**Level** *Strict*	**Level** *Vague*
Erste Responsible Stock America	39.04%	39.46%	Return	5.72%	5.78%
ESG Mix 10	42.61%	40.25%	Risk	3.17	3.18
ESG Mix 30	18.35%	20.29%	Cost	1,61%	1.62%
			ESG	69.55	69.45

Table 2 Revised portfolios and their characteristics, Source: own calculations

The requested changes are not significant, so no change in funds is expected. Risk mitigation is provided by the lower equity fund share. ESG Mix 10 strengthens its position at the expense of ESG Mix 30 mainly due to the higher ESG. The cost preference does not play a major role as equity funds with lower risk (return) usually carry a lower cost burden. As expected, in the case of precise preference, the return was reduced by a maximum of 0.3 percentage points, while in the case of vague preference the tolerance "cushion" was only partially used in order to simultaneously ensure adequate relative positions of all investment characteristics. Depending on the preference requirements, the interactive process may continue in a similar vein until the portfolio composition is accepted.

4 Conclusion

In a capital market full of uncertainties, investors are often unable or unwilling to express their preferences in a precise manner. However, multi-objective optimization methods do not capture this aspect. Thus, this paper focuses on integrating vague preferences into the portfolio selection process. For this purpose, a suitable interactive multi-objective method is selected, which is appropriately extended to account for vague information. Empirical experience with ESG mutual funds confirms the appropriateness of this step. The tolerance band of the change in deterioration of the selected objective functions, integrated via a so-called one-sided triangular fuzzy number into the STEM algorithm, allows for a "finer" search for an acceptable solution via simultaneous multi-objective optimization of the tolerance level. The STEM method becomes a more versatile tool for supporting investment decision-making.

This research opens up partial solitary theoretical and empirical questions. Further research could focus on studying the linearity of preferences, or possibly recalculating the weights in an interactive procedure. Of course, the proposed methodology would deserve a more detailed application (not only) in an investment case study in order to highlight its applicability in the nowadays so common uncertain conditions.

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

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

Benayoun R, de Montgolfier J, Tergny J, Metra S, Laritchev O I (1971). Linear programming with multiple objective functions: STEP Method (STEM). Mathematical programming 1:366–375. https://doi.org/10.1007/BF01584098

Chen L H, Huang L (2009) Portfolio optimization of equity mutual funds with fuzzy return rates and risks. Expert Systems with Applications 36:3720–3727. https://doi.org/10.1016/j.eswa.2008.02.027

Committee of European Securities Regulators (2010) CESR's guidelines on the methodology for the calculation of the synthetic risk and reward indicator. https://www.esma.europa.eu/sites/default/files/library/2015/11/10_673.pdf

Erste Asset Management (2025). ESGenius® Analysis. Available at: https://www.erste-am.at/en/private-investors/our-solutions/fund-glossary/esgenius-analyse [cited 2025-06-01]

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

Greco S, Ehrgott M, Figueira J R (2016) Multiple Criteria Decision Analysis: State of the Art Surveys. Springer, New York

Gupta M, Bhattacharjee D (2010) Multi objective problem in fuzzy environment where resources are triangular fuzzy number. European Journal of Scientific Research 46:99-106

Investment center (2025) Responsible investment. Available at: https://www.investicnicentrum.cz/en/products/esg-products?issuers=21915489&issuers=21915496&issuers=21915490&esg=true [cited 2025-06-01]

Miettinen K (1998) Nonlinear Multiobjective Optimization. Springer, New York

Konno H, Yamazaki H (1991) Mean-absolute deviation portfolio optimization model and its applications to the Tokyo stock market. Management Science 37:519–531. https://doi.org/10.1287/mnsc.37.5.519

Liu Y-J, Zhang W-G (2013) Fuzzy portfolio optimization model under real constraints. Insurance: Mathematics and Economics: 53:704–711. https://doi.org/10.1016/j.insmatheco.2013.09.005

Lwin K T, Qu R, MacCarthy B L (2017) Mean-VaR portfolio optimization: A nonparametric approach. European Journal of Operational Research 260:751–766. https://doi.org/10.1016/j.ejor.2017.01.005

Xin B, Chen L, Chen J, Ishibuchi H, Hirota K, Liu, B (2018) Interactive multiobjective optimization: a review of the state-of-the-art. IEEE Access, 6:41256-41279. https://doi.org/10.1109/ACCESS.2018.2856832

Yoshimoto A (1996) The mean-variance approach to portfolio optimization subject to transaction costs. Journal of the Operations Research Society of Japan 39:99–117. https://doi.org/10.15807/jorsj.39.99

Zopounidis C, Pardalos P M (2010) Handbook of Multicriteria Analysis. Springer, Berlin

Received: 2025-06-05 Last revision: 2025-06-05 Accepted: 2025-07-07

Minimizing elevation difference in route construction: a geospatial and computational analysis

Jindra Lacko [1], Michal Černý [2]

Abstract. Our study examines a class of routing problems involving minimizing of differences in terrain altitude; a class of problems encountered in traffic use cases, such as railroad construction. We present a graph-based approach to this problem, leveraging a square grid derived using a real world altimetry. We discuss two classes of undesirable behavior of a grid based approach on synthetic data, and a workable solution of the problem using a real world dataset.

Keywords: routing, spatial analysis, graph networks, computational complexity

JEL Classification: C44
AMS Classification: 90C15

1 Introduction and problem formulation

1.1 High-level problem description.

This study contributes to the problem of designing ''good'' paths in complex terrains (landscapes), where construction of paths is costly, such as mountains with high passes and deep valleys. To describe the landscape formally, we assume to be given a smooth function $h : \mathbb{R}^2 \to \mathbb{R}$ on a compact domain D. We interpret (x, y) as spatial coordinates (e.g. GPS location) and $h(x, y)$ as the terrain height (e.g. meters above the sea level). Let two spatial points $(x_{\text{START}}, y_{\text{START}}), (x_{\text{END}}, y_{\text{END}}) \in D$ be given. The general task is to find a smooth curve $(x(t), y(t))_{t\in[0,1]}$ such that $(x(0), y(0)) = (x_{\text{START}}, y_{\text{START}})$, $(x(1), y(1)) = (x_{\text{END}}, y_{\text{END}})$, $(x(t), y(t)) \in D$ for every t, and such that the resulting *path* — the curve $(x(t), y(t), h(x(t), y(t)))_{t\in[0,1]}$ in $\mathbb{R}^3$ — has some good properties to be discussed below.

1.2 The criterion of quality of a path.

In practical problems, we usually meet various criteria for the quality of the path, such as length, elevation changes, or curvature. In this text, the elevation difference (or *slope*) of the path plays the main role. We say that a path is η-*feasible* if its slope does not exceed η. Formally, we require the *slope* $h^*(t) := \lim_{\Delta\to 0}(h(x(t+\Delta), y(t+\Delta)) - h(x(t), y(t)))/\sqrt{(x(t+\Delta) - x(t))^2 + (y(t+\Delta) - y(t))^2}$ to satisfy $|h^*(t)| \leq \eta$ for every t. Among all η-feasible paths we are interested in finding the shortest one, i.e., the one minimizing $\int_0^1 \sqrt{(\mathrm{d}x(t)/\mathrm{d}t)^2 + (\mathrm{d}y(t)/\mathrm{d}t)^2}\,\mathrm{d}t$. This is not the only possible criterion; a particular choice of the criterion to choose the ''best'' path from the set of all η-feasible paths depends on a particular application. For example, a natural alternative of the path length could be minimizing $\int_0^1 \sqrt{(\mathrm{d}x(t)/\mathrm{d}t)^2 + (\mathrm{d}y(t)/\mathrm{d}t)^2 + (\mathrm{d}h(x(t), y(t))/\mathrm{d}t)^2}\,\mathrm{d}t$ instead. Another example is minimization of overall elevation changes $\int_0^1 |h^*(t)|\,\mathrm{d}t$, preventing the path from being bumpy.

Other possible criteria include the financial cost of road construction (Vázquez-Méndez et al., 2018). Since a single objective is not always practical a bi-objective optimization framework using Pareto front can be applied (Hirpa et al., 2016). A two step approach can be applied by generating various promising paths in the first phase and refining them into final alignments by fitting the required curved sections (Li and Pu, 2016).

1.3 Discretization.

Our task is solvable by discretization. For simplicity assume that the domain D is a box of the form $D = X \times Y$, where $X = [\underline{x}, \overline{x}]$ and $Y = [\underline{y}, \overline{y}]$ are intervals. Let us discretize such domain D it into mn equal-sized boxes $B_{ij} := [\underline{x} + (i-1)\Delta_x, \underline{x} + i\Delta_x] \times [\underline{y} + (j-1)\Delta_y, \underline{y} + j\Delta_y]$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, where $\Delta_x := (\overline{x} - \underline{x})/n$ and $\Delta_y := (\overline{y} - \underline{y})/m$. A box B_{ij} is associated with average terrain height $h_{ij} := (\Delta_x\Delta_y)^{-1} \int_{B_{ij}} h(x, y)\,\mathrm{d}x\mathrm{d}y$ (or, for simplicity, just using the midpoint value $h_{ij} := h(\underline{x} + (i - \frac{1}{2})\Delta_x, \underline{y} + (j - \frac{1}{2})\Delta_y)$.)

[1] Department of Econometrics, Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, Czechia, jindra.lacko@vse.cz
[2] Department of Econometrics, Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, Czechia, cernym@vse.cz

1.4 Graph G.

Let $G(V, E)$ be an undirected graph with vertices (i, j), such that an edge $\{(i, j), (i', j')\}$ is present in E iff boxes $B_{ij}, B_{i'j'}$ are adjacent (i.e., if $B_{ij} \cap B_{i'j'}$ is a line segment) and the terrain slope between them does not exceed the threshold η. The slope condition can be stated more formally: $|h_{ij} - h_{i'j'}|/\Delta_x \leq \eta$ in case $j = j'$, and $|h_{ij} - h_{i'j'}|/\Delta_y \leq \eta$ in case $i = i'$.

1.5 How to measure the distance.

For two adjacent boxes $B_{ij}, B_{i'j'}$, the edge $\{(i, j), (i', j')\}$ of G is associated with the distance either Δ_x if $j = j'$ and $i' = i \pm 1$, or Δ_y if $i = i'$ and $j' = j \pm 1$. An alternative natural approach would assign distances $\sqrt{\Delta_x^2 + (h_{ij} - h_{i'j'})^2}$ in the former case and $\sqrt{\Delta_y^2 + (h_{ij} - h_{i'j'})^2}$ in the latter.

1.6 Remark: Is discretization necessary?

The continuous formulation of the problem—to find the optimal (= shortest) path satisfying the η-feasibility condition—amounts to finding the ''optimal'' smooth functions $x(t), y(t)$. However, even if they exist, there is no theoretical guarantee that they are elementary, i.e., that they can be described by a standard formula. Hence, a natural question arises: *how to describe $x(t), y(t)$ once they do not have a closed-form formula?* The discretization is a natural approach; we can choose $\Delta_x, \Delta_y \searrow 0$ and get a tabelation of $x(t), y(t)$ with arbitrarily high precision.

1.7 Some algorithmic tasks.

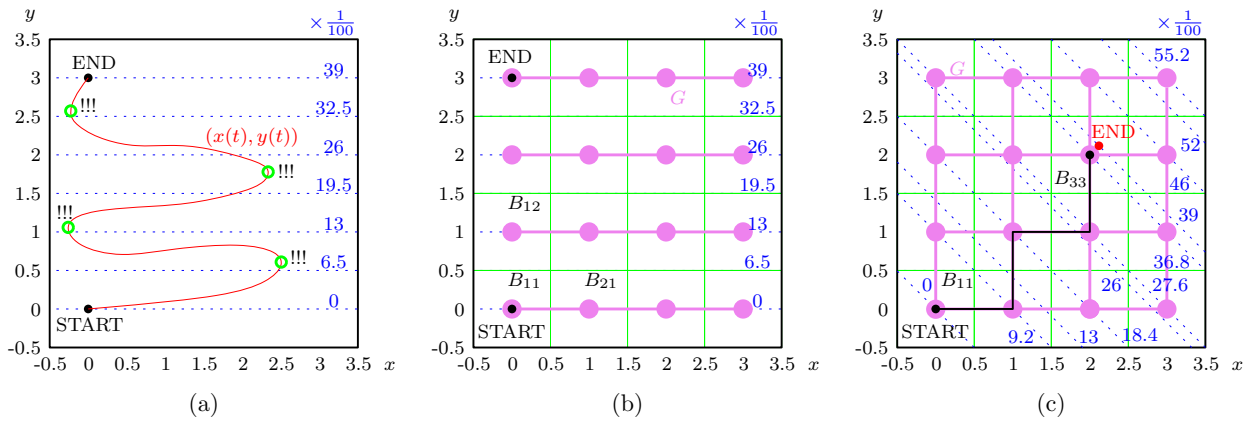
If the discretization of the domain $D = X \times Y$ into boxes B_{ij} is dense enough, the discretized problem can be seen as an approximation to the outlined problem with continuous formulation. Using the representation by graph G, it is easy to answer some straightforward questions. For example: *given (x_{START}, y_{START}), (x_{END}, y_{END}) and the slope threshold $\eta \geq 0$, does there exist an η-feasible path?* (It suffices to test if boxes $B^0 \ni (x_{\mathrm{START}}, y_{\mathrm{START}})$ and $B^1 \ni (x_{\mathrm{END}}, y_{\mathrm{END}})$ are in the same component of G.) *Finding the shortest η-feasible path* is also easy: it is reducible to Dijkstra's algorithm.

1.8 The main contribution.

The main goal of this text is to show some strengths and weaknesses of the discretization method. In particular: the main weakness is that the discretization method can completely fail, even in a very simple setup, as illustrated in Sections 2, 3. On the other hand, in cases when it works fine, the main advantage is computational speed: as the graph G is sparse (in particular, it is easily seen that $|E| \leq 2|V| = 2mn$), we get complexity $O(|V| \log |V|)$ for Dijkstra's method, making the method easy-to-use even for very dense discretization grids.

2 A failure of the discretization method: The rotation effect

Let us consider $\eta = 0.12$; this is what we know from real roads, that the slope is often allowed to be at most 12% (both upwards and downwards). Consider a very simple terrain described by the linear function $h(x, y) = 0.13y$, and let $(x_{\mathrm{START}}, y_{\mathrm{START}}) = (0, 0)$, $(x_{\mathrm{END}}, y_{\mathrm{END}}) = (0, 3)$. This situation is depicted in Figure 1(a). The dotted blue lines depict some contour lines of the terrain $h(x, y)$.



(a) (b) (c)

Figure 1: A sample instance illustrating Section 2.

Plot (a) of Figure 1 shows that there exists *no* 0.12-feasible path: as we insist on smoothness of the curve $(x(t), y(t))$, it is easily seen that every candidate smooth curve between $(0, 0)$ and $(0, 3)$ has a point t_0 with slope $h^*(t_0) = 0.13$; see the green points denoted by triple exclamation marks.

Plot (b) shows an example of the discretization grid with $\Delta_x = \Delta_y = 1$. The starting point is in box B_{11}, the endpoint is in box B_{14}. Graph G is depicted in purple: it has four components and, obviously, in G there exists no path from B_{11} to B_{14}. For example, it is easily seen that $h_{i1} = 0$ for all i, $h_{i2} = 0.13$ for all i, and thus the slope condition does not allow the addition of edges of the form $\{(i, 1), (i, 2)\}$. This property of graph G remains true even if we make the grid more dense. Thus, in this case, the answer ''no feasible path exists in G'' is correct.

However, the situation changes when we rotate the entire scene by $\pi/4$, see plot (c) of Figure 1. We would expect that rotation does not change the terrain (it just changes the angle how a camera observes the terrain) — and thus the answer of the method should be the same. The bad news is that this intuition is false.

Indeed, after the $(\pi/4)$-rotation we have $h(x, y) = 2^{-1/2}x + 2^{-1/2}y$, $(x_{\text{START}}, y_{\text{START}}) = (0, 0)$, $(x_{\text{END}}, y_{\text{END}}) = (3 \cdot 2^{-1/2}, 3 \cdot 2^{-1/2}) \approx (2.1, 2.1)$. Now $h_{11} = 0$ and $h_{12} = h_{21} = 0.13 \cdot 2^{-1/2} \approx 0.092$; thus the difference of elevations is below the threshold $\eta = 0.12$ and both edges $\{(1, 1), (1, 2)\}$ and $\{(1, 1), (2, 1)\}$ are present in G. By a similar argument it follows that G contains all possible edges. Thus there exists a path between B_{11} and $B_{33} \ni (x_{\text{END}}, y_{\text{END}})$. One example of such path is shown in Figure 1(c) in black.

Even worse, this undesirable property remains true even if we make the grid more dense. This is the case where the method fails: *by making the grid more and more dense, the method constructs a sequence of finer and finer ''approximations'' of a continuous path between* (x_{START}, y_{START}), (x_{END}, y_{END}) *which does not exist at all.*

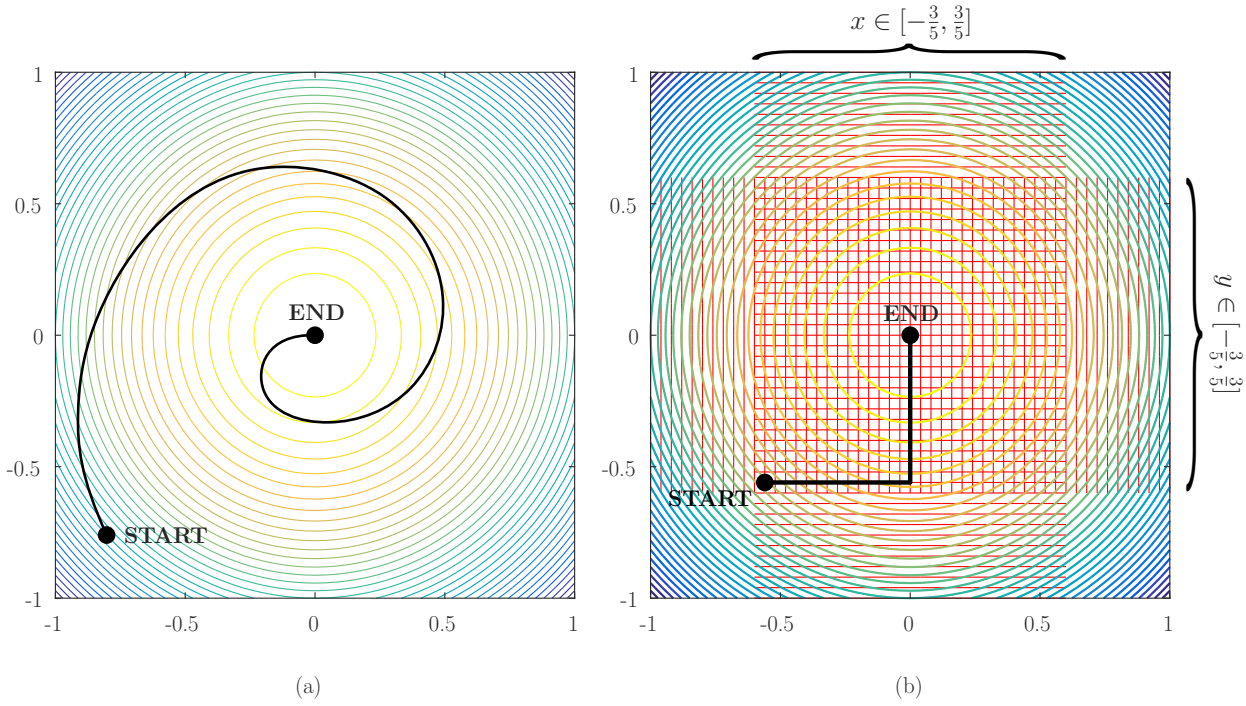


Figure 2: The peak effect. Plot (a): contour lines of a mountain with a possible path to the peak in $(0, 0)$ satisfying the slope condition $|h^*| \le \eta$. Plot (b): edges of the graph G (in red) and a sample G-path.

3 Another failure of the discretization method: The peak effect

In this example we consider a ''mountain'' $h(x, y) = -\frac{1}{10}(x^2 + y^2)$ on the domain $D = [-1, 1] \times [-1, 1]$ with peak in $(x, y) = (0, 0)$, see Figure 2. Let $(x_{\text{END}}, y_{\text{END}}) = (0, 0)$. Intuitively, no matter where we start, a path to the peak always exists: it suffices to follow a screw-shaped path having a small angle with respect to a contour line, see Figure 2(a).

Let us analyze how the graph $G(V, E)$ looks, still working with our sample value $\eta = 0.12$. Recall that an edge of the form $\{(i, j), (i + 1, j)\}$ is in E if $|h_{ij} - h_{i+1,j}|/\Delta_x \le \eta$. This condition is a discrete approximation of the continuous condition $\left|\frac{\partial h}{\partial x}(x_i, y_j)\right| \le \eta$, where $(x_i, y_j) = (\underline{x} + (i - \frac{1}{2})\Delta_x, \underline{y} + (j - \frac{1}{2})\Delta_y)$ is the midpoint of box B_{ij}. Reasoning loosely now, assume that the grid is dense enough that the difference between the continuous and

discrete condition can be neglected. Then we can write

$$\{(i,j),(i+1,j)\} \in E \Leftrightarrow \left|\frac{\partial h}{\partial x}(x_i, y_j)\right| \leq 0.12 \Leftrightarrow \left|-\frac{1}{5}x_i\right| \leq 0.12 \Leftrightarrow -\frac{3}{5} \leq x_i \leq \frac{3}{5}$$

(but the first equivalence shouldn't be understood literally). This condition tells us which boxes B_{ij} are incident with horizontal edges of G (edges of G are plotted in red in Figure 2(b)). For the vertical edges, we get a similar condition: $\{(i,j),(i,j+1)\} \in E \Leftrightarrow -\frac{3}{5} \leq y_i \leq \frac{3}{5}$. The figure shows that G is disconnected. Moreover, a path to the peak $(0,0)$ exists only when we start in a box B_{ij} with $(x_i, y_j) \in [-\frac{3}{5}, \frac{3}{5}] \times [-\frac{3}{5}, \frac{3}{5}]$. Otherwise G contains no path to the peak. (And this property remains true even if we try to make the grid even more dense.) In addition, the shortest paths in G are trivial — one of them is plotted in Figure 2(b) in black. Obviously, there is no way how the method could represent the screw-shaped path outlined in Figure 2(a).

We used the terrain with the specific shape $h(x,y) = -\frac{1}{10}(x^2+y^2)$ just for an illustration of the peak effect with a very simple analysis. A (slightly) more realistic landscape from Figure 3 shows a similar phenomenon in the neighborhood of the peaks, too.

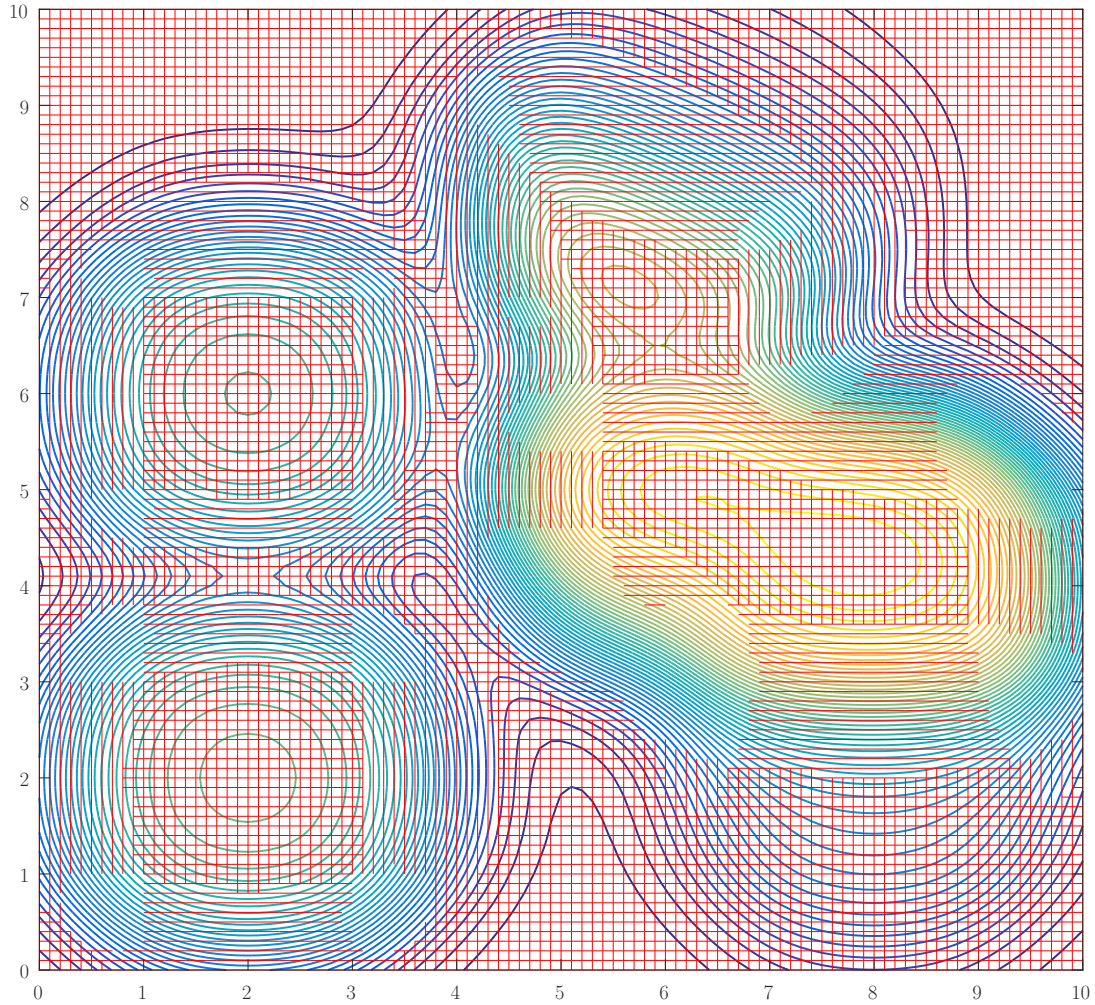


Figure 3: Edges of graph G (in red) in a more complex terrain (blue = low contour lines, yellow = high contour lines).

4 Towards an improvement: Additional edges in G

To avoid at least some of the undesirable properties of the discretization method, it is natural to consider the graph G with some additional edges. We still identify the G-vertices (i,j) with boxes B_{ij} as previously. Now we consider a possible edge $\{(i,j),(i',j')\}$ for *any* pair of boxes B_{ij}, $B_{i',j'}$, $(i,j) \neq (i',j')$, not necessarily adjacent. A natural rule to consider is as follows: the edge $\{(i,j),(i',j')\}$ should be present in E iff (a version of) the η-slope condition holds true *along the entire line segment* $L := \{\lambda(x_i, y_j) + (1-\lambda)(x_{i'}, y_{j'}) \mid \lambda \in [0,1]\}$. Let us test

this condition again by discretization of the line segment L. Setting $(x^k, y^k) := (x_i, y_j) + k\Delta(x_{i'} - x_i, y_{j'} - y_j)$ with $\Delta := \sqrt{(x_{i'} - x_i)^2 + (y_{j'} - y_j)^2}/\max\{|i' - i|, |j' - j|\}$, we test if $|h(x^k, y^k) - h(x^{k-1}, y^{k-1})|/\Delta \leq \eta$ for every $k = 1, \ldots, \max\{|i' - i|, |j' - j|\}$. If the inequality holds true for all k, the edge $\{(i, j), (i', j')\}$ is added into G.

Remark. The construction of the augmented graph G is an extension of graph G as described previously. For example, all red edges from Figure 3 are present in G, plus many other ones (not visible in the figure).

Remark. The method can produce many redundant edges that can be removed: for example, it can happen that all edges $\{(i, j), (i, j + 1)\}$, $\{(i, j + 1), (i, j + 2)\}$ and $\{(i, j), (i, j + 2)\}$ are present in G. Such cases can be tested and redundant edges can be removed.

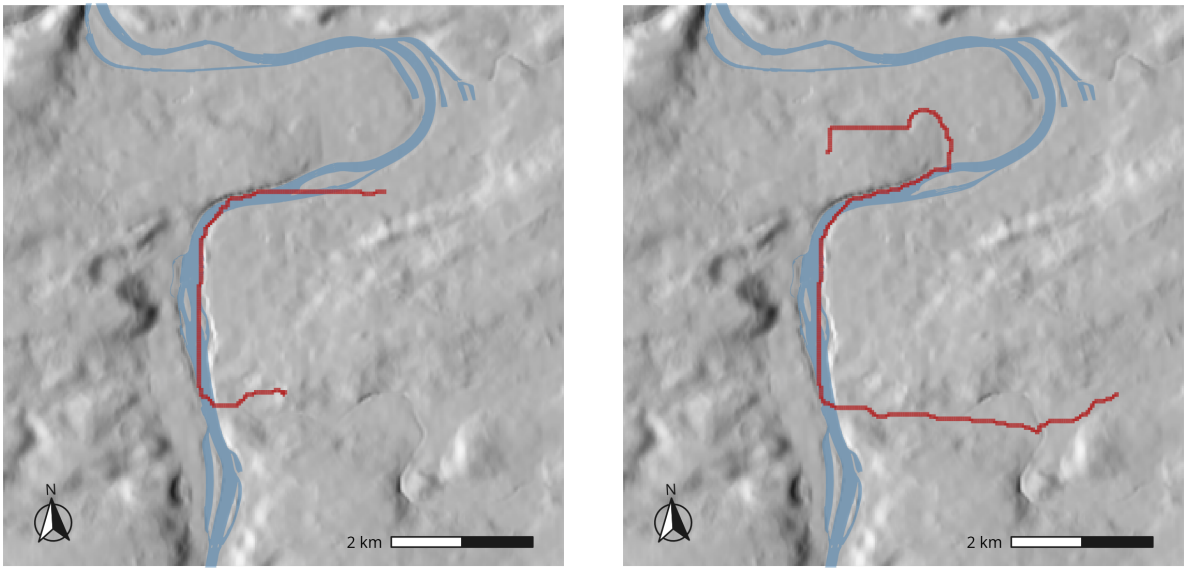
In general we lose the spartisity property: the inequality $|E| \leq 2|V|$ does not hold true any more, and the complexity of Dijkstra's method increases to $O(|E| \log |V|) = O(m^2 n^2 \log(mn))$.

5 The discretization method can work fine: A real-world example

Finally, we consider the discrete case within the context of the spatial data ecosystem of the statistical programming language `R`, using the software package `sfnetworks` (van der Meer et al., 2022) (which extends the structures introduced in `sf` (Pebesma, 2018) and `igraph` (Csardi and Nepusz, 2005) packages). We create a square grid B_{ij} covering the central part of Prague, capital of the Czech Republic, with $n = m = 256$, such that $\Delta_x = \Delta_y = 1/32$ km (i.e., the area of a grid box B_{ij} is 1/1024 km²). The heights h_{ij} are the altitudes from the EU-DEM (GISCO - Eurostat, 2016) digital elevation model in the centers of boxes B_{ij}. An edge $\{(i, j), (i', j')\}$ of G is associated with weight $|h_{ij} - h_{i'j'}|$ if $|h_{ij} - h_{i'j'}| \leq 0.12\Delta_x$; otherwise this edge is not present in G.

Finding the minimum-weight path using Dijkstra's algorithm now corresponds to finding a path with ''minimum up-down-up-down bumps'', regardless of its kilometric length.

In our example we used two sets of origin–destination pairs: (a) Karlín to Karlov (a path connecting two buildings of Faculty of Mathematics and Physics, Charles University) and (b) Sparta to Slavia (a route connecting two well-known rival football clubs), see figure 4.



(a) Karlín to Karlov

(b) Sparta to Slavia

Figure 4: Routing Examples Using a Real World Grid

In both examples, a significant part of the route follows the Vltava river, which we intentionally included in the grid as a valid region. This is because the river surface is by far the flattest area in Prague. And, indeed, the method found this flat path correctly.

Acknowledgements

J. L. was supported by grant No. IGA F4/18/2025, Faculty of Informatics and Statistics, Prague University of Economics and Business. M. Č. was supported by CSF project 25-18028S.

References

Csardi G, Nepusz T (2005) The Igraph Software Package for Complex Network Research. InterJournal Complex Systems:1695 https://igraph.org

EU-DEM - GISCO - Eurostat. https://ec.europa.eu/eurostat/web/gisco/geodata/digital-elevation-model/eu-dem#LAEA

Hirpa D, Hare W, Lucet Y, et al (2016) A bi-objective optimization framework for three-dimensional road alignment design. Transportation Research Part C: Emerging Technologies 65:61–78. https://doi.org/10.1016/j.trc.2016.01.016

Li W, Pu H, Schonfeld P, et al (2016) Methodology for optimizing constrained 3-dimensional railway alignments in mountainous terrain. Transportation Research Part C: Emerging Technologies 68:549–565. https://doi.org/10.1016/j.trc.2016.05.010

Meer L van der, Abad L, Gilardi A, Lovelace R (2022) sfnetworks: Tidy Geospatial Networks https://CRAN.R-project.org/package=sfnetworks

Pebesma E (2018) Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10:439. https://doi.org/10.32614/RJ-2018-009

Vázquez-Méndez ME, Casal G, Santamarina D, Castro A (2018) A 3D Model for Optimizing Infrastructure Costs in Road Design. Computer-Aided Civil and Infrastructure Engineering 33:423–439. https://doi.org/10.1111/mice.12350

Received: 2025-06-03 Last revision: 2025-07-02 Accepted: 2025-07-02

Towards Evaluation of the Czech Primary Education and Its Effect on Civic Engagement and Governance

Martin Flegl[1], Kristýna Vltavská[2], Andres Acero[3]

Abstract.
The Czech primary education faces several challenges mainly linked to socioeconomic inequalities, which affect students' outcomes, low literacy levels, and outdated teaching methods. As significant differences in educational quality can be observed across the country, it is necessary to develop instruments to enhance quality in education and eliminate inefficient resource allocations. Furthermore, the Czech Republic has historically faced challenges in fostering widespread political engagement, with many citizens showing limited participation in formal democratic processes. To analyze the role of education in civic engagement, a two-stage DEA cross-efficiency model was constructed, where Stage 1 evaluates primary education quality and Stage 2 evaluates the impact of primary education quality on civic engagement and governance. The analysis used information about education and social conditions in the Czech Republic and information related to civic engagement and governance in regions published by DataPAQ and the System for evaluating and comparing the quality of life of the population of the Czech Republic, respectively. As socio-economic disparities exist between different regions and within individual regions as well, the analysis is presented for 206 Czech municipalities with extended jurisdiction. Results of such micro-level analysis enable the creation of precise decision-making policies.

Keywords: Czech Republic, data envelopment analysis, primary education, cross-efficiency analysis, political and community activities.

JEL Classification: C44, C61, I21
AMS Classification: 90C05

1 Introduction

Primary schools in the Czech Republic are established by municipalities or by voluntary associations of municipalities whose remit includes educational responsibilities (CSI, 2024). Primary education in the Czech Republic is compulsory and accessible to all children. Historically, primary (as well as regional) education has been significantly underfunded, which led to a reform of regional education funding (including primary schools). Its main objectives are to use financial instruments to support quality in education, eliminate inefficient and unequal allocation of state funds, and reach the OECD average level (measured by the share of total education expenditure to gross domestic product) (MEYS, 2020).

One can identify both positives or strengths, as well as negatives or challenges that primary education in the Czech Republic will face. According to the Strategy for Education Policy of the Czech republic up to 2030 (MEYS, 2020), the four main positives include a high degree of school autonomy and decentralized management (allowing schools to adapt education to local conditions and student needs, or to react flexibly to changes), efforts to modernize education and focus it on competencies needed for the 21st century, the existence of Framework Educational Programmes, and the initiated funding reform. However, primary education in the Czech Republic also faces several negative aspects. These include inequalities in education, which primarily means the dependence of student outcomes on parents' socioeconomic status, or the low basic literacy levels among many students. Another negative is the content, methods, and assessment of education, exemplified by overloading students with information rather than fostering deeper understanding, or the continued prevalence of frontal teaching (lecture-style instruction). A further negative issue is the existence of ethnic segregation or segregation based on

[1] Tecnologico de Monterrey, School of Engineering and Sciences, Calle Puente 222, Coapa, Arboledas del Sur, Tlalpan, 14380, Mexico City, Mexico, martin.flegl@tec.mx. ORCID: 0000-0002-9944-8475.
[2] Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic, kristyna.vltavska@vse.cz. ORCID: 0000-0002-8579-4027.
[3] Tecnologico de Monterrey, School of Engineering and Sciences, Mexico, andres.acero@tec.mx. ORCID: 0000-0002-1730-1892.

socioeconomic background. Lastly, other negative aspects include the high administrative burden on schools and the shortage of qualified and high-quality teachers in some regions and schools. The Strategy 2030 (MEYS, 2020) also outlines plans to address these negatives or challenges in the Czech primary education.

In the Czech Republic, there is also a significant disparity in the quality of primary education between different regions and within individual regions as well. The most affected regions are Karlovarský and Ústecký, which have long demonstrated the poorest results in national and international comparisons (e.g. PISA, TIMSS). Students in these regions achieve lower scores, a higher proportion enter vocational training leading to an apprenticeship certificate, and these regions have the highest proportion of unqualified teachers. In these regions, the negative influence of family socioeconomic background on children's educational outcomes is also stronger. The goal of the Strategy 2030 is to continuously reduce these regional disparities through targeted financial and methodological support for schools and teachers (MEYS, 2020).

Data Envelopment Analysis (DEA) is a non-parametric methodology widely used for efficiency and performance analyses. The DEA has a long history of efficiency evaluations in education related to regional (Agasisti et al., 2022; Flegl et al., 2023), institutional (Henriques and Marcenaro-Gutierrez, 2021; Lin and Yu, 2023) and educational (Mendoza-Mendoza et al., 2023) perspectives. Considering the regional perspective, Flegl et al. (2023) evaluated efficiency of primary and secondary education in 32 Mexican states and its impact on regional, economic and social development. The analysis revealed that higher education quality leads to better regional development, but such higher development is not reflected in higher efficiency. Agasisti et al. (2022) assessed efficiency of 10,600 Brazilian elementary public schools and observed differences between its five regions and 26 states. The authors discovered significant differences across the country that are related to socioeconomic conditions. Camanho et al. (2021) analysed efficiency of Italian upper secondary schools regarding 20 Italian regions grouped into four macro-regions. The analysis confirmed disparity between the North and South of the country in the case of students' educational achievements.

Education plays a crucial role in shaping civic and social engagement, as literacy is a prerequisite for meaningful participation in civic life. In this case, quality education forms students' skills in understanding complex issues, forms opinions and helps engaging in discussions (Hendricks, 2025). Further, education helps address civic engagement inequality as people with limited access to quality education struggle to participate in political and community activities (Schröder and Neumayr, 2023). Le and Nguyen (2021) adds that education cultivates political interest, however better educated people are not more likely to vote in elections. Within the Czech Republic context, the country historically experiences low levels of political involvement, as people are often disengaged from political processes (Vráblíková, 2025).

Considering the above-mentioned negative aspects of the Czech primary education (such as identified inequalities in education system), and low civic engagement, the objective of the presented article is to investigate the impact of education on civic engagement in the Czech Republic. For this purpose, a two-stage cross-efficiency DEA methodology was adopted, using data from 206 Czech municipalities with extended jurisdiction (MEJs).

2 Materials and methods

2.1 Cross efficiency Data Envelopment Analysis

Data Envelopment Analysis evaluates efficiency of a set of n homogeneous and comparable DMUs, where $\mathrm{DMU}_j, (j = 1,2,\dots,n)$ consumes m inputs $x_{ij}, (i = 1,2,\dots,m)$ to produce s outputs $y_{rj}, (r = 1,2,\dots,s)$. The linearized output-oriented constant return to scales (CCR) model for DMU_o is defined as follows:

$$\begin{aligned} &\min e_o^* = \sum_{i=1}^{m} v_i x_{io} \\ s.t.\ &\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0 \\ &\sum_{r=1}^{s} \mu_r y_{ro} = 1 \\ &\mu_r, v_i \geq \varepsilon. \end{aligned} \tag{1}$$

where $v_i, (i = 1,2,\dots,m)$ and $\mu_r, (r = 1,2,\dots,s)$ represent the weights of the inputs and outputs, and ε is the so-called non-Archimedean element important to eliminate zero weights. Solving model (1) for each $\mathrm{DMU}_o, (j = 1,2,\dots,n)$ gives a series of efficiency scores e_o^* that are based on optimal weights v_i^{o*} and μ_r^{o*}. The efficiency scores e_o^* can take values between 0 and 1, and DMU_o is considered 100% efficient if $e_o^* = 1$, otherwise it is considered as inefficient (Li et al., 2021).

In many cases, the basic one-stage production process described above is not suitable for characterized complex processes. To evaluate the impact of primary education on civic engagement and governance, we use the two-stage process as depicted in Figure 1. To extend the traditional one-stage evaluation process, we consider that stage 1 consists of D outputs $z_{dj}, (d = 1,2, \dots, D)$, which become the inputs of the second stage. Finally, the second stage of the process produces s outputs y_{rj} (Cooper et al., 2011; Kao and Hwnag, 2008).

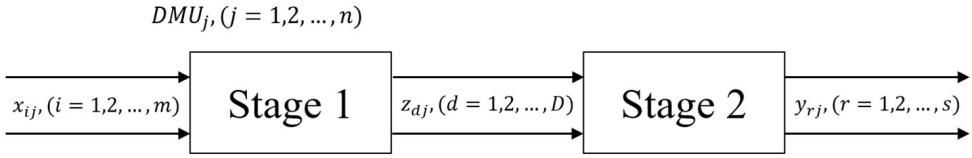


Figure 1 Two-stage production process

Further, to enhance the discrimination power of the model, the cross-efficiency DEA approach can be used. In this case, DMU's efficiency is assessed from multiple perspectives, considering both self-evaluation and peer evaluation (Li et al., 2021). This is particularly relevant when the analysis does not include dynamic evaluation across several periods, making the analysis more robust. So, the cross-efficiency of DMU_j is calculated regarding $\mathrm{DMU}_o, (j = 1,2, \dots, n)$ as following

$$e_j^o = \sum_{i=1}^{m} v_i^{o*} x_{ij} \Big/ \sum_{r=i}^{s} \mu_r^{o*} y_{rj} \tag{2}$$

As Du et al. (2014) pointed out, cross-efficiency score of $\mathrm{DMU}_j, (j = 1,2, \dots, n)$ is calculated as the arithmetic mean value of all its o cross-efficiency scores as

$$e_j = \sum_{o=1}^{n} e_j^o / n. \tag{3}$$

2.2 Data and model structure

The constructed two-stage DEA model evaluates primary education quality in Stage 1, whereas Stage 2 evaluates the impact of primary education quality on civic engagement and governance. The analysis is carried out for 206 Czech municipalities with extended jurisdiction (MEJs) and reflects situation in 2021/2022, although variables associated with elections reflect situation in 2018 and 2020.

The input part of Stage 1 consists of 3 variables: 1) Primary school funding per pupil: this index contains expenditures from the state and municipalities directed to the pupils' education, calculated per year and pupil in primary schools. This includes salaries of pedagogical and non-pedagogical staff and teaching assistants, other non-investment expenditures, non-investment contributions (equipment, operation, etc.), investment transfers and municipal expenditures on education that cannot be directly attributed to contributory organizations, and subsidies. 2) Qualified teaching: this input represents a share of lessons taught in primary schools and multi-year grammar schools by teachers who meet the required education according to the Act on pedagogical workers (MSMT, 2025). 3) Number of full-time teacher positions per pupil at primary school: this input reflects the number of teachers in primary schools and multi-year grammar schools converted to full-time positions per pupil.

The output part of Stage 1 includes 2 variables: 1) Testing results: this index is calculated from the results of students in the Czech School Inspectorate testing in grades 5 and 9 and from the results of the unified entrance exam. 2) Educational success: this index is calculated as inverted index of educational failure, which is composed of non-completion of primary education, repetition of a grade, failure and absence. The information was collected from DataPAQ regional data viewer tool (https://www.datapaq.cz) that collects data on education and social conditions in the Czech Republic. The outputs of Stage 1 represent the intermediate outputs D that enters as inputs to Stage 2. The variables were selected based on similar studies elaborated in primary education (Agasisti et al., 2022; Camanho et al., 2021; Hendricks, 2025; Lin and Yu, 2023).

In the case of Stage 2 outputs, we use data related to civic engagement and governance in regions published by the System for evaluating and comparing the quality of life of the population of the Czech Republic in the regional dimension (https://www.zivotvregionech.cz/). The selected variables include: 1) Voter turnout in municipal elections: it is calculated as the proportion of the number of voters who cast their votes to the number of registered voters in each region in 2018. 2) Voter turnout in regional elections: it is calculated as the proportion of the number of voters who cast their votes to the number of all registered voters in each region in 2020. 3) Number of candidates for a mandate: this variable measures the number of candidates for a mandate in municipal elections, calculated as

the ratio of the total number of eligible candidates to the total number of mandates in 2018. Higher numbers of candidates for a mandate indicate a higher interest of the region's residents in active participation in territorial administration. 4) Association activity: this indicator measures the number of associations per 1,000 inhabitants of a given region in 2021. The selected variables correspond to DEA applications in civic engagement and governance, such as Peiró-Palomino et al. (2021).

We used the CCR output-oriented model for both stages, and epsilon was set equal to .3. The MaxDEA 7 Ultra software was used for all calculations.

3 Results

The results are divided into three main sections: first, we describe the results regarding the primary education efficiency (Stage 1); second, the obtained results regarding the civic engagement and governance efficiency (Stage 2) will be presented; third, the relation between both stages will be analyzed.

The average efficiency of the primary education was 0.545 with standard deviation (StDev) 0.160. This indicates lower education efficiency as the operation process in the Czech primary education can be improved by 45.5%. Further, 68% of MEJs are evaluated within the range of 0.385 and .705. The best-evaluated MEJs were Luhačovice (with a cross-efficiency 0.978 and StDev 0.057), Praha (0.976, StDev 0.043), Černošice (0.937, StDev 0.055), Říčany (0.910, StDev 0.037) and Šlapanice (0.0865, StDev 0.033). On the other hand, the 5 worst-evaluated MEJs were Žatec (0.0.156, StDev 0.030), Rumburk (0.189, StDev 0.053), Most (0.226, StDev 0.019), Chomutov (0.243, StDev 0.020) and Králíky (0252, StDev 0.073). As can be observed, the results of the best and worst evaluated MEJs revelated very low variability, as the StDev shows a stable cross evaluation between all MEJs.

Considering the civic engagement and governance stage, the average cross-efficiency was 0.490 with StDev 0.154. The results obtained in Stage 2 again reveal very low efficiency across the evaluated MEJs, which is -0.055 lower compared to Stage 1 (with similar StDev -0.006). In this case, 68% of the MEJs reported efficiency between 0.336 and 0.645. The highest civic engagement and governance efficiency was achieved in Žatec (0.998, StDev 0.029), Frýdlant (0.956, StDev 0.066), Rumburk (0.917, StDev 0.025), Vítkov (0.895, StDev 0.087) and Jeseník (0.868, StDev 0.080). On the other side, the five worst-evaluated MEJs were Jablunkov (0.252, StDev 0.011), Luhačovice (0.262, StDev 0.020), Blansko (0.274, StDev 0.020), Rosice (0.275, StDev 0.018) and Kravaře (0.278, StDev 0.015).

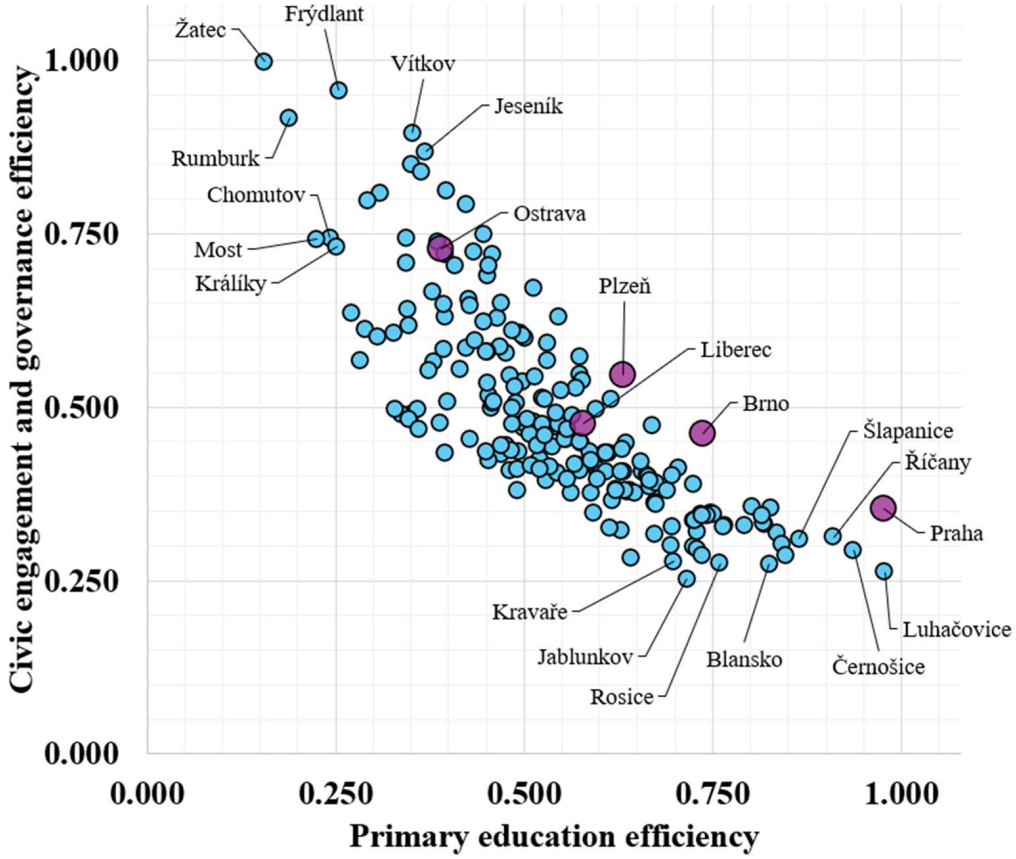


Figure 2 Relation between primary education efficiency and civic engagement and governance efficiency

Figure 2 displays the relation between primary education efficiency and civic engagement and governance efficiency. The analysis discovers an inverse relationship, as higher efficiency in primary education leads to lower civic engagement and governance efficiency. The highest evaluated MEJS in Stage 1 reports one of the lowest efficiencies in Stage 2, and vice versa. Taking example of both sides, Luhačovice was the best evaluated MEJ in Stage 1, but was ranked as the 2nd worst in Stage 2. Similarly, Žatec was ranked as the worst MEJ in Stage 1, but reached the top evaluation in Stage 2. If we take a closer look at the data, in Stage 1, the best-evaluated MEJs had on average -1.71% lower primary school funding per pupil, +0.99% higher qualified teaching and -4.43% number of full-time teacher positions per pupil. However, these MEJs reported +666.17% better testing results and +224.72% higher educational success. However, such significantly higher outcomes of the educational process were not proportionally reflected in civic engagement and governance. Data indicates that the best-evaluated MEJs in Stage 1, had only +32.31% higher voter turnout in municipal elections, +52.22% higher voter turnout in regional elections and +34.93% higher association activity, with -14.44% lower number of candidates for a mandate.

If we consider the results in Stage 2, the MEJs with the highest civic engagement and governance efficiency reported +3.83% higher primary school funding per pupil, -10.96% lower qualified teaching and +4.89% number of full-time teacher positions per pupil, with -54.64% worse testing results and -72.78% lower educational success. Those outcomes resulted in -14.30% lower voter turnout in municipal elections, -13.58% lower voter turnout in regional elections, with +19.52% higher number of candidates for a mandate and +3.52% higher association activity.

4 Conclusion

The presented analysis used the cross-efficiency Data Envelopment Analysis methodology to evaluate relationship between primary education efficiency and civic engagement and governance efficiency in the Czech Republic. The analysis was performed using data for 206 Czech municipalities with extended jurisdiction. The analysis revealed very low average efficiency in both stages (0.545 and 0.490 respectively), which indicates high space for improvements. The results further revealed a negative relation between education efficiency and civic engagement and governance efficiency. On the one side, the analysis confirmed the positive effect of education on civic engagement, as teaching quality and literacy play an essential role in political interest and community activities. However, on the other hand, considering the efficiency point of view, this positive effect was not observed. One reason may be the brain drain of individuals from the municipalities to bigger cities where they obtained their education and found employment during or after completing their studies. In such cases, they participate in elections in their new cities (e.g. with a voter identification card) and are also more involved in civic associations in the areas where they live.

References

Agasisti T, de Oliveira Ribeiro C, Sanches Montemor D (2022) The efficiency of Brazilian elementary public schools. International Journal of Educational Development 93: 102627. https://doi.org/10.1016/j.ijedudev.2022.102627

Camanho AS, Varriale L, Barbosa F, Sobral T (2021) Performance assessment of upper secondary schools in Italian regions using a circular pseudo-Malmquist index. European Journal of Operational Research 289(3): 1188-1208. https://doi.org/10.1016/j.ejor.2020.07.050

Cooper W, Seiford LM, Zhu J (2011) Handbook on Data Envelopment Analysis. New York: Springer.

CSI (2024) Kvalita vzdělávání v České republice - Výroční zpráva 2023/2024 [Quality of education in the Czech Republic - Annual Report 2023/2024]. Česká školní inspekce. https://www.csicr.cz/CSICR/media/Prilohy/2024_p%c5%99%c3%adlohy/Dokumenty/VZ_CSI_2023-2024_everze.pdf (accessed 13 May 2025).

Du J, Cook WD, Liang L, Zhu J (2014) Fixed cost and resource allocation based on DEA cross-efficiency. European Journal of Operational Research 235(1): 206-214. http://dx.doi.org/10.1016/j.ejor.2013.10.002

Flegl M, Avilés-Sacoto SV, Güemes-Castorena D, Avilés-Sacoto EC (2023) A State-Level Analysis of Mexican Education and Its Impact on Regional, Economic, and Social Development: Two-Stage Network DEA Approach. Journal on Efficiency and Responsibility in Education and Science 16(4): 275-286. https://doi.org/10.7160/eriesj.2023.160402

Hendricks A (2025) What is Education's Impact on Civic and Social Engagement?. Simply stakeholders. https://simplystakeholders.com/civic-engagement-education/ (accessed 14 May 2025).

Henriques CO, Marcenaro-Gutierrez OD (2021) Efficiency of secondary schools in Portugal: A novel DEA hybrid approach. Socio-Economic Planning Sciences 74: 100954. https://doi.org/10.1016/j.seps.2020.100954

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(1): 418-429. https://doi.org/10.1016/j.ejor.2006.11.041

Le K, Nguyen M (2021) Education and political engagement. International Journal of Educational Development 85: 102441. https://doi.org/10.1016/j.ijedudev.2021.102441

Li F, Wu H, Zhu Q, Liang L, Koou G (2021) Data envelopment analysis cross efficiency evaluation with reciprocal behaviors. Annals of Operations Research 302: 173–210. https://doi.org/10.1007/s10479-021-04027-x

Lin Y-C, Yu M-M (2023) Performance evaluation of compulsory education system in Taiwan: A modified dynamic network data envelopment analysis approach. Studies in Educational Evaluation 78: 101280. https://doi.org/10.1016/j.stueduc.2023.101280

Mendoza-Mendoza A, De La Hoz-Domínguez E, Visbal-Cadavid D (2023) Classification of industrial engineering programs in Colombia based on state tests. Heliyon 9(5): e16002. https://doi.org/10.1016/j.heliyon.2023.e16002

MEYS (2020) Strategy for the Education Policy of the Czech Republic up to 2030+. Ministry of Education, Youth and Sports. https://msmt.gov.cz/uploads/brozura_S2030_en_fin_online.pdf (accessed 13 May 2025).

Peiró-Palomino J, Picazo-Tadeo AJ, Tortosa-Ausina E (2021) Measuring well-being in Colombian departments. The role of geography and demography. Socio-Economic Planning Sciences 78: 101072. https://doi.org/10.1016/j.seps.2021.101072

Schröder JM, Neumayr M (2023) How socio-economic inequality affects individuals' civic engagement: a systematic literature review of empirical findings and theoretical explanations. Socio-Economic Review 21(1): 665-694. https://doi.org/10.1093/ser/mwab058

Vráblíková K (2025) Czech Republic—From Apathy to Mass Mobilisations, 1989–2022. In: Crăciun, C., Rammelt, H.P. (eds) Power and Protest in Central and Eastern Europe. Palgrave Studies in European Political Sociology. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-77888-9_5

Received: 2025-06-08 Last revision: 2025-07-15 Accepted: 2025-07-29

Efficiency Dynamics in the Czech Brewing Industry: A Post-Pandemic Comparison of Small, Medium, and Large Breweries

Martina Kuncová[1], Roman Fiala[2]

Abstract. This paper builds on previous analyses of efficiency in the Czech brewing industry using Data Envelopment Analysis (DEA) models. It focuses on differences in technical efficiency among micro, small, and medium-to-large breweries in the post-COVID-19 period (2021–2022). The study applies input-oriented DEA models with both constant and variable returns to scale (CCR and BCC) to evaluate how effectively breweries convert economic inputs (labour costs, debt, equity) into financial outputs (sales, profit). While the results suggest a trend of increasing efficiency with firm size, statistical tests indicate that these differences are not significant. The findings provide practical insights for managers and policymakers and highlight opportunities for performance improvement across different types of breweries.

Keywords: DEA model, brewing industry, efficiency, small breweries, medium-sized enterprises, large breweries, post-pandemic analysis

JEL Classification: C44, C61, C67
AMS Classification: 90B50, 90C08

1 Introduction

Beer is one of the few alcoholic beverages that have been consumed for centuries. The Czech Republic has a deep-rooted brewing tradition, with beer production documented as early as 1088 (Basařová and Hlaváček, 1999; Danišová, 2020). Over the centuries, brewing developed in parallel with the development of towns, especially during the founding of royal towns in the 12th and 13th centuries. However, during industrial consolidation in the 19th and 20th centuries, the number of breweries declined significantly, from approximately 1,000 in 1841 to less than 500 in 1925 (Danišová, 2020). By the 1960s, fewer than 100 breweries remained. Since the 1990s, the Czech brewing landscape has undergone dynamic changes. While the number of industrial breweries has remained relatively stable (52 in 2019), the craft beer sector has grown rapidly—from 27 microbreweries in 1999 to over 450 in 2019 (Tripes, 2019; Tripes and Dvořák, 2017). By 2024, the total number of active breweries approached 570, ranking the Czech Republic among the top European countries by brewery density (pividky.cz, © 2011–2025). The Czech beer market is characterized by the dominance of a few large-scale producers alongside a rapidly growing number of micro and craft breweries. Since 2010, the number of craft breweries in the Czech Republic has increased significantly, with more than 350 new establishments founded between 2010 and 2020. Despite the already high density of breweries—one for every 20,000 inhabitants—more than 65% of craft brewery owners believe there is still room for new entrants on the market (Březinová, 2021).

Globally, Anheuser-Busch InBev (Belgian-American concern) was the largest beer company in the world in 2024 with sales of approximately US$59.4 billion, the second largest company was Heineken Holding with sales of $32.8 billion (Conway, 2024). Heineken operates almost worldwide, with a significant presence in Western Europe. As Conway (2025) notes, the beer market in Europe has faced several years of difficulties caused by the COVID-19 pandemic and, more recently, the war in Ukraine and subsequent inflation, which has caused many breweries to close or prices to rise. Although sales started to pick up in 2024, the alcoholic beer sector has been stagnating for several years as consumers reduce consumption and seek other alcoholic beverages or non-alcoholic beers. The leading producer of beer in Europe is Germany and with 76 million hectoliters consumed, Germany was also the biggest beer consuming country in Europe. But when comparing beer consumption per capita, the Czech Republic is still the leader with about 128 liters in 2023. In 2023, a person in Czechia consumed, on average, 133.2 liters of beer. This was a decrease of over nine liters compared to the previous year (Statista.com, 2025).

[1] Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Econometrics, W. Churchill Sq. 4, 13067, Prague 3, kuncovam@vse.cz
[2] College of Polytechnics Jihlava, Department of Economic Studies, Tolstého 16, 58601 Jihlava, roman.fiala@vspj.cz

This paper builds on previous applications of Data Envelopment Analysis (DEA) to the Czech brewing industry and examines how small, medium, and large breweries adapted to post-pandemic market conditions. Our findings (Fiala and Kuncová, 2021) from the year 2021 for the Czech breweries have suggested that size alone does not determine performance; instead, managerial and operational practices play a critical role. By applying input-oriented DEA models with both constant and variable returns to scale, we assess differences in the efficiency. The findings offer practical insights for strategic management and efficiency optimization across diverse breweries.

2 Efficiency measurements in the Beer Industry and DEA models

The beer industry is characterized by diverse production scales, ranging from small craft operations to large industrial breweries. Measuring efficiency across such heterogeneous units requires robust analytical tools that can accommodate differences in input-output structures and scale. In this context, DEA has become a widely adopted method for evaluating performance and benchmarking efficiency in both domestic and international brewery studies. Bernetti et al. (2020) analyzed the efficiency of 163 Italian microbreweries. They used input-oriented DEA models with three inputs (number of employees, debt-to-equity ratio, total debt) and one output (sales and service revenue), considering both CRS and VRS models. Sellers-Rubio (2018) focused on the estimation of advertising efficiency in the Spanish brewing industry. The input-oriented DEA models were used to analyze six brewing enterprises between 2007 and 2014 to calculate advertising efficiency, considering six inputs (four different advertising expenditures, number of employees, capital) and two outputs (total sales revenue, total beer sales). Ezan (2015) used a two-stage DEA model to measure the efficiency of 500 beer producing enterprises. The first stage involved profitability analysis, with the output variables (sales and operating profits) serving as inputs in the second stage, known as sales analysis. In the second stage, three input variables (total assets excluding financial and real estate investments, equity and total number of employees) were considered, while four indicators (earnings per share, average share price, return on invested capital, net profit) were used as outputs.

The Czech breweries were analyzed by several researchers. Kasem et al. (2015) conducted research on 14 Czech breweries focusing on efficiency evaluation using DEA models. They used indicators from the Global Reporting Initiative and KPIs along with data sets from environmental, financial and sustainability reports. The models used four inputs: number of employees, average employee wages and bonuses, amount of waste, and percentage of women in leadership positions in the company. Two outputs were considered: cash flow and economic value added (EVA). Hřebíček et al. (2015) demonstrated the applicability of DEA models in assessing the sustainability and performance of Czech breweries using standardized KPIs aligned with ESG principles. Their work showed that breweries could significantly benefit from efficiency evaluation when formulating long-term operational strategies. Another study by Kasem and Trenz (2020) proposed a three-stage sustainability assessment system for 89 Czech brewing companies. The first phase involved data collection and calculation of sustainability value added (SVA) based on the concepts of environmental, social and administrative value added. The second phase included a DEA analysis and a sustainability assessment model. To automate the processes in the first two phases, the authors used the Web-based Enterprise Performance Rating and Sustainability Reporting Information System (WEBRIS) web portal to collect data. Finally, an analysis was performed based on questionnaires and data from 89 Czech breweries. Březinová (2021) identified factors influencing the establishment of small breweries, noting that while these entities often lack economies of scale, they may compensate through innovation, local branding, and adaptability. However, their efficiency varies widely, reinforcing the need for tools like DEA to capture their performance accurately.

In our previous studies (Fiala and Kuncová, 2021; Kuncová and Fiala, 2023) we also examined the relationship between brewery size and relative efficiency. The older study focused on whether larger breweries tend to achieve higher relative efficiency, using DEA analysis and suggesting a weak but positive relationship. The follow-up 2023 study extended the analysis over five years and confirmed that while many small breweries operate efficiently, size still plays a modest role, particularly in years with greater market volatility such as 2019. The study of Sekničková and Kuncová (2023) was focused on the evaluation of breweries with respect to profit, cost and production efficiency and examined the impact of the COVID-19 pandemic in the Czech Republic on the whole brewing sector.

DEA models were introduced by Charnes, Cooper, and Rhodes (1978) to evaluate the performance of diverse entities such as countries, regions, companies, schools, hospitals, insurance companies, and military units (Charnes et al., 1985). DEA models calculate the efficiency of decision-making units (DMUs). The key principle is to estimate the efficient frontier, which represents the optimal relative ratios of inputs and outputs for the compared DMUs. A DMU is efficient if it resides on the efficient frontier, while inefficient DMUs fall below it and can enhance their efficiency by adjusting inputs or outputs (Cook and Seiford, 2009).

Let us denote $\boldsymbol{Y} = (y_{rj}, r = 1, \dots, R, j = 1, \dots n)$ a non-negative matrix of outputs and $\boldsymbol{X} = (x_{sj}, s = 1, \dots, S, j = 1, \dots n)$ a non-negative matrix of inputs. The efficiency score of the unit j_0 is calculated using the model:

$$\begin{aligned} \text{Maximise} \quad & U_{j_0} = \frac{\sum_{r=1}^{R} u_r y_{r,j_0}}{\sum_{s=1}^{S} v_s x_{s,j_0}} \\ \text{subject to} \quad & \frac{\sum_{r=1}^{R} u_r y_{r,j}}{\sum_{s=1}^{S} v_s x_{s,j}} \leq 1, \quad j = 1, \dots, n, \\ & u_r \geq \varepsilon, \qquad r = 1, \dots, R, \\ & v_s \geq \varepsilon, \qquad s = 1, \dots, S, \end{aligned} \tag{1}$$

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 U_{j_0} equals 1 for the efficient units and lower than 1 for the inefficient units. In model (1), the objective function is non-linear, which necessitates the linearization of the model. Typically, two approaches are employed for linearization based on the orientation of the linear model: input-oriented and output-oriented. This article, like similar papers, focuses on input-oriented models. The linearized version of the input-oriented model, commonly referred to as the CCR model (Charnes et al, 1978), is presented as follows:

$$\begin{aligned} \text{Maximise} \quad & U_{j_0} = \sum_{r=1}^{R} u_r y_{r,j_0} \\ \text{subject to} \quad & \sum_{s=1}^{S} v_s x_{s,j_0} = 1, \\ & \sum_{r=1}^{R} u_r y_{r,j} - \sum_{s=1}^{S} v_s x_{s,j} \leq 0, \quad j = 1, \dots, n, \\ & u_r \geq \varepsilon, \quad r = 1, \dots, R, \\ & v_s \geq \varepsilon, \quad s = 1, \dots, S. \end{aligned} \tag{2}$$

The CCR input-oriented models assume constant returns to scale (CRS). It means that if the DMU with an input/output combination $(\boldsymbol{x}, \boldsymbol{y})$ is efficient, then the unit with an input/output combination $(\alpha\boldsymbol{x}, \alpha\boldsymbol{y})$, where $\alpha > 0$ is also efficient. On the other hand, it is possible to use variable returns to scale (VRS) or BCC models. The multiplicative form of the BCC input-oriented model (3) uses μ as a free variable and $U_{j_0}^V$ as the efficiency score for j_0-th unit under evaluation. The $U_{j_0}^V$ equals one for the efficient units. The inefficient units have an efficiency score lower than 1.

$$\begin{aligned} \text{Maximise} \quad & U_{j_0}^V = \sum_{r=1}^{R} u_r y_{r,j_0} + \mu \\ \text{subject to} \quad & \sum_{s=1}^{S} v_s x_{s,j_0} = 1, \\ & \sum_{r=1}^{R} u_r y_{r,j} - \sum_{s=1}^{S} v_s x_{s,j} + \mu \leq 0, \quad j = 1, \dots, n, \\ & u_r \geq \varepsilon, \quad r = 1, \dots, R, \\ & v_s \geq \varepsilon, \quad s = 1, \dots, S, \end{aligned} \tag{3}$$

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 decided for 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.

As mentioned above, on both models the returns to scale (RTS) are important. If the results of the CCR and BCC model are available, it is also possible to calculate the so-called Scale efficiency (SE). SE is usually calculated as the ratio of the efficiency score of the unit under the assumption of CRS (CCR model) to the efficiency score of the unit under the assumption of VRS (BCC model) and refers to the degree to which a DMU is operating at an optimal scale (Aparicio and Santín, 2024). SE is calculated for each DMU and it is always between zero and one. The values below one suggest the potential for improvement in resource utilization. It also can measure how much of the inefficiency is due to operating at the wrong scale (too small or too large).

3 Data

For the efficiency analysis we chose two years period (2021-2022) to be able to compare the results with our previous research (Kuncová and Fiala, 2023). For the evaluation of the economic efficiency of breweries that operated (according to the classification of economic activities NACE) in the section 11-Manufacturing of beverages, 1105-Manufacture of beer, data from the Albertina CZ Gold Edition database was used. According to the previous analysis based on similarly focused research, we included personnel costs, foreign resources, and equity as inputs, and profit/loss for the accounting period and sales of goods, products, and services as outputs to the DEA models.

For 2021, data was available for 113 enterprises, but for 11 of them the number of employees was not provided. For 2022, there were 106 enterprises of which 10 were missing data on the number of employees. Table 1 describes the distribution of enterprises by employee category in both years. After the aggregation into categories we had 43/40 micro enterprises (up to 10 employees in 2021/2022), 36/33 small enterprises (10-49 employees), 16 medium enterprises in both years (50-249 employees) and 7 large enterprises (250 or more employees).

employees cat.	1 - 5	6 - 9	10 - 19	20 - 24	25 - 49	50 - 99	100 - 199	250 - 499	500 - 999	2 500 - 2 999	n/a
2021	37	6	14	8	14	11	5	2	4	1	11
2022	34	6	14	8	11	11	5	2	4	1	10

Table 1 Number of enterprises based on the employees' category

It is evident that the largest number of enterprises is in the category of micro-enterprises with up to 10 employees - we decided not to combine them with enterprises with 10-49 employees, which is the second largest number. As mentioned above, this is a situation typical in the Czech Republic, where there are many small breweries, the number of which varies slightly from year to year, while the number of medium and large breweries remains constant, as in our sample. The statistics of the data set of the year 2021 (3 inputs, 2 outputs) are in Table 2.

Company size			micro	small	middle	large
Number of companies			43	36	16	7
CZK millions	**I1: personnel costs**	**avg.**	2,481.9	6,856.5	39,020.8	542,438.3
		min.	19	412	14,033	92,048
		max.	28,877	17,294	66,385	2,019,000
	I2: foreign resources	**avg.**	12,432.7	19,281.9	53,060.7	2,175,009.1
		min.	148	614	9,170	265,436
		max.	107,474	111,835	170,038	8,422,000
	I3: equity	**avg.**	6,385.6	23,834.3	131,488.6	5,374,158.3
		min.	-19,349	-8979	10,892	280,943
		max.	124,235	166,289	348,525	21,124,000
	O1: profit/loss	**avg.**	-1,629.4	303	6,890.9	829,558.1
		min.	-19370	-8,675	-11,396	-15,535
		max.	5,486	10,213	21,825	4,481,000
	O2: sales	**avg.**	7,776.5	25,449.6	156,150.8	4,098,648.4
		min.	102	1,412	53,698	261,501
		max.	102,390	90,418	400,704	16,335,000

Table 2 DEA models' data statistics for the year 2021

For micro and small enterprises, it is possible to find negative equity, and in each category, there is a brewery making a loss in a given year. The same situation can be seen from the 2022 data (Table 3). DEA models usually suppose the data to be non-negative, but in Frontier Analyst software a shifting procedure is used to avoid this problem (on the other hand we could exclude these DMUs - however, this would lose a third of the enterprises in the sample, so we decided to keep them in the analysis).

In all categories of breweries, we can observe a year-on-year deterioration of average values for individual indicators. This is most evident in microbreweries, where all average values have deteriorated except for foreign resources. For small companies, there has been a decline in the average for equity, while for large companies there has been a decline in the profit/loss indicator (in each year it is one of the large breweries that has reported a loss

as an economic result, in Pivovary Group CZ in one year and Pivovar Staropramen in 2022). One of the reasons has already been mentioned above - the shift of consumers away from beer towards other types of beverages. On the other hand, the good news is the relatively stable number of breweries year-on-year.

Company size			micro	small	middle	large
Number of companies			40	33	16	7
CZK millions	I1: personnel costs	avg.	1,775.1	7,082.8	38,383.3	624,617.3
		min.	6	505	459	119,851
		max.	28,218	18,431	80,719	2,215,000
	I2: foreign resources	avg.	15,714	25,443.6	54,788.8	2,109,208.6
		min.	251	1,087	476	355,695
		max.	204,761	174,099	168,773	7,665,000
	I3: equity	avg.	3,887.5	23,794.7	191,674.3	5,655,657.6
		min.	-86,809	-17,699	34	185,658
		max.	83,759	169,805	1,223,467	17,558,000
	O1: profit/loss	avg.	-1,631	458.8	2,903.8	639,960.1
		min.	-40,476	-12,599	-45,282	-2,202,000
		max.	2,574	11,320	38,219	5,915,000
	O2: sales	avg.	7,670.1	29,239.5	156,366.1	4,820,621.6
		min.	2	1,874	1,253	773,582
		max.	122,130	102,078	447,093	19,612,000

Table 3 DEA models' data statistics for the year 2022

4 Results

In the first part of the analysis, we focused on calculating DEA models (CCR and BCC) for each category of firms separately. The results are summarized in Table 4 - expressed as percentages given the different numbers of enterprises in each category. It can be seen from the results that there has been a year-on-year increase in the % of efficient breweries in all categories except large companies, but this is to be expected, as we explain below.

CCR-I (2021)	micro	small	middle	large	BCC_I (2021)	micro	small	middle	large
0-0.5	35%	28%	19%	14%	0-0.5	21%	25%	6%	0%
0.5-0.99	32%	44%	43%	29%	0.5-0.99	26%	33%	50%	14%
efficient	33%	28%	38%	57%	efficient	53%	42%	44%	86%
CCR-I (2022)	micro	small	middle	large	BCC_I (2022)	micro	small	middle	large
0-0.5	38%	27%	6%	0%	0-0.5	23%	24%	6%	0%
0.5-0.99	25%	34%	50%	43%	0.5-0.99	23%	24%	44%	29%
efficient	38%	39%	44%	57%	efficient	55%	52%	50%	71%

Table 4 DEA results for both years

The distribution of microbreweries is very polarized (in CCR_I model): almost the same percentage are highly inefficient as fully efficient. This fact suggests a high heterogeneity in this category – as also noted by Březinová (2021), For small breweries, the percentage with efficiency above 0.5 and efficient in CCR-I is already higher compared to inefficient. But in both categories the results look better in BCC-I models where more than 50% of enterprises are efficient. For medium-sized enterprises, the majority were rated as fairly efficient, especially in the BCC-I models. In this case, however, the results may already be partially biased due to the number of DMUs (16) and the number of inputs and outputs (5 in total). This is then a major problem for large enterprises. Large breweries have the highest share of fully efficient units (57% in CCR-I and more in BCC-I), even though there are only 7 enterprises, which is problematic given the dimensionality (3 inputs + 2 outputs). Therefore, we decided to merge the categories of medium and large enterprises into one to increase the number of DMUs.

The results of the CCR_I models after merging are shown in Figure 1. It can be seen that as the number of employees increases, the share of inefficient breweries decreases and the share of firms with medium or higher efficiency increases (especially in 2022). The share of efficient ones is similar in all three categories, or slightly higher for microbreweries.

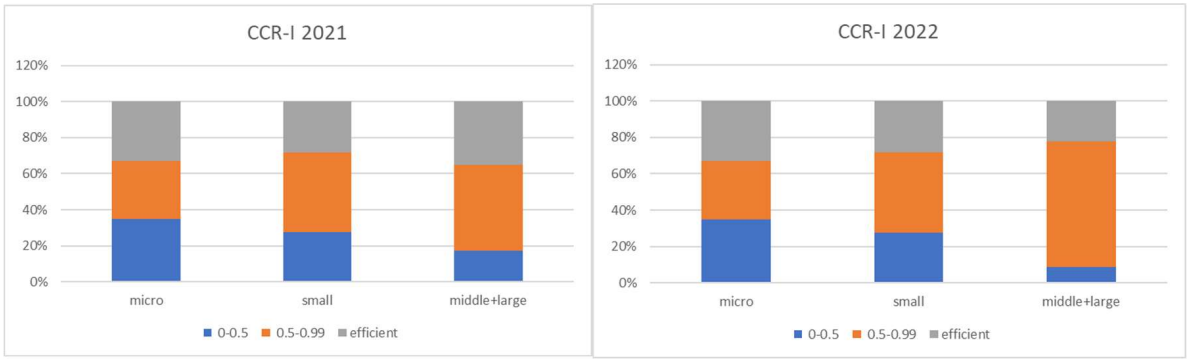


Figure 1 Results of CCR-I models for both years

In addition to the CCR and BCC models, scale efficiency was calculated as the ratio of CCR to BCC scores, enabling a more detailed view of whether inefficiencies stem from pure technical inefficiency or from operating at a non-optimal scale. The distribution of scale efficiency (SE) across company sizes in 2021 and 2022 is summarized in Figure 2. A value of SE = 1 indicates that a firm is operating at an optimal scale, while values less than 1 suggest scale inefficiency. In 2021, microbreweries showed a polarized distribution: 33% were fully scale efficient, but 12% had SE below 0.5, indicating possible challenges in scaling operations efficiently. Small enterprises performed relatively better: over 58% operated near efficiency (SE between 0.8–0.99) and only 3% were severely inefficient. Medium and large enterprises, when combined, showed a balanced pattern with 35% fully efficient, another 35% near efficiency, and no firms in the lowest SE band. In 2022, scale efficiency improved for small and larger firms. The share of fully efficient small firms increased from 28% to 36%, and medium+large enterprises had no scale-inefficient units below SE = 0.8. Micro firms remained polarized - although 35% were fully efficient, the share with very low scale efficiency rose slightly to 15%.

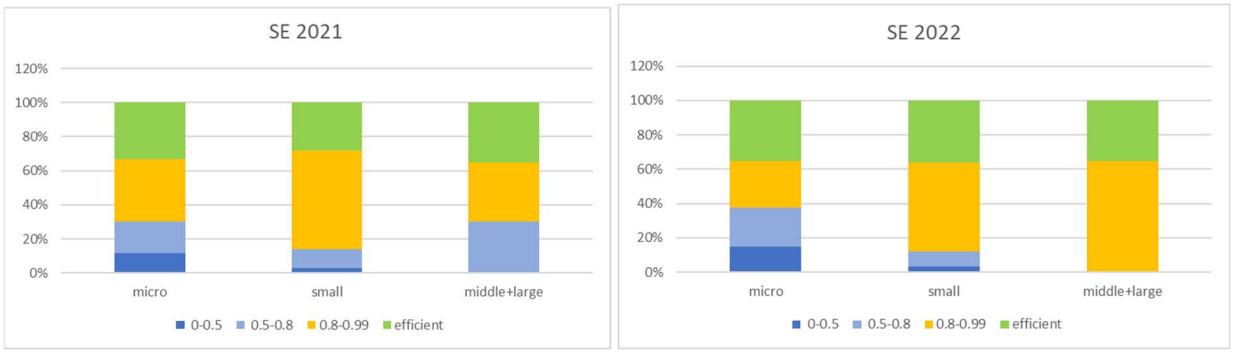


Figure 2 Results of SE for both years

The results can also be looked at in terms of the average efficiency of each group. Although the average efficiency in the medium and large firm category is slightly higher than in the other two categories (Table 5), the difference is not statistically significant (Kruskal-Wallis test, p-value 0.533 and 0.804 for CCR-I and BCC_I averages, respectively). For the years' comparison based on the SE average efficiency, small enterprises consistently achieved the highest average scale efficiency (0.918 in 2021 and 0.929 in 2022), suggesting that they tend to operate closer to their optimal scale compared to micro or larger firms, which showed a notable increase, from 0.901 in 2021 to 0.977 in 2022, indicating an overall improvement in scale optimization. On the other hand, microenterprises experienced a slight decline, from 0.837 to 0.802, further reinforcing the previously observed polarization within this group. This pattern confirms that while technical efficiency remains relatively stable, scale-related inefficiencies are more pronounced in micro firms, whereas small and larger firms have shown improvements

year	**2021**			**2022**		
average efficiency	**micro**	**small**	**middle + large**	**micro**	**small**	**middle + large**
CCR-I	0.661	0.655	0.745	0.651	0.723	0.786
BCC_I	0.783	0.708	0.820	0.783	0.772	0.873
SE	0.837	0.918	0.901	0.802	0.929	0.977

Table 5 Average DEA efficiencies for both years

.

5 Conclusions

The analysis presented in this contribution focuses on differences in efficiency across small, medium, and large breweries in the post-COVID-19 period in the Czech Republic. DEA models were used to evaluate how efficiently individual breweries use economic and capital inputs (personnel costs, external resources, equity) to achieve economic outputs (sales, profit). The measurement thus corresponds to technical efficiency in the economic sense. Data from 2021 and 2022 show that larger breweries (medium and large) have on average higher technical efficiency than small and microbreweries, which is consistent with our previous research (Kuncová and Fiala, 2023). However, the differences are not statistically significant, which may be due to the variable behaviour of firms during and after the pandemic and different adaptive capacities. Micro-breweries show the greatest variability in efficiency. While microbreweries outperformed small breweries in 2021, the opposite was true in 2022, suggesting a lack of consistent efficiency advantage between the two groups. Both the CCR-I and BCC-I DEA models provided a consistent picture of slightly increasing efficiency with firm size. It is important to note that the analysis covered only selected indicators and a sample of about one-sixth of the Czech market, where all large breweries were included, but not all micro and small breweries (data unavailability). However, the fact that the average efficiency of all categories is relatively high and has increased slightly in 2022 can be seen as a positive result.

Acknowledgements

The paper was supported by the grant No. F4/18/2024 of the Faculty of Informatics and Statistics, Prague University of Economics and Business and by the contribution of long-term institutional support of research activities by the College of Polytechnics Jihlava.

References

Aparicio J, Santín D (2025) Global scale efficiency in data envelopment analysis. International Transactions in Operational Research 32: 2474-2496. https://doi.org/10.1111/itor.13501

Basařová G, Hlaváček I (1999) České pivo, NUGA, Praha

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

Březinová M (2021) Beer industry in the Czech Republic: Reasons for founding a craft brewery. Sustainability 13(17): 9680. https://doi.org/10.3390/su13179680

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

Charnes A, Cooper W W, Golany B, Seiford L, Stutz J (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. Journal of econometrics, 30: 91–107. https://doi.org/10.1016/0304-4076(85)90133-2

Conway J (2024) Sales of the leading beer companies worldwide. https://www.statista.com/statistics/257670/sales-of-the-leading-beer-companies-worldwide/

Conway J (2025). European beer market - statistics & facts. https://www.statista.com/topics/8020/european-beer-market/#topicOverview

Cook W D, Seiford, L M (2009) Data envelopment analysis (DEA) – Thirty years on. European Journal of Operational Research 192: 1–17. https://doi.org/10.1016/j.ejor.2008.01.032

Danišová I (2020) Zlatý mok očima statistiky. Statistika & My 10/2020: 20–21.

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. https://doi.org/10.10.35603/sws.iscss.2021/s04.18

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

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

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

Sekničková J, Kuncová M (2023) The Impact of the COVID-19 Pandemic in the Brewing Industry with Regard to Profitability, Cost and Production EfficiencyAnalysis of the efficiency of Czech breweries in the years 2016-2020. In: 41st International Conference on Mathematical Methods in Economics 2023, Conference Proceedings, Prague University of Economics and Business, Prague, pp 365–372. https://mme2023.vse.cz/mme_2023_proceedings.pdf

Pividky.cz (© 2011–2025) Seznam pivovarů. https://pividky.cz/seznam-pivovaru

Sellers-Rubio R (2018) Advertising efficiency in the Spanish beer industry: spending too much? International Journal of Wine Business Research 30:, 410–427. https://doi.org/10.1108/IJWBR-11-2017-0066

Statista.com (2025) Annual consumption of alcoholic beverages per capita in Czechia 2011-2023, https://www.statista.com/statistics/1311266/czechia-annual-consumption-of-alcoholic-beverages-per-capita/

Tripes S (2019) Leaders of Czech Craft Breweries and Their Ability to Attract the Customer. In: 15th European Conference on Management, Leadership and Governance, Conference Proceedings. University of Porto, Portugal, pp 383–390. https://www.proquest.com/openview/e4a18b7cb7ef756ac47446574c2de0c0/1?pq-origsite=gscholar&cbl=1796418

Tripes S, Dvořák J (2017) Strategic forces in the Czech brewing industry from 1990-2015. Acta Oeconomica Pragensia 25: 3–38. https://doi.org/10.18267/j.aop.580

Received: 2025-06-30 Last revision: 2025-08-13 Accepted: 2025-08-13

Automated Trading Systems as a Modern Tool in Citizen Investing

Petr Brzobohatý[1], Aliaksei Sidoryk[2]

Abstract. In an era of growing interest in investing and the automation of financial processes, automated trading systems (ATS) are gaining increasing importance. These systems enable investors to apply algorithmic strategies without the need for constant market supervision. The aim of this work was to design and implement an open ATS that allows users to create, configure, and backtest trading strategies on historical data. The developed system is built on modern technologies — the backend is implemented in Python, the frontend in React, and the solution also integrates an AI-based model for predicting market movements. In the practical part, the application is tested on AI strategies and compared with other types of asset investment. The results confirm that combining traditional rule-based approaches with predictive models can significantly improve the decision-making process in algorithmic trading.
Keywords: automated trading system, algorithmic trading, artificial intelligence, backtesting, investing

JEL Classification: C44
AMS Classification: 90C15

1 Introduction

Inflation in recent years has significantly eroded the purchasing power of not only Czech citizens (ČNB, 2025; ČSÚ, 2025). As a result, inflation forces people to think of family wealth as a portfolio of assets that could generate real returns (Brodani et al., 2025). The most typical way to invest money is through real estate (ČNB, 2024; CONSEQ, 2025).

1.1 Real estate

Real estate is perceived as a safer alternative to stocks, even though it is not always liquid. Property prices often rise during periods of high inflation because people seek ways to preserve the value of their money. Real estate is seen as a stable investment that resists inflation over the long term. Higher inflation also increases construction costs, which pushes property prices upward (CONSEQ, 2025). Therefore, in an inflationary environment, both investors and ordinary citizens more frequently direct their savings into real estate (Brodani et al., 2025).

1.2 Commodity

A second option is commodities such as gold, silver, or timber, which often serve as a hedge against inflation (KB, 2025; Kurzy.cz, 2025). These investments can help preserve wealth even in uncertain times. Gold, like real estate, has historically resisted inflation. Although gold does not generate any yield, over the past year it has outperformed all indices and is also exempt from VAT when purchased (Zlaté rezervy, 2025). Real estate and commodities are generally suitable during high inflation (CONSEQ, 2025).

1.3 Financial products

A third alternative is financial products such as stocks and bonds, which, on the other hand, thrive in an environment of low and stable inflation. Although stocks have historically been profitable, many Czech citizens avoid them for several reasons (Brodani et al., 2025; KB, 2025). One of the main reasons is the previously mentioned low financial literacy, which leads to fear of risk and a lack of understanding of how the stock market works. Historical distrust also plays a role, stemming from negative experiences such as the coupon privatization in the 1990s or financial crises like the dot-com bubble in 2000 or the 2008 crisis—distrust that parents unknowingly passed on to their children (Brodani et al., 2025).

[1] Tomas Bata University, Faculty of Applied Informatics, Nad Stráněmi 4511, Zlín 5, brzobohaty@utb.cz.
[2] Tomas Bata University, Faculty of Applied Informatics, Nad Stráněmi 4511, Zlín 5, a_sidoryk@utb.cz.

An improving offering of stock and real estate funds is now accessible to the average user via internet banking of any bank or broker (an intermediary who facilitates the buying and selling of securities) (Czech news center, 2025). As Czech household savings have grown over the past five years, this topic has become relevant again (Brodani et al., 2025). Trading fees with modern brokers range from 0% to 0.3% , which is significantly lower than the fees of bank-managed stock or real estate funds, where investors face entry, management, and often exit fees ranging from 2% to 4% (Czech news center, 2025). This difference has a major impact on overall returns, especially in long-term investing, as lower investment costs are a key factor in effective appreciation (Traxler, 2023).

1.4 Cash

Citizens often choose a conservative approach by holding cash, which yields little or no return but provides quick access to funds in case of unexpected expenses or investment opportunities (Traxler, 2023). It also serves as a buffer when the value of other assets declines. A major obstacle is that the Czech population still lacks sufficient investment education. Basic financial literacy is not adequately taught in schools or within families (CONSEQ, 2025).

While these traditional asset classes form the foundation of most investment portfolios, recent technological advances have introduced entirely new tools for market participation. One of the most prominent innovations is automated trading systems (ATS), a technology that allows investors to implement algorithmic strategies with minimal manual intervention.

2 Automated trading systems

2.1 Introduction to automated trading systems

Today, ordinary users have access to tools that allow them to trade practically anytime and anywhere. One such tool is ATS. ATS are designed to enable investors to use algorithmic strategies without constant market monitoring (Sidoryk, 2025). In practice, this means that trading decisions—buying and selling assets—are made automatically by the system based on predefined rules or data-driven predictive models. A major advantage of these systems is the elimination of the emotional component of trading, which often leads to irrational decisions and financial losses. Solutions are emerging that allow more advanced users to design and test their own strategies, adjust parameters, and analyze performance using historical data. Artificial intelligence (AI) plays a significant role here, as it can identify patterns and suggest optimal decisions based on large datasets. In the future, AI should also be able to advise on specific investment steps in terms of risk and money management, tailored to the investor's profile (Sidoryk, 2025).

This article focuses on the design and implementation of a custom ATS that combines traditional rule-based approaches with AI predictive algorithms. The goal is to create a platform that is open, flexible, and intuitive enough for less experienced users. The proposed system enables not only the creation and configuration of strategies but also their testing on historical data and direct comparison of multiple strategies. By combining these features, the tool can serve not only for education and hypothesis testing but also as a foundation for real-world deployment in financial markets. Investment funds are held in broker accounts, so trust in the broker is crucial (Czech news center, 2025).

2.2 Automated trading systems and their principles

ATS are software tools that enable trading on financial markets without the need for direct user intervention. Trades are executed based on algorithms and predefined rules that determine when to open or close a position, how to manage risk, or what market conditions must be met to activate a specific strategy (Sidoryk, 2025).

The core components of an ATS include: automatic opening and closing of positions, application of risk management rules (e.g., Stop Loss, Take Profit), backtesting capabilities on historical data and 24/7 operation without the need for human supervision.

ATS can be designed for various types of strategies, such as **momentum strategies**, which follow the current trend and aim to capitalize on it; **mean reversion strategies**, which anticipate a return of the price to its average; or **breakout strategies**, which enter positions when key levels are breached. We have turn to one single strategy combining classical indicators (e.g., SMA, RSI, MACD) (Sidoryk, 2025) with elements of artificial intelligence. The result is a flexible web application that allows advanced users to create, test, and compare strategies on historical data. Thanks to its modular architecture and the use of modern technologies (Python, React, TensorFlow), the system is easily extendable and adaptable (Sidoryk, 2025).

2.3 Advantages and disadvantages of automated trading systems

One of the greatest advantages of ATS is their ability to trade without emotions—all decisions are based on objective rules, which minimizes errors caused by fear or greed. Another key benefit is the speed with which the system reacts to market conditions, as well as scalability—strategies can be deployed simultaneously across multiple markets.

On the other hand, there are also risks. ATS depends on technological infrastructure—any error in the code, connection failure, or issue on the broker's side can lead to losses. Strategies that were successful on historical data may not perform well in the future, which is why it is essential to regularly test, adjust, and monitor the systems. Moreover, excessive reliance on automation can weaken a trader's ability to independently evaluate market situations (Sidoryk, 2025).

2.4 Technologies used and system architecture

ATS was developed as a two-layer architecture:

- Frontend is built using the ReactJS library. It provides a user interface for configuring trading strategies, analyzing and backtesting them, visualizing metrics and charts, and also allows customization of the application according to user preferences.
- Backend is implemented in Python, specifically using FastAPI for building the API. Communication takes place via a REST interface, secured with token-based authentication (Sidoryk, 2025).

2.5 The importance of backtesting

One of the most important features of any ATS is backtesting—the process of testing a strategy on historical data. This allows users to verify how the strategy would have performed in the past and optimize its parameters before deploying it in live trading.

The system also includes comparison of multiple strategies simultaneously, enabling evaluation of key metrics such as: total PnL (total profit and loss), average PnL per trade, max Drawdown – important parameter for conservative investors, sharpe Ratio and win Rate. These metrics help users better understand the behavior of a strategy and identify its strengths and weaknesses.

The user has the option to select a trading strategy, time frame, and instrument (ticker of the stock) on which they want to test the strategy. After making their selection, they can initiate a backtest, the results of which are displayed in three main sections: performance metrics, an interactive capital evolution chart and a trade table.

In our backtest, two stocks were selected: the stock of Tesla Inc. with the ticker TSLA (Yahoo Finance, 2025a) and the stock of company Apple with the ticker AAPL (Yahoo Finance, 2025b).

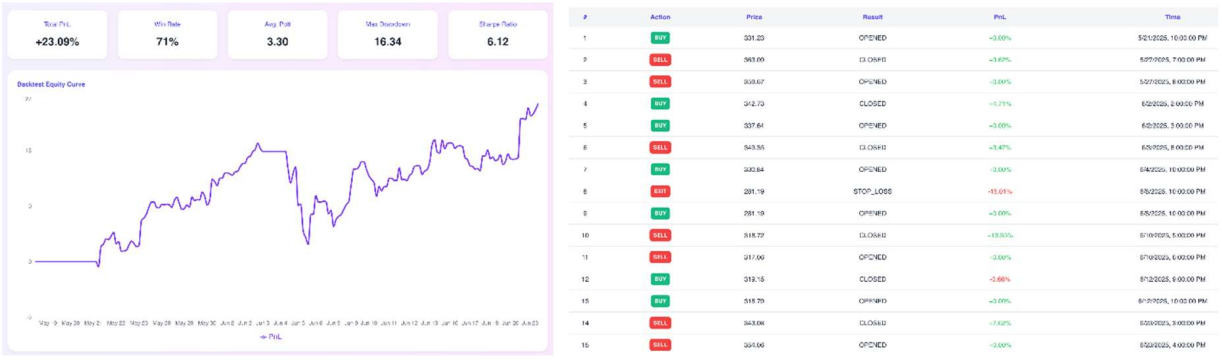


Figure 1 Equity curve with a metrics and the trade table for the stock Apple

2.6 Artificial intelligence

The system includes a predictive module that uses a combination of convolutional layers (CNN) and LSTM networks built on TensorFlow. The model processes historical prices and technical indicators, and classifies each market point into one of three categories as buy, sell or hold. The "Hold" signal is an im-portant part of the strategy. In such cases, the system takes no action, as in trading it is often better to wait than to make a move at the wrong time (Sidoryk, 2025). The requirements of the ATS strategy for entering or exiting a position were not met, so the system stays in HOLD signal what does mean an inactivity.

3 Methodology

3.1 Type of research

This study employs a comparative quantitative approach, involving the analysis of annual returns of four asset types based on data from the 2025 AKAT ČR survey on the investment behavior of Czech citizens during periods of high inflation (Brodani et al., 2025). These assets include real estate, commodities, financial products, and cash. The annual returns of these assets, after deducting fees (VALUO, 2025; CONSEQ, 2025; ČNB, 2024; Zlaté rezervy, 2025; Traxler, 2023), are compared with the annual return of an ATS (Sidoryk, 2025). The aim is to design and validate the effectiveness of the ATS strategy in comparison with traditional investment assets, integrating rule-based strategies with predictive algorithms powered by artificial intelligence.

3.2 Data and sample selection

To design the system, publicly available historical financial data were analyzed (e.g., prices of stocks (Yahoo Finance, 2025a; Yahoo Finance, 2025b), commodities (KB, 2025; Kurzy.cz, 2025), and real estate (ČNB, 2024; CONSEQ, 2025)), and compared together using two investment scenarios. The data were sourced from reliable platforms such as stock market databases, central banks (ČNB, 2025; ČSÚ, 2025), and brokerage services (Czech news center, 2025). Time frame for comparing and testing were set last year (from 06/2024 to 06/2025).

The first investment scenario involves a one-time investment of CZK 120,000 (USD 5,210 as of June 2024) into the individual asset classes discussed in the introduction.

The second scenario involves 12 regular monthly investments of CZK 10,000 (USD 433), totaling CZK 120,000 annually. With an average annual exchange rate of CZK 23.10/USD, this corresponds to approximately USD 5,195 (ČSÚ, 2025).

4 Results and discussion

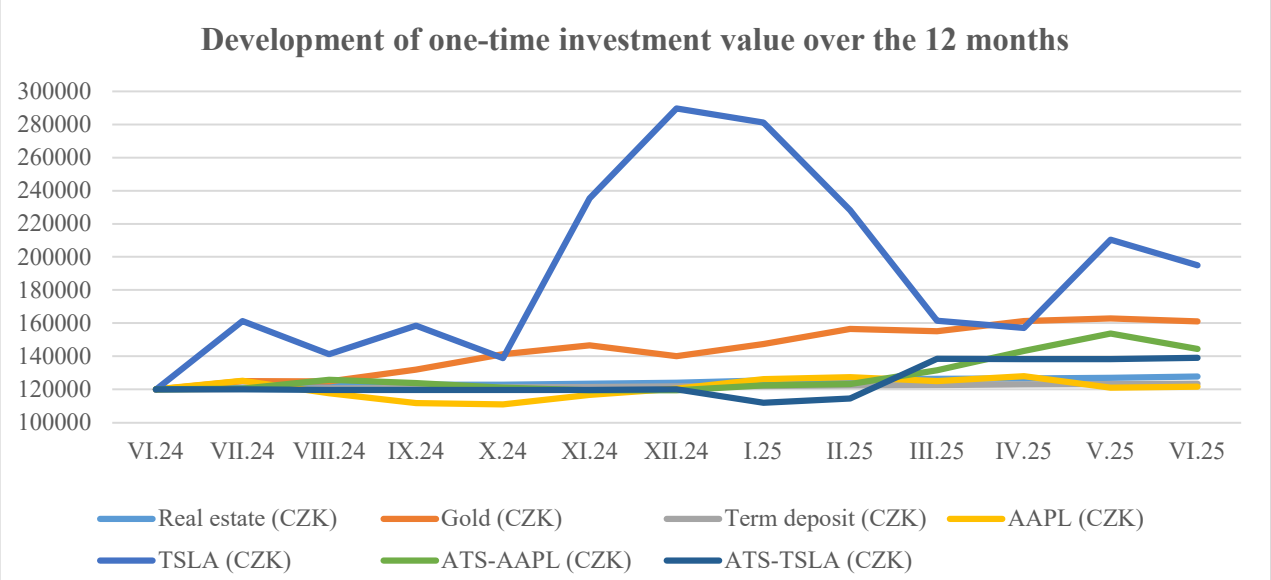


Figure 2 Development of one total investment value over the 12 months

Figure 2 illustrates the high volatility of Tesla stock and its impressive year-over-year gain of +60% (Yahoo Finance, 2025a). In contrast, Apple stock underperformed, showing almost no change in CZK terms compared to the previous year (+1.3%) (Yahoo Finance, 2025b). This indicates that although the value of Apple shares increased in USD, the appreciation of the Czech koruna during the observed period offset the gains, meaning that selling this investment in June 2025 would not have generated a profit in CZK (ČSÚ, 2025). Gold delivered a remarkable annual return of +34%, despite the weakening of the US dollar (KB, 2025; Kurzy.cz, 2025). Term deposits yielded 3% year-over-year, and with expected interest rate cuts in the future, returns are likely to decline further (Traxler, 2023). The performance of AI-driven strategies is also noteworthy. The ATS-AAPL strategy

achieved a +20% return, while ATS-TSLA returned +16%. The final profits are relatively close, as the AI strategy is designed to minimize drawdown, helping investors avoid emotional reactions during portfolio declines and reducing the likelihood of panic selling (Sidoryk, 2025). The final investment option includes real estate and increasingly popular real estate funds. According to the VALUO real estate index, the annual return on real estate was nearly 10% (VALUO, 2025). Real estate funds earned 6.6% (CONSEQ, 2025). However, after adjusting property prices for fees of 6%, the net return on real estate drops to 4%, while real estate funds yielded 4.8% after fees.

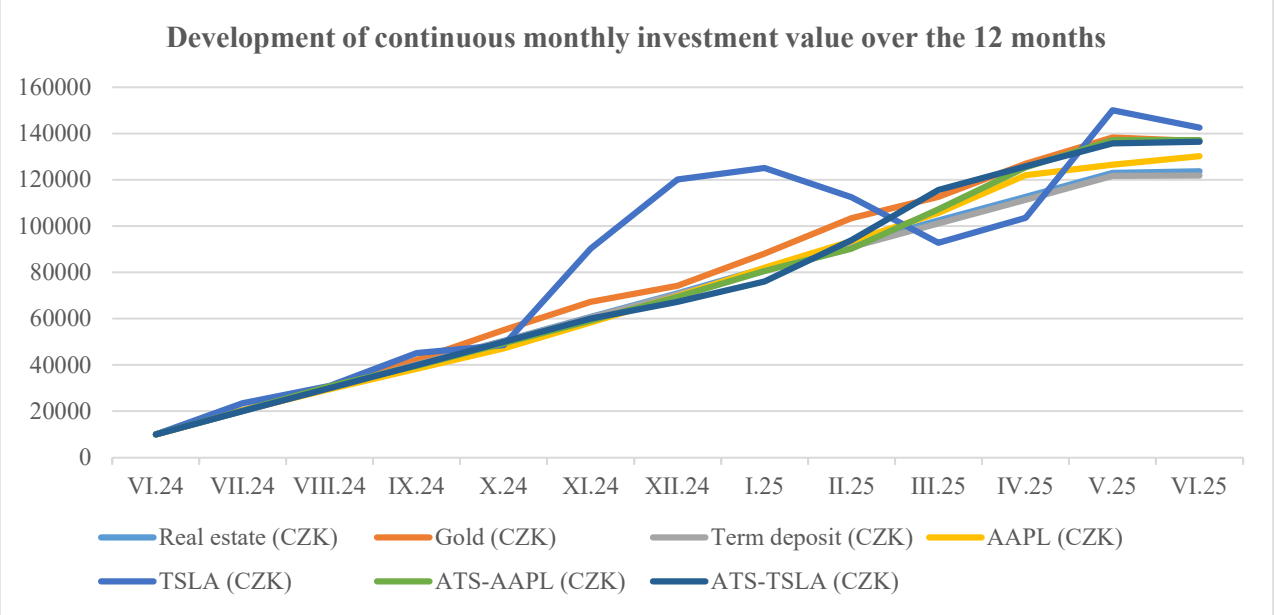


Figure 3 Development of continuous monthly investment value over the 12 months

In the case of continuous monthly investment, the situation does not change dramatically (Figure 3). The real estate fund yielded 3.1% through monthly investments, but after deducting the management fee (-1.8%), the total annual return was only 1.3% (CONSEQ, 2025). Term deposits generated just 1.6% annually with regular contributions (Traxler, 2023). If the return is lower than that of government bonds, one might question why an investor should take on the additional risk of investing in real estate. Over the same period, an investor who regularly invested in gold earned 14% (KB, 2025). The most profitable investments through monthly investing were Tesla stock (+18.8%) (Yahoo Finance, 2025a) and the AI-powered ATS-TSLA strategy (+13.8%). In the case of Apple stock, the AI strategy outperformed traditional regular monthly investing, delivering +14.3% compared to +8.5% (Yahoo Finance, 2025b). Practical testing has shown that combining classical indicators (e.g., SMA, RSI, MACD) with AI-based predictions can lead to better results and greater strategy robustness (Sidoryk, 2025).

5 Conclusion

The results show significant differences in the performance of individual investment instruments over the period from June 2024 to June 2025. The highest returns were achieved by Tesla stock (+60%) and gold (+34%), while Apple stock remained almost unchanged in CZK terms. This outcome also highlights the importance of currency risk—even if the value of shares increases in USD, the appreciation of the Czech koruna can significantly reduce or completely eliminate returns in the domestic currency.

Artificial intelligence within ATS strategies demonstrated the ability to stabilize returns and reduce drawdowns, with ATS-AAPL achieving +20% and ATS-TSLA +16%. These strategies are designed to minimize drawdown, helping investors avoid emotional decisions during market downturns.

Real estate and real estate funds offered solid gross returns (10% and 6.6%, respectively), but after accounting for fees, net returns dropped to 4% and 4.8%. Regular monthly investing produced similar results, with the most profitable options again being Tesla (+18.8%) and ATS-TSLA (+13.8%). In the case of Apple stock, the AI strategy outperformed traditional monthly investing (14.3% vs. 8.5%).

The findings confirm that diversification across different asset classes (stocks, gold, real estate, cash) and access to modern technologies such as AI-powered ATS can significantly contribute to effective investment performance, even during periods of economic uncertainty. At the same time, it is essential to consider currency effects and fee structures, which can have a substantial impact on final returns.

The authors plan to expand the system in the future by adding more models, live-trading features, and more advanced portfolio management. They are also considering the addition of a robo-advisory feature—a digital investment service that uses algorithms and automation to manage a user's funds held with their broker. It works by determining the user's investment profile based on a questionnaire about their goals, risk tolerance, and investment time horizon. The algorithms then propose a portfolio structure and automatically rebalance it to stay aligned with the original strategy. Thanks to automation, the costs are lower than those of traditional financial advisors.

References

Brodani J, Sladkovský J, Krajčovičová N (2025) Průzkum investičního chování Čechů [online]. AKAT ČR. https://www.akatcr.cz/DesktopModules/EasyDNNNews/DocumentDownload.ashx?portalid=0&moduleid=1674&articleid=4348&documentid=1982 [cit. 10. 10. 2025].

Conseq (2025) Přehled fondů [online]. Conseq Invest. https://www.conseq.cz/investice/prehled-fondu [cit. 10. 10. 2025].

Czech News Center (2025) Velké srovnání brokerů: Kde investovat s nejnižšími vstupními poplatky a nejlepšími službami? [online]. E15.cz. https://www.e15.cz/galerie/byznys/finance-a-bankovnictvi/303750/velkesrovnani-brokeru-kde-investovat-s-nejnizsimi-vstupnimi-poplatky-a-nejlepsimi-sluzbami?foto=1 [cit. 10. 10. 2025].

ČNB (2024) Český trh obytných nemovitostí: Včera, dnes a zítra [online]. Česká národní banka. https://www.cnb.cz/cs/cnbnews/aktuality/cnBlog-Cesky-trh-obytnych-nemovitosti-vcera-dnes-a-zitra [cit. 10. 10. 2025].

ČNB (2025) Inflace u 2% cíle [online]. Česká národní banka. https://www.cnb.cz/cs/menova-politika/inflacni-cil/tema-inflace/index.html [cit. 10. 10. 2025].

ČSÚ (2025) Inflace, spotřebitelské ceny [online]. Český statistický úřad. https://csu.gov.cz/inflace-spotrebitelske-ceny [cit. 10. 10. 2025].

Evropa V Datech (2024) Finanční gramotnost [online]. Index prosperity. https://www.indexprosperity.cz/2024/financni-gramotnost/ [cit. 10. 10. 2025].

KB (2025) Jak na investování do zlata [online]. Komerční banka. https://www.kb.cz/cs/obcane/kb-radce/zhodnoceni-financi/jak-na-investovani-do-zlata [cit. 10. 10. 2025].

Kurzy.cz (2025) Investiční zlato – aktuální cena a vývoj [online]. Kurzy.cz. https://zlato.kurzy.cz/investicni-zlato/ [cit. 10. 10. 2025].

Sidoryk, A. (2025) Návrh a implementace programu pro správu automatických obchodních systémů. Bakalářská práce. Zlín: Univerzita Tomáše Bati ve Zlíně.

Traxler, J (2023) Lepší než spořicí účet? Kdy se vyplatí depozitní fondy [online]. Peníze.cz. https://www.penize.cz/podilove-fondy/447567-lepsi-nez-sporici-ucet-kdy-se-vyplati-depozitni-fondy [cit. 10. 10. 2025].

Valuo (2025) Nemovitostní index VALUO [online]. Valuo. https://www.valuo.cz/index [cit. 10. 10. 2025].

Yahoo Finance (2025) Historical Data for Apple Inc. (AAPL) [online]. Yahoo Finance. https://finance.yahoo.com/quote/AAPL/history/ [cit. 10. 10. 2025].

Yahoo Finance (2025) Historical Data for Tesla Inc. (TSLA) [online]. Yahoo Finance. https://finance.yahoo.com/quote/TSLA/history/ [cit. 10. 10. 2025].

Zlaté Rezervy (2025) Vstupní poplatky a náklady investice do zlata [online]. Zlaté rezervy. https://www.zlaterezervy.cz/ [cit. 10. 10. 2025].

Received: 2025-05-23 Last revision: 2025-07-08 Accepted: 2025-07-08

The Solving of Optimal Routes for Multi-Warehouses Distribution by Returning Vehicles

Jaromír Zahrádka[1]

Abstract. This article proposes a solution to the vehicle routing problem for multi-store distribution of ordered goods when each vehicle returns to the point of departure. A wholesale company that offers food products to retail customers has multiple warehouses. It is assumed that each warehouse can distribute the entire range of products offered. The wholesale company solves the problem of distributing ordered goods from s warehouses to n customers as efficiently as possible. The goods ordered by each customer are placed in a certain number of transported containers. All stores and customer points are determined by GPS coordinates and all distances between warehouses and customer locations are known from truck navigation systems. The target of the solution is to heuristically select customers to the warehouses and plan the number of vehicles used and their routes between the distribution warehouses and the customer locations so that the total distance or travel time is as short as possible. The full procedure of solution is developed and implemented in Matlab code.

Keywords: Customer, goods, location, Matlab, optimization, point of departure, vehicle routing problem, warehouse, wholesale

JEL Classification: C64
AMS Classification: 68W04, 90C11, 05C20

1 The Vehicle Routing Problem with Multi-Store Distribution by Returning Vehicles

The vehicle routing problem for multistore distribution by returning vehicles (VRPMSR) describes a practical situation where a wide range of food products can be delivered from several wholesale warehouses to smaller stores in a large area. Smaller stores are customers of the wholesaler. The wholesaler has more fully stocked warehouses available in its territory from which it distributes goods to customers. All stores have the entire range of food products to be delivered and there are an enough number of vehicles available. The task specification requires that each vehicle used returns to its point of departure. The VRPMSR solution consists of planning the routes of trucks delivering goods so that the total distance travelled by all vehicles is as small as possible.

1.1 Background of resolved topic

The original exact solutions of the traveling salesman problem and the vehicle tracking problem are contained in (Eksioglu et al., 2009; Toth and Vigo, 1998). The subsequent modifications and applications of the exact methods are considered in (Jonak et al., 2020; Winston, 1994). I mention my articles in which the solution of the TSP and VRP are adapted for the case of using time windows and for the case of applying priority, for example, in the distribution of chilled and frozen goods (Zahrádka, 2023; Zahrádka, 2024). This presented article follows on from (Zahrádka, 2024) and includes the procedure that was created here to solve the current problem.

1.2 Mathematical Formulation of the VRPMSR

We calculate that n customers are served from s warehouses, i.e. depots. The customers sites (vertices) are marked by numbers $1, \dots, n$, and depots are marked by numbers $n+1, \dots, n+s$. The vehicle routing problem with multiple warehouses (VRPMSR) can be presented as the subsequent graph problem. Let $G = (V, E)$ be a complete directed graph where $V = \{1, \dots, n, n+1, \dots, n+s\}$ is the nodes set and E is the set of all oriented arcs. Nodes $i = 1,2, \dots, n$ correspond to the customers, each with a number q_i of ordered containers. The values q_i are arranged into components of the row vector $\boldsymbol{q} = (q_1,\ q_2,\ \dots\ , q_n)$. Each oriented $arc(i,j)$ is associated with non-negative

[1] University of Pardubice, Department Automation and Mathematics, Studentská 95, 53210 Pardubice, jaromir.zahradka@upce.cz.

values d_{ij} of travel distance (in meters), and c_{ij} of travel time (in sec) from node i to node j. For easier reference we assume, $I = \{1, \dots, n\}$, and $I_D = I \cup \{n+1, \dots, n+s\}$. The distance matrix $\boldsymbol{D} = \left(d_{ij}\right)_{i,j \in I_D}$ and time distance matrix $\boldsymbol{C} = \left(c_{ij}\right)_{i,j \in I_D}$, both matrices can be generally asymmetrical.

The VRPMSR consists of finding a collection of simple cycles, each corresponding to a vehicle route with minimal sum of the distances of the cycle arcs, such that:
a) each cycle starts and ends in the same warehouse,
b) any number of cycles can start and end in each warehouse.
b) each vertex (customer) $j \in I$ is visited by only one cycle exactly once,
c) the sum of delivered containers during each cycle does not exceed the vehicle capacity Q.

The delivered containers contain goods with only higher or only lower delivery priority, e.g. frozen goods (priority 1) and chilled goods (priority 2). As a result, all customers are divided according to the priority of the delivered goods too. The priorities of customers are arranged in a row vector $\boldsymbol{p} = (p_1, p_2, \dots, p_n)$.

We assume that for each customer $i \in I$, the variable m_i denotes the service time (in seconds) associated with unloading goods and dealing with the customer. The customer service times are arranged in a row vector $\boldsymbol{m} = (m_1, m_2, \dots, m_n)$.

The solution of the VRPMSR is a matrix $\boldsymbol{X} = \left(x_{ij}\right)_{ij \in I_D}$ of binary elements 0 or 1. The value $x_{ij} = 1$ means that the arc from i node to j is included in one cycle, and the value $x_{ij} = 0$ means that the corresponding arc is not included in any one cycle. The number of all x_{ij} variables is $(n+s)^2$.

1.3 An Outline Solution of the VRPMSR

The aim is to propose a practically applicable heuristic solution to the VRPMSR problem. The propose solution consists in dividing the customers into s groups, where each group of customers will be served from one warehouse. The initial division into groups can be created according to the nearest warehouse, and for this, a distance matrix can be used. The customer groups can be adjusted according to other indicators, for example, so that the group corresponding to each warehouse includes as many closest customers as possible to achieve full utilization of the trucks. In each group the exact local solution of the vehicle routing problem is performed using software taken from (Zahrádka 2023; Zahrádka 2024). The software presented in (Zahrádka 2023; Zahrádka 2024) is involved in the solution of the VRPMSR in Matlab code.

The essence of the proposed solution is that in each created customer group around one relevant warehouse, local optimization is performed. The partial s results $\boldsymbol{X_k}, k = \{1, \dots, s\}$ are then merged into the resulting matrix $\boldsymbol{X}$. During preparation for local optimization, the local distance and local time distance matrices $\boldsymbol{D_k}$ and $\boldsymbol{C_k}$, are created by omitting unnecessary rows and columns of the matrices $\boldsymbol{D}$ and $\boldsymbol{C}$. Similarly, by omitting the relevant elements of the vectors $\boldsymbol{q}$, $\boldsymbol{p}$ and $\mathbf{m}$ the corresponding vectors $\boldsymbol{q_k}$, $\boldsymbol{p_k}$ and $\boldsymbol{m_k}$ are created for the local solutions.

1.4 Implementation of the VRPMSR Solution in Matlab Code

The Matlab implementation of the VRPMSR solution presented above is contained in the M-function *VRPMSR_SOLVER.m* which is listed in the Appendix of this article.
The M-function *VRPMSR_SOLVER.m* is started using a special setup M-script, which is not published here. Using the setup M-script, all input data are introduced into the Matlab environment. They are the variables *n, s, D, C, Q, q, p, m, xE, yN, E0, N0.*

On the line No. 2, the initiatory values of the output variables *X, y, TotLgth, TotDur*, are introduced. Using the commands on the line No. 3, the vertices are sorted into groups according to the closest warehouses thus creating the vector g. The commands on lines No. 4 to No. 18 create data files for performing optimizations in local groups around individual warehouses.

Inside the for-cycle, which is created by the commands on lines No. 19 to No. 41, the optimization in individual groups is performed. The M-function VRPMW_SOLVER (line No. 29) is used to perform local optimizations. This M-function was published by the author in (Zahrádka 2024). During the above for-cycle, the output values *X, y, TotLgth, TotDur* are gradually created using the commands on lines No. 33, 34, 41.

Drawing a corresponding figure with all cycles that are the global solution of the VRPMSR is provided by calling the M-function *Plot_VRP_MME25* in the line No. 43. This M-function *Plot_VRP_MME25* is not published here.

The total distance *TotLgth* (km) traveled by the vehicles and the total time *TotDurHour* (hhmmss) of the vehicles' journey are calculated and displayed using the commands on lines No. 44 to No. 46.

2 Applied Tasks

In the east region of Czech Republic and in the surrounding area, 3 warehouses and 27 customers were selected for the applied task demonstration. In Table 1 the customer and warehouse locations GPS coordinates E_i and N_i are listed. In the table, q_i means the number of containers for the customer i and the value p_i specifies the priority of goods for customer i. The value $p_i = 1$ means higher priority of delivered goods and $p_i = 2$ means the lower one. The customer and store locations can be seen in Figure 1. The considered cargo capacity of each truck is assumed to be 18 containers. The service time m_i at each customer's location is assumed by 20 min and plus 60 sec for each unloaded container q_i, i.e. $m_i = (20 + q_i)\ 60$.

	Customers									
i	1	2	3	4	5	6	7	8	9	10
E_i (°)	16.24160	16.08722	14.69055	15.27025	16.66231	16.98885	14.90985	17.11870	15.58698	14.85888
N_i (°)	50.57513	50.27706	49.41570	49.95616	49.72949	49.95986	50.42491	49.46988	49.61639	49.99393
q_i	7	1	5	4	4	3	3	6	3	3
p_i	1	2	1	2	1	2	1	2	1	2
	Customers									
i	11	12	13	14	15	16	17	18	19	20
E_i (°)	14.74052	15.63217	15.36320	15.598478	15.30779	15.04006	15.22476	16.65998	15.99651	15.90839
N_i (°)	50.15575	50.35722	50.43471	49.42432	50.73916	50.76432	49.42156	49.49256	49.84431	50.02656
q_i	7	1	1	7	1	5	1	6	3	5
p_i	1	2	1	2	1	2	2	2	1	1
	Customers							Stores		
i	21	22	23	24	25	26	27	28	29	30
E_i (°)	14.4257	16.5257	16.4257	16.1565	16.3101	17.3520	17.3535	14.6359	16.2264	16.7487
N_i (°)	49.7105	50.0569	50.8749	50.7268	49.6891	50.8844	50.4806	51.0259	49.3061	50.6544
q_i	3	5	1	1	3	5	3	-	-	-
p_i	1	2	1	2	1	2	1	-	-	-

Table 1 The GPS coordinates E_i , N_i of customers/stores locations, numbers of containers $q_i$and priorities p_i.

2.1 The Applied Task VRPMSR with Respecting Priorities

By running the *VRPMSR_SOLVER.m* function with above-given input parameters including given priorities a solution which consists of 7 cycles is found. All cycles can be seen in Figure 1. This means that the delivery of 97 containers to customers can be ensured by seven trucks, and the total distance traveled by vehicles is 2 005.077 km. The sum of driving times of all seven trucks, including the service times is 44:18:16 (hh:mm:ss). The list of customers on the cycles, numbers of delivered containers, priorities, total lengths, and durations can be found in Table 2.

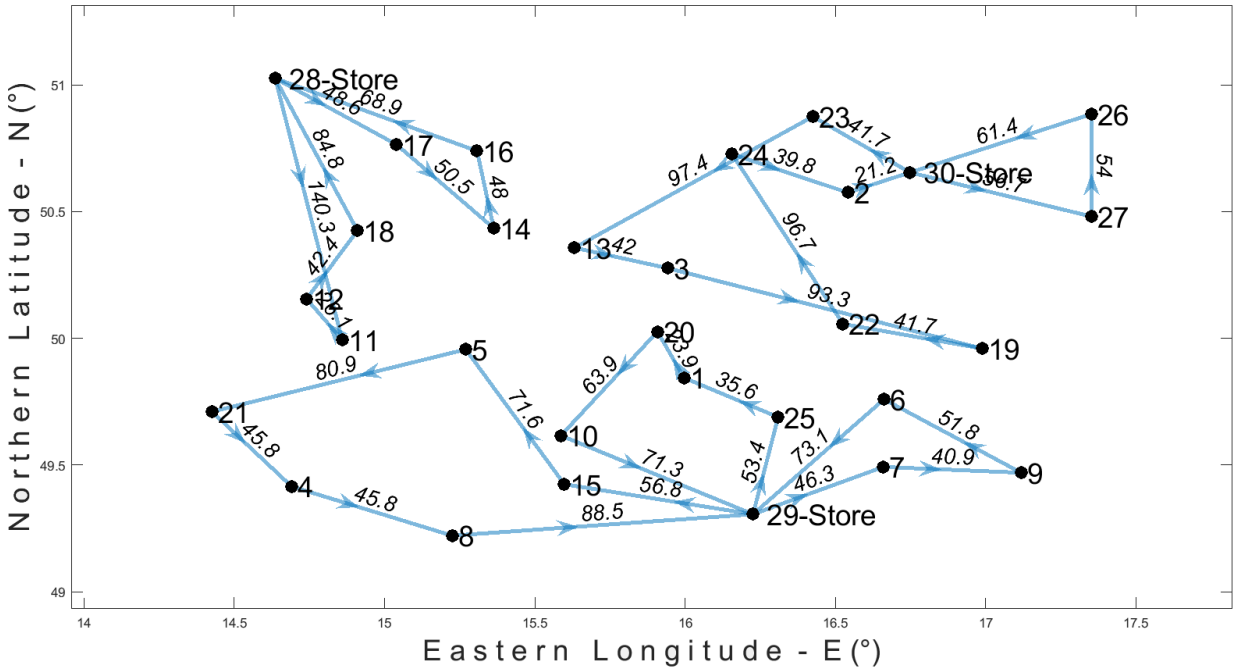


Figure 1 The optimal solution VRPMSR for minimal traveling distance with respect priorities.

No.		Start Store	Customers							Stop Store	Lengths (km)	Duration hh:mm:ss)
1 Cycle	**i**	**28**	**17**	**14**	**16**	**-**	**-**	**-**	**-**	**28**	509.627	11:03:35
	q_i	-	1	7	5	-	-	-	-	-		
	p_i	-	1	2	2	-	-	-	-	-		
2 Cycle	**i**	**28**	**11**	**12**	**18**	**-**	**-**	**-**	**-**	**28**		
	q_i	-	7	1	6	-	-	-	-	-		
	p_i	-	1	2	2	-	-	-	-	-		
3 Cycle	**i**	**29**	**25**	**1**	**20**	**10**	**-**	**-**	**-**	**29**	849.466	18:58:04
	q_i	-	3	7	5	3	-	-	-	-		
	p_i	-	1	1	2	2	-	-	-	-		
4 Cycle	**i**	**29**	**15**	**5**	**21**	**4**	**8**	**-**	**-**	**29**		
	q_i	-	1	1	3	4	6	-	-	-		
	p_i	-	1	1	1	2	2	-	-	-		
5 Cycle	**i**	**29**	**7**	**9**	**6**	**-**	**-**	**-**	**-**	**29**		
	q_i	-	3	3	3	-	-	-	-	-		
	p_i	-	1	1	2	-	-	-	-	-		
6 Cycle	**i**	**30**	**23**	**13**	**3**	**19**	**22**	**24**	**2**	**30**	645.984	14:16:27
	q_i	-	1	1	1	3	5	1	1	-		
	p_i	-	1	1	1	1	2	2	2	-		
7 Cycle	**i**	**30**	**27**	**26**	**-**	**-**	**-**	**-**	**-**	**30**		
	q_i	-	5	5	-	-	-	-	-	-		
	p_i	-	1	2	-	-	-	-	-	-		
										Total	2 005.077	44:18:16

Table 2 The customers cycles by returning vehicles with respecting priorities.

2.2 The Applied Task VRPMSR without Priorities

By running the *VRPMSR_SOLVER.m* function with above-given input parameters excluding priorities a solution which consists of 7 cycles is found. All cycles can be seen in Figure 2. The total distance traveled by vehicles is 1 902.001 km. The sum of driving times of all trucks, including the service times is 42:32:11 (hh:mm:ss). The list of customers on the cycles, numbers of delivered containers, total lengths, and total durations can be found in Table 3. The comparison of tasks with and without prioritization shows that the total distance traveled with prioritization is 103 km longer, i.e. 5.4%, and the total travel time of all vehicles is 1 hour 46 min and 5 sec longer.

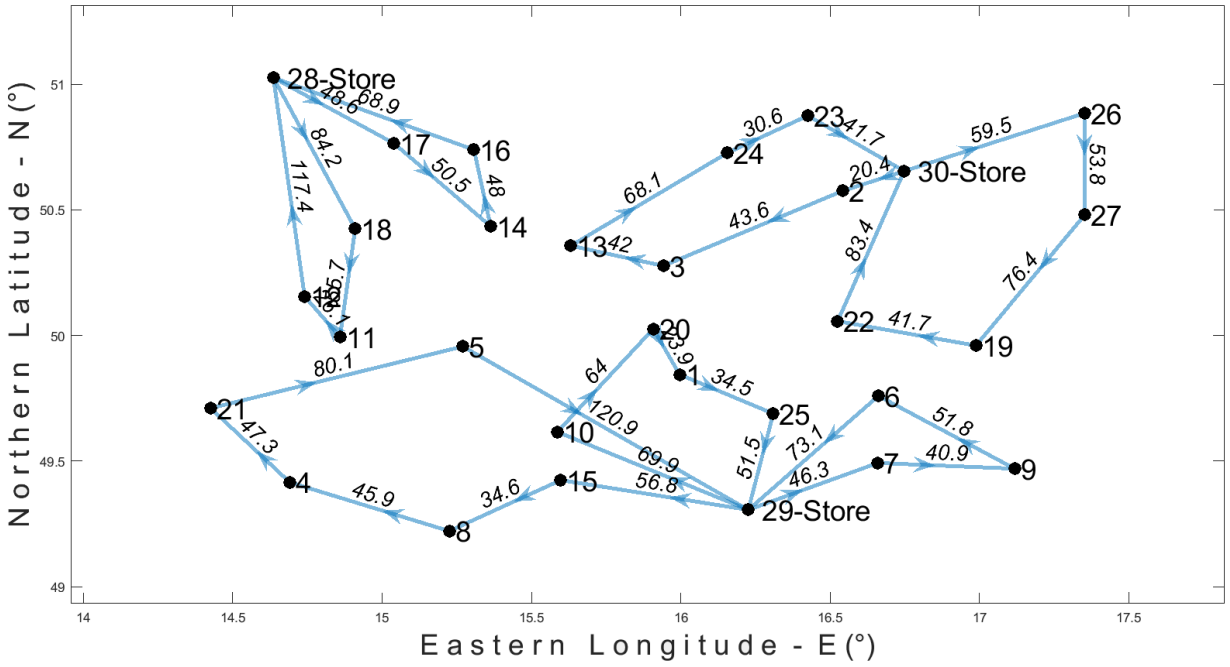


Figure 2 The optimal solution VRPMSR for minimal traveling distance without priorities.

No.		Start Store	Customers							Stop Store	Lengths (km)	Duration hh:mm:ss)
1 Cycle	**i**	**28**	**17**	**14**	**16**	**-**	**-**	**-**	**-**	**28**	499.344	10:53:51
	q_i	-	1	7	5	-	-	-	-	-		
	p_i	-	1	1	1	-	-	-	-	-		
2 Cycle	**i**	**28**	**18**	**11**	**12**	**-**	**-**	**-**	**-**	**28**		
	q_i	-	6	7	1	-	-	-	-	-		
	p_i	-	1	1	1	-	-	-	-	-		
3 Cycle	**i**	**29**	**10**	**20**	**1**	**25**	**-**	**-**	**-**	**29**	841.367	18:48:28
	q_i	-	3	5	7	3	-	-	-	-		
	p_i	-	1	1	1	1	-	-	-	-		
4 Cycle	**i**	**29**	**15**	**8**	**4**	**21**	**5**	**-**	**-**	**29**		
	q_i	-	1	6	4	3	1	-	-	-		
	p_i	-	1	1	1	1	1	-	-	-		
5 Cycle	**i**	**29**	**7**	**9**	**6**	**-**	**-**	**-**	**-**	**29**		
	q_i	-	3	3	3	-	-	-	-	-		
	p_i	-	1	1	1	-	-	-	-	-		
6 Cycle	**i**	**30**	**2**	**3**	**13**	**24**	**23**	**-**	**-**	**30**	561.290	12:49:52
	q_i	-	1	1	1	1	5	-	-	-		
	p_i	-	1	1	1	1	1	-	-	-		
7 Cycle	**i**	**30**	**26**	**27**	**19**	**22**	**-**	**-**	**-**	**30**		
	q_i	-	5	5	3	5	-	-	-	-		
	p_i	-	1	1	1	1	-	-	-	-		
										Total	1 902.001	42:32:11

Table 3 The customers cycles by returning vehicles, and without priorities.

3 Conclusion

The aim to create a functional procedure for solving the of vehicle routing problem for multiple warehouses with prioritizing customers, and with returning of vehicles to the point of departure was fulfilled. The M-script *VRPMSR_SOLVER.m* was implemented in the Matlab code.

We can also compare the obtained result of the task presented with the results obtained in (Zahrádka 2024) , where it is assumed that place of arrival can be any warehouse, including any warehouse other than the point of departure. For otherwise the same assignments, the optimization result in (Zahrádka, 2024) is better. The total distance traveled is in the case of prioritization by 165.492 km shorter (i.e. 8.2%), and the travel time by 2 hours 45 min and 32 sec shorter. The total distance traveled without prioritization is by 212.737 km shorter (i.e. 11.2%), and the total travel time by 2 hours 26 min and 35 sec shorter.

From an economic point of view, it is necessary to consider whether saving the distance traveled is more advantageous than the costs associated with driver accommodation and paying for business trips when traveling between warehouses. The solved comparative tasks show that even with increased distances traveled, it is more advantageous to return all vehicles to their points of departure.

Acknowledgements

The paper was supported by the grant No. CZ.01.1.02/0.0/0.0/21_374/0027244 "Technology development for intelligent traffic flow management - 2nd part - optimization and extension" of Czech Ministry of Industry and Trade.

References

Eksioglu B, Vural AV, Reisman, A (2009) The vehicle routing problem: A taxonomic review. Computers & Industrial Engineering, 57, 1472-1483.

Jonak R, Smutný Z, Simunek M, Dolezel M (2020) Rout and Travel Time Optimization for Delivery and Utility Services: An Industrial Viewpoint. Acta Informatica Pragensia, (2) 9, 200-209.

Math Works. Inc. (2025). Optimalization Toolbox™. User´s *Guide. Natick.*

Winston W L (1994) Operations Research. Applications and Algorithms. Duxbury: Duxbury Press.

Toth P, Vigo D (1998) Exact algorithms for vehicle routing. Boston: Kluwer Academic Publisher.

Zahrádka J (2023) The Exact Solution of Vehicle Routing Problem by Mixed Integer Linear Programming in Matlab. In J. Zahrádka, Proceedings of 41st International Conference on Mathematical Methods in Economics (pp, 433-438). Prague: Czech Society for Operations Research.

Zahrádka J (2024) The Problem of Optimal Delivery of Frozen and Child Goods with Given Priority from Multiple Warehouses. In J. Zahrádka, Proceedings of 42nd International Conference on Mathematical Methods in Economics (pp, 331-336). Prague: Czech Society for Operations Research.

Appendix

```
function [X, y, TotLgth, TotDur] = VRPMSR_SOLVER(n, s, D, C, Q, q, p, m, xE, yN, E0, N0)
X=zeros(n+s); y=zeros(1,n); TotLgth=0; TotDur=0;
for i=1:n D(n+1:n+s,i),[Mg,Ig]=min(D(n+1:n+s,i)), g(i)=n+Ig, end
for i=1:s for j=1:n; if g(j)==n+i IndG(i,j)=j; end; end
verG=[IndG,zeros(s)];
for i=1:s verG(i,n+i)=n+i; end
for i=1:s
    IverG(i,1:n)=sum(IndG)-IndG(i,1:n);
    for j=1:s IverG(i,n+j)=n+j; end
    IverG(i,n+i)=0
end
for i=1:s IverG(i,:)=abs(sort(-IverG(i,:))); end
for i=1:s
    DG(i,:,:)=D; CG(i,:,:)=C; qG(i,:)=q; G(i,:)=m; pG(i,:)=p; Iverkk=IverG(i,:);
    for ip=1:n+s if Iverkk(n+s+1-ip)==0, Iverkk(n+s+1-ip)=[]; end; end
    pom=size(Iverkk); pom(:,1)=[]; iG(i)=pom; pom=[];
    nG(i)=n+s-iG(i)-1;
end
for k=1:s
  DGk=D; CGk=C;
  for i=1:iG(k)
    DGk(IverG(k,i),:)=[]; DGk(:,IverG(k,i))=[];
    CGk(IverG(k,i),:)=[]; CGk(:,IverG(k,i))=[];
  end
  qGk=q; mGk=m; pGk=p;
  for i=s:iG(k)
  qGk(IverG(k,i))=[]; mGk(IverG(k,i))=[]; pGk(IverG(k,i))=[];
  end
  [Xk, yk, TotLgthk, TotDurk] = VRPMW_SOLVER_Z(nG(k), 1, DGk, CGk, Q, qGk, pGk, mGk)
  verGk=verG(k,:);
  for i=1:n+s if verGk(n+s+1-i)==0 verGk(n+s+1-i)=[]; end; end;
  verGk=sort(verGk);
  for i=1:nG(k) y(verGk(i))=yk(i); end
  for i=1:nG(k)+1 for j=1:nG(k)+1 X(verGk(i),verGk(j))=Xk(i,j); end; end
  TotDurHourk=TotDurk/3600; TotLgthk=TotLgthk/1000;
  disp('Total Lengths of Cycles Around k-Store in km:'); disp(TotLgthk);
  TotDurHourk=TotDurk/3600; AllWorkTimehhmmssk=hours(TotDurHourk);
  AllWorkTimehhmmssk=hours(TotDurHourk); disp('AllDistances:');
  AllWorkTimehhmmssk.Format='hh:mm:ss';
  disp('Total Duration Cycles Around k-Store:'),disp(AllWorkTimehhmmssk);
  TotLgth=TotLgth+TotLgthk;TotDur=TotDur+TotDurk;
end
[] = Plot_VRP_MME25(X, n, s, Q, q, t0, C, D, xE, yN, m, TotLgth, TotDur, E0, N0);
TotDurHour=TotDur/3600; TotLgth=TotLgth/1000;
disp('Total Lengths of All Cycles:'); disp(TotLgth);
AllWorkTimehhmmss=hours(TotDurHour); AllWorkTimehhmmss.Format='hh:mm:ss';
disp('Total Duration of All Cycles:'), disp(AllWorkTimehhmmss),
end
```

Received: 2025-05-31 Last revision: 2025-07-09 Accepted: 2025-07-10

Modeling CO_2 Emissions and Economic Growth: A Spatial Panel Approach to the N-Shaped EKC

Michaela Chocholatá[1]

Abstract. This paper examines the relationship between economic growth and environmental degradation in the context of the Environmental Kuznets Curve (EKC) hypothesis, paying particular attention to its N-shaped variant. Using annual panel data for 29 European countries from 2017 to 2022, the study employs spatial panel econometric techniques to address spatial dependence and temporal dynamics. The empirical analysis is based on a spatial panel data model (a fixed effects spatial lag model) estimated via maximum likelihood, which incorporates spatial lags of the dependent variable in order to capture both direct and indirect effects. The results support the existence of an N-shaped EKC, suggesting that, after an initial decline in CO_2 emissions associated with economic growth, emissions may rise again at higher income levels. This development may indicate that efforts to reduce CO_2 emissions have been exhausted. This paper's key contribution lies in its application of a spatial panel data model, which captures spatial dependencies and spillover effects and enhances the EKC's empirical analysis.

Keywords: Environmental Kuznets Curve (EKC), Spatial Econometrics, CO_2 Emissions, Gross domestic product (GDP), Spatial Panel Data Analysis

JEL Classification: O44, C33, C21
AMS Classification: 91B76, 62P20

1 Introduction

In recent decades, the dual imperatives of economic growth and environmental sustainability have attracted considerable global attention. As the adverse effects of climate change and global warming become increasingly evident, the urgency to address environmental degradation has intensified. This global challenge has prompted international cooperation, as evidenced by the establishment of agreements such as e.g., the Paris Agreement (European Commission, 2019), which calls for concerted efforts to reduce carbon emissions and limit global temperature rise.

The Environmental Kuznets Curve (EKC) hypothesis (Grossman and Krueger, 1991) is a central framework for understanding the relationship between economic development and environmental impact. The EKC hypothesises an inverted U-shaped relationship between economic growth and environmental degradation. In the initial phase of economic growth, there is a tendency for environmental quality to deteriorate as a result of increased industrial activity and resource consumption. However, beyond a certain income threshold, further economic growth is associated with improved environmental outcomes, driven e.g., by technological innovation and regulatory frameworks (Wang et al., 2024).

Among the various indicators utilised to test the EKC hypothesis, carbon dioxide (CO_2) emissions are among the most prevalent, given their central role in climate change. Despite the extensive attention this hypothesis has received, the empirical findings remain inconclusive and often contradictory. As highlighted by Kusumawardani and Dewi (2020), the evidence supporting the EKC is largely ambiguous, and no definitive consensus has emerged regarding its universal applicability. Wang, Zhu, and Zhang (2021) analysed the relationship between economic development and municipal solid waste generation in Chinese cities based on EKC models. They confirmed the N-shaped and inverted N-shaped relationships between them in different regions. Ayık and Özer (2025) give the comprehensive literature survey of studies examining the CO_2 – income relationship to assess the validity of the EKC pattern based on various panels of regions/countries using different econometric approaches. Chang and Yu (2023) conducted a comprehensive investigation into impact analysis and its asymptotic inference within spatial autoregressive models, employing both cross-sectional and panel data frameworks. They introduced a range of impact measures to quantify spatial interaction and feedback effects across different spatial dimensions.

[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 objective of the present study is to evaluate the relationship between annual data on total CO_2 emissions per capita and GDP per capita for 29 European countries from 2017 to 2022. This analysis is conducted utilising the estimation of a spatial panel data model (fixed effects spatial lag model) by hypothesizing about the N-shaped relationship.

The remainder of the paper is structured as follows. The second section is dedicated to the methodology employed, while the third section presents the data and the empirical results of the analysis. The final section, section four, offers a conclusion.

2 Methodology

The EKC hypothesis assumes an inverted U-shape in the relationship between environmental degradation and economic growth (quadratic functional form, wherein per capita income and its square are the key explanatory variables in the model). That is, the positive relationship between these variables, observed in the early stages of economic development, turns negative in the later stages of development.

As noted earlier, the non-linear relationship between environmental degradation and economic growth may also be N-shaped (cubic functional form, in addition to per capita income and its square, it also contains the third power of per capita income). This pattern suggests that after an initial decline in emissions - caused by technological and structural changes - emissions rise again. This trend may indicate that efforts to reduce CO_2 emissions have been exhausted or that new sources of environmental burdens are emerging as economic growth continues.

With regard to the empirical part of the paper, we will present the N-shaped relationship for panel data:

$$ln(CO_2)_{it} = \mu_i + \beta_1 ln(\mathrm{GDP})_{it} + \beta_2 (ln(\mathrm{GDP})_{it})^2 + \beta_3 (ln(\mathrm{GDP})_{it})^3 + u_{it} \quad (1)$$

Where $ln(.)$ denotes the natural logarithm of the dependent variable (emissions of CO_2 per capita) and the independent variables (GDP per capita and its powers) – see e.g., Jankiewicz and Szulc (2021), Ayık and Özer (2025). Symbol i is an index for the cross-sectional dimension (i = 1, ..., N), and t is an index for the time dimension (t = 1, ..., T), u_{it} represents an independently and identically distributed error term for i and t with zero mean and variance σ^2. Symbol μ_i denotes a spatial specific effect[2], no spatial interaction effects are included (Elhorst, 2014), β_1, β_2 and β_3 are parameters to be estimated.

In recent years, there has been a marked expansion in the use of spatial panel data models within the spatial econometrics' literature. Spatial panels, which consist of time-series observations across multiple spatial units, such as municipalities, regions, states, or countries, enable more robust modelling of spatial interdependencies and temporal dynamics than traditional cross-sectional approaches (for more information about spatial econometric models for both cross-sectional and panel data see e.g., Elhorst (2010, 2014) and Furková (2013, 2019)).

This paper utilizes spatial panel econometric methods to capture both spatial interdependence and temporal dynamics, thereby enhancing the precision and depth of the analysis of economic relationships. While some models may simply pool time-series cross-sectional data, they more commonly incorporate controls for spatial and/or temporal fixed or random effects. Based on the empirical part of the paper and due to limited space, we only present the fixed effects[3] spatial lag (SAR) model considering the extension of model formulation (1):

$$ln(CO_2)_{it} = \mu_i + \beta_1 ln(\mathrm{GDP})_{it} + \beta_2 (ln(\mathrm{GDP})_{it})^2 + \beta_3 (ln(\mathrm{GDP})_{it})^3 + \lambda \sum_{j=1}^{N} w_{ij} ln(CO_2)_{jt} + u_{it} \quad (2)$$

Where w_{ij} are elements of the spatial weights matrix **W**, capturing the spatial structure and interactions (for specification of spatial matrix, see e.g., Elhorst, 2014)[4], λ is the spatial autoregressive parameter indicating the strength of spatial dependence, μ_i denotes a unit-specific fixed effects and β_1, β_2 and β_3 are parameters to be estimated. The literature identifies two primary methodologies for estimating models that incorporate spatial interaction effects in panel data context: one relies on the maximum likelihood (ML) estimation technique, while the other employs the generalized method of moments (GMM). Thanks to the inclusion of spatial lags of the dependent variable,

[2] Time-specific effects are not taken into account (Elhorst, 2014).
[3] Denotes situation where the spatial specific effects are treated as fixed effects (Elhorst, 2014).
[4] Furthermore, we suppose that the spatial matrix **W** is time invariant and that the panel is balanced.

model (2) allows us to quantify the magnitude of the so-called direct and indirect effects (Chang and You, 2025; Jankiewicz and Szulc, 2021; LeSage and Pace, 2009).

As pointed out by Elhorst (2014), in the estimation of spatial panel models with fixed effects, the parameters can be consistently estimated using ML method in a three-step procedure, beginning with the demeaning of variables to eliminate fixed effects. While the fixed effects themselves (μ_i) can only be consistently estimated when the time dimension (T) is sufficiently large, this limitation does not affect the consistency of the estimated $\boldsymbol{\beta}$ parameters. This is because the estimator of $\boldsymbol{\beta}$ in the demeaned equation is independent of the fixed effects, effectively mitigating the incidental parameters problem. As such, when the primary focus is on $\boldsymbol{\beta}$ rather than on μ_i, the estimation remains robust even in panels with shorter time spans.

3 Data and Empirical Results

The empirical analysis in this study is based on annual panel data for 29 European countries[5] from 2017 to 2022 obtained from two primary sources. First, data on annual total per capita carbon dioxide (CO_2) emissions[6], measured in tonnes, were retrieved from the Our World in Data platform (Global Carbon Budget, 2024). Second, data on GDP per capita in purchasing power parity (PPP, current international dollars) was sourced from the World Bank database (World Bank, 2023). Additionally, the spatial component of the analysis utilized the shapefile *NUTS_RG_20M_2021_3035*, representing European countries at the NUTS 0 level according to the NUTS 2021 classification. This file was downloaded from the Eurostat GISCO database (Eurostat, 2021) and 29 European countries were selected in the open-source software tool GeoDa. Besides the GeoDa, all the analyses were conducted using the software R (packages "sf", "spdep", "splm", "plm", "readxl"), for graphical illustration of the EKC shape the Matplotlib library of Python was used.

Figure 1 presents two quantile maps illustrating the spatial distribution of CO_2 emissions per capita (tonnes per person) on the left, and GDP per capita (PPP, current international dollars) on the right, for the year 2022[7]. On the left-hand map, CO_2 emissions per capita vary significantly across Europe. Countries shown in lighter shades have lower emission intensities, while those in darker hues represent higher emissions per person. The map reveals that low emission levels are not exclusive to either highly developed or less developed countries - both types of economies can be found in the lighter quantiles. In contrast, elevated CO_2 emissions per capita are observed in several central, northern, and eastern European countries. The highest emissions are particularly notable in Luxembourg, followed by the Czech Republic, Poland, Germany, Estonia, Ireland, and Belgium, all appearing in the darkest colour category.

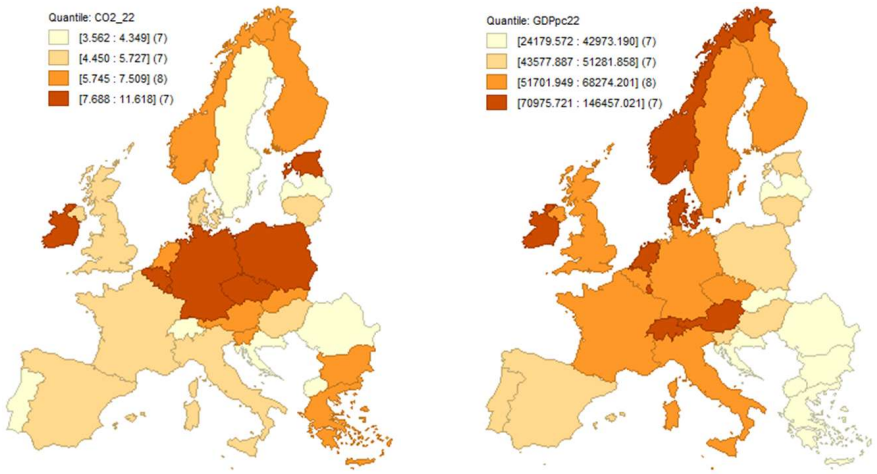


Figure 1 Spatial structure of the CO_2 emissions measured in tonnes per person and GDP per capita, PPP (current international dollars) in 2022, Source: own calculation in GeoDa

[5] Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland and United Kingdom.
[6] Land-use change is not included.
[7] Due to limited space, we are unable to present maps showing the 2017–2021 data.

The right-hand map depicts the distribution of GDP per capita. Here, a clearer regional pattern emerges. Southern and eastern European countries generally fall into the lower quantiles, shown in light colours, indicating lower income levels per person. Central European countries occupy the middle tiers, while the highest GDP per capita levels are found in Luxembourg, Ireland, Norway, Switzerland, Denmark, and Netherlands, which are shaded most darkly. This map suggests a more cohesive clustering of economic development levels across Europe compared to CO_2 emissions.

Building upon evidence of substantial unobserved heterogeneity among countries, the analysis employs a fixed effects modelling strategy. Specifically, the panel dataset comprises 29 European countries observed over six years ($n = 29$, $T = 6$, $nT = 174$), displaying clear structural differences in economic and environmental profiles. Initial pooled maximum likelihood estimation yielded significant GDP terms coefficients (e.g. the cubic term with an estimated coefficient of 0.480, $p < 0.01$)[8], but subsequent spatial dependence diagnostics indicated caution. The standard LM test for spatial lag dependence produced a highly significant result (LM = 20.682, $p < 0.01$) and the LM test for spatial error dependence produced a similarly significant result (LM = 19.080, $p < 0.01$), indicating the presence of spatial interactions. However, robust LM tests, which account for the mutual presence of lag and error dependence, did not identify a dominant form of spatial dependence. This diagnostic profile supports the appropriateness of the fixed effects model as a baseline specification, which allows intercepts to vary across countries while mitigating bias from time-invariant unobserved factors. Assuming an N-shaped relationship between CO_2 emissions and GDP, the fixed effects model controls for unobserved country-specific heterogeneity and provides a more reliable estimation of non-linear income-environment dynamics within the panel context. Furthermore, taking into account the geographical location of individual countries, we estimated the model (2) via ML method. A spatial weights matrix **W** was constructed based on a contiguity criterion, with the queen criterion used to define neighbourhoods. The matrix was then row-standardised, which helps to interpret coefficients as average neighbouring effects. In this study, both the dependent variable and the independent variables were expressed as natural logarithms in their demeaned forms. Table 1 provides the estimates of β_1, β_2, β_3 and λ parameters. All the estimated parameters as well as all calculated impacts were statistically significant. Figure 2 shows the N-shaped EKC, which is based on the $\boldsymbol{\beta}$ parameter estimates in Table 1.

Parameter	Parameter estimate	Direct impact	Indirect impact	Total impact
β_1	73.616 ***	78.363 ***	38.902 ***	117.265 ***
β_2	-6.739 ***	-7.172 ***	-3.561 ***	-10.733 ***
β_3	0.205 ***	0.218 ***	0.108 **	0.326 ***
λ	0.372 ***	-	-	-

Table 1 Parameter estimates and average impact measures - Fixed effects SAR model, Note: Symbols *** and ** indicate the statistical significance at 1 % and 5 % significance level, respectively.
Source: own calculations in software R based on demeaned *ln* data

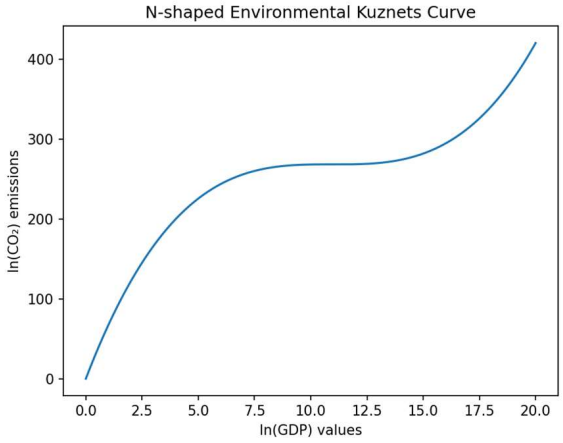


Figure 2 EKC for the considered countries in the period of 2017–2022 (determined based on parameter estimates – Table 1), Source: own calculations in Python software

[8] Results are available from the author on request.

The results of the fixed effects spatial lag model (SAR) estimation demonstrate a statistically significant spatial dependence in CO_2 emissions among the 29 European countries under study. The estimated spatial autoregressive coefficient is $\lambda = 0.372$ ($p < 0.01$), indicating that approximately 37.2 % of the variation in a country's emissions is attributable to the emissions levels of its neighbouring countries. The analysis of the economic determinants reveals a non-linear, N-shaped relationship between economic growth and CO_2 emissions. The coefficient for the natural logarithm of GDP is positive and significant (73.62, $p < 0.01$), suggesting that initial increases in GDP are associated with higher emissions. The negative coefficient for the squared ln(GDP) -6.74 ($p < 0.01$) indicates a subsequent deceleration in emissions growth, while the positive coefficient for the cubed ln(GDP) 0.205 ($p < 0.01$) suggests that at high levels of GDP, emissions begin to rise again.

Further decomposition of the spatial effects reveals that a 1 % increase in GDP leads to an average direct effect of a 78.36 % ($p < 0.01$) increase in emissions, accompanied by an indirect (spillover) effect of 38.90 % ($p < 0.01$) from neighbouring countries, resulting in a total effect of 117.26 % ($p < 0.01$). Similarly, the squared GDP term exerts a direct effect of -7.17 % ($p < 0.01$) and an indirect effect of -3.56 % ($p < 0.01$), with a total effect of -10.73 % ($p < 0.01$), highlighting the diminishing impact of GDP growth at higher levels. The cubed GDP term exhibits a smaller, although statistically significant, positive total effect of 0.326 % ($p < 0.05$). Since a spatial lag model was estimated, the ratio between the direct and total effects is the same for each independent variable, amounting to 66.8%. These findings underscore the critical role of spatial spillovers in shaping CO_2 emissions patterns and emphasize the necessity of accounting for regional interdependencies in the formulation of effective environmental and economic policies.

4 Conclusion

This paper examines the relationship between economic growth and CO_2 emissions for 29 European countries from 2017 to 2022, applying a fixed effects spatial lag (SAR) model. The results confirm significant spatial dependence, with 37.2% of emissions variation explained by neighbouring countries' emissions levels. The findings support an N-shaped relationship between GDP per capita and CO_2 emissions, where emissions initially rise with economic growth, decline at higher income levels, and increase again at very high GDP levels.

Spatial effects were further decomposed into direct and indirect impacts, all of which were statistically significant. On average, 66.8% of the total impact was attributable to direct effects, while the remainder reflected spillover effects from neighbouring countries. These results emphasize the importance of considering spatial spillovers in environmental policy design, highlighting the need for coordinated regional strategies to effectively address emissions reduction.

Acknowledgements

This work was supported by the Grant Agency of Slovak Republic – VEGA grant no. 1/0047/23 „The importance of spatial spillover effects in the context of the EU's greener and carbon-free Europe priority."

References

Ayık U, Özer H (2025) Comparative analysis of the environmental Kuznets curve hypothesis for E7 countries on the relationship between CO2 emissions and economic growth. Environment, Development and Sustainability. https://doi.org/10.1007/s10668-025-06131-5

Chang H-Y, Yu J (2023) Impact analysis for spatial autoregressive models: With application to air pollution in China. Statistica Sinica 33:2693–2714. https://doi.org/10.5705/ss.202021.0119

Elhorst JP (2010) Spatial panel data models. In: Fischer MM, Getis A (eds) Handbook of applied spatial analysis: Software tools, methods and applications, Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03647-7_19

Elhorst JP (2014) Spatial econometrics: From cross-sectional data to spatial panels. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40340-8

European Commission (2019, December 11) The European Green Deal (COM(2019) 640 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0640

Eurostat (2021) NUTS_RG_20M_2021_3035 – Nomenclature of territorial units for statistics (NUTS), 1:20 million scale, EPSG:3035. https://ec.europa.eu/eurostat/web/gisco/geodata/statistical-units/territorial-units-statistics

Furková A (2013) Spatial dependence econometrics models. Ekonomika a informatika 11:41–52

Furková A (2019) Spatial panel data models and estimation possibilities in R. In: Zborník zo seminára Využitie kvantitatívnych metód vo vedeckovýskumnej činnosti a v praxi XIII. Ekonomická univerzita v Bratislave, Vydavateľstvo EKONÓM, Bratislava, pp 29–34. https://fhi.euba.sk/www_write/files/katedry/kove/zborniky/seminar_19_Zuberec-2019.pdf

Global Carbon Budget (2024) Annual CO_2 emissions (per capita). Population data based on various sources (2024) with major processing by Our World in Data. https://ourworldindata.org/co2-and-greenhouse-gas-emissions

Grossman GM, Krueger AB (1995) Economic growth and the environment. The Quarterly Journal of Economics 110:353–377. https://doi.org/10.2307/2118443

Jankiewicz M, Szulc E (2021) Analysis of spatial effects in the relationship between CO_2 emissions and renewable energy consumption in the context of economic growth. Energies 14:5829. https://doi.org/10.3390/en14185829

Kusumawardani AR, Dewi RK (2020) The effect of income inequality on carbon dioxide emissions: A case study of Indonesia. Heliyon 6:e04772. https://doi.org/10.1016/j.heliyon.2020.e04772

LeSage JP, Pace RK (2009) Introduction to spatial econometrics. Chapman and Hall/CRC, New York. https://doi.org/10.1201/9781420064254

Wang K, Zhu Y, Zhang J (2021) Decoupling economic development from municipal solid waste generation in China's cities: Assessment and prediction based on Tapio method and EKC models. Waste Management 133:37–48. https://doi.org/10.1016/j.wasman.2021.07.034

Wang Q, Wang X, Li R, Jiang X (2024) Reinvestigating the environmental Kuznets curve (EKC) of carbon emissions and ecological footprint in 147 countries: A matter of trade protectionism. Humanities and Social Sciences Communications 11:160. https://doi.org/10.1057/s41599-024-02639-9

World Bank (2023) GDP per capita, PPP (current international $). https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

Received: 2025-06-08 Last revision: 2025-08-09 Accepted: 2025-08-12

When Does Reoptimization Pay Off? A Hierarchical Approach to Vehicle Route Planning

Michal Konopa[1], Jana Klicnarová[2,1]

Abstract. In today's fast-paced economic environment, minimizing costs and time is a central concern for operational efficiency and financial sustainability. With the rapid advancement of technology, we now have greater capabilities to optimize and solve increasingly complex logistical problems. This paper focuses on the solution of the Vehicle Routing Problem (VRP) with geographical partitioning, where the goal is to determine multiple delivery routes to ensure optimal delivery routes.
Modern computational tools enable us to solve complex routing problems and dynamically update delivery paths in real time. Reoptimization is a common strategy in VRP to adapt to changes in custom er demand or network structure. However, the conditions under which reoptimization provides substantial benefit remain underexplored. In this study, we present a critical evaluation of reoptimization within a hierarchical VRP framework, where routes are planned on two levels: (i) a fixed backbone infrastructure dividing the customer base, and (ii) weekly local tours serving dynamically selected customers based on demand.

Keywords: Vehicle Routing Problem, Route Reoptimization, Geographical Partitioning, Customer Demand Dynamics, Hierarchical Planning, Simulation-Based Evaluation, Route Stability, Delivery Network Optimization

JEL Classification: C44
AMS Classification: 90B06

1 Introduction

Efficient distribution systems are essential for reducing costs and maintaining service quality in modern logistics. The Vehicle Routing Problem (VRP), which involves determining optimal delivery routes for a fleet of vehicles, remains a fundamental yet computationally challenging problem with wide-ranging industrial applications. Over the past decades, advancements in optimization algorithms and computational tools have significantly improved our ability to address complex routing tasks. However, many practical systems operate under conditions of uncertainty and change, particularly with regard to customer demand and network structure (Toth and Vigo, 2014; Pillac et al., 2013).

In such dynamic environments, route reoptimization, i.e., the periodic revision of existing delivery plans - has become a widely adopted strategy to maintain efficiency under changing operational conditions. Although this approach is widely used in practice, a critical question remains: under what circumstances does reoptimization really provide measurable benefit? In this study, we consider structural network dynamics across weekly planning cycles, rather than real-time or online routing scenarios. Our model assumes full knowledge of weekly demand at the time of route planning. Specifically, how frequently should delivery plans be revised in order to strike an effective balance between routing efficiency and operational stability (Zhang et al., 2022)?

To explore this question, we consider a hierarchical planning framework in which routes are organized on two levels. First, a stable backbone of delivery zones is defined through geographical partitioning of the customer base. Second, within each zone, weekly delivery tours are generated to serve customers with actual demand. This structure reflects many real-world delivery systems and serves as a testbed for our critical evaluation of reoptimization strategies. The simulated environment is inspired by mid-size last-mile delivery systems with a stable depot and fluctuating customer demand, such as parcel distribution or e-grocery delivery. This setting provides a relevant foundation for studying long-term planning strategies without the complexity of real-time dispatching or capacity constraints. Rather than proposing a new routing framework, our focus lies in quantifying the benefits and trade-offs of reoptimization under varying levels of network dynamism.

[1] University of South Bohemia in České Budějovice/Faculty of Agriculture and Technology, Department of Data Science and Computing Systems, Studentska 1668, 370 05 Ceske Budejovice, Czech Republic, konopm05@fzt.jcu.cz.
[2] The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, CZ-182 08, Prague 8, Czech Republic, janaklic@fzt.jcu.cz.

The primary contribution of this paper is a simulation-based study that evaluates the trade-offs associated with different reoptimization frequencies. We systematically analyze how the benefit of reoptimization varies with the volatility of the customer network and quantify the relationship between reoptimization frequency, total distance traveled, and route stability. Our experiments are conducted using two distinct heuristics for zone formation: a widely used OR-Tools-based min-max approach and a custom angular partitioning heuristic. This dual-heuristic evaluation provides a broader perspective on the effectiveness of reoptimization. All scenarios are initially evaluated under a spatially homogeneous customer distribution.

We hypothesize that more frequent reoptimization improves total routing efficiency but reduces structural stability in zone assignments, and that the strength of this trade-off depends on both the selected heuristic and the level of customer volatility.

2 Methodology

We adopt a simulation-based approach to evaluate the impact of route reoptimization in a two-level delivery planning system. The baseline scenario assumes a stable geographical partitioning of the customer base into multiple delivery zones, each served by a dedicated route. These zones are defined once and remain fixed throughout the planning horizon. Within each zone, weekly delivery tours are dynamically generated to serve only those customers who have placed an order in a given week.

2.1 Customer Network and Demand Modeling

Let C denote the set of all customers and $T \in \mathbb{N}$ the number of planning periods (weeks). Each customer $i \in C$ is assigned to a frequency class $f_i \in \{verystable, stable, occasional, rare\}$, which determines their weekly order probability p_{f_i}. Let $C_t \subseteq C$ be the subset of customers who place an order in week $t \in \{1, \dots, T\}$.

The customer network is defined over a fixed depot and a set of customers with geographic coordinates in $\mathbb{R}^2$. To capture temporal demand patterns, we model customer behavior as a stochastic process where each customer $i \in C$ places an order in week t independently with probability p_{f_i}, according to their assigned class. The class assignment f_i may evolve over time following a Markov transition model, calibrated to preserve class proportions in expectation.

The customer base size remains roughly stable, with individual customers entering and exiting over time; however, weekly demand realizations vary due to the stochastic nature of ordering behavior.

2.2 Route Planning and Reoptimization

The delivery planning system consists of a fixed set of zones Z, each associated with a base route. Let $C_{t,z}$ denote the set of customers assigned to zone $z \in Z$ who are active in week $t \in \{1, \dots, T\}$; we assume that $\cup_{z \in Z} C_{t,z} = C_t$ and $C_{t,z1} \cap C_{t,z2} = \emptyset$ for all $z_1, z_2 \in Z$ and all $t \in \{1, \dots, T\}$. For each zone and week, a delivery tour is constructed by solving a Traveling Salesman Problem (TSP) over $C_{t,z}$. The resulting tour length is denoted $d_{t,z}$.

Each week, delivery tours are constructed independently within each zone. We employ a two-step routing procedure: (i) identifying the set of active customers in each zone, and (ii) solving a TSP over these customers using an exact solver with a 1-second time limit. This time-constrained solution is typically near-optimal for moderate instance sizes. Tours are recomputed weekly to reflect the current demand realization.

A reoptimization cycle of length $K \in \mathbb{N}$ determines how often the Z is recomputed. Specifically, reoptimization occurs in weeks $t \in \{1, 1+K, 1+2K, \dots\}$, while base zones remain fixed between such updates. Weekly tours, however, are always updated based on current demand, regardless of K.

By varying the reoptimization frequency K, we assess its impact on total distance traveled and route stability. This evaluation is conducted separately for the OR-Tools-based and angular heuristics.

2.3 Heuristic Approaches for Backbone Route Construction

To construct the backbone delivery zones, we compare two distinct heuristics with different design philosophies. The OR-Tools Min-Cost Routing heuristic formulates the zone construction task as a capacitated multi-vehicle routing problem over active customers, using Google's OR-Tools solver with cost penalties for span and vehicle usage (Perron and Furnon, 2023). The algorithm iteratively increases the number of routes until a feasible solution is found under a strict distance constraint (1000 km per route).

The Angular heuristic employs a geometric sector-based partitioning of the space around the depot. Customers are assigned to angular sectors based on location and expected demand, with sector boundaries dynamically adjusted so that all resulting delivery routes comply with the same strict 1000 km distance constraint as used in the OR-Tools-based approach. This heuristic emphasizes demand balance and spatial contiguity (Sirait and Lubis, 2023; Liu et al., 2023). Unlike OR-Tools, which offers strong global optimization at the cost of interpretability, the angular approach enables explicit evaluation of route stability via sector assignments.

2.4 Evaluation Metrics

The total distance over the planning horizon is computed as:

$$D = \sum_{t=1}^{T} \sum_{z \in Z} d_{t,z}$$

where $d_{t,z}$ denotes the weekly tour length for zone z.

For the angular heuristic, we additionally measure route stability based on the Jaccard similarity between angular intervals assigned to individual customers before and after reoptimization:

$$J_j^{(t,K)} = \frac{\left|A_j^{(t)} \cap A_j^{(t+K)}\right|}{\left|A_j^{(t)} \cup A_j^{(t+K)}\right|}$$

where $A_i^{(t)}$ is an angular sector (interval on the unit circle modulo 2Π) assigned to customer j at time t. The average route stability is then defined as:

$$S_t^K = \frac{1}{N_t} \cdot \sum_{i=1}^{N_t} J_i^{(t,K)}$$

where N_t is the number of customers with defined assignments in both weeks.

Stability is not computed for the OR-Tools heuristic due to the absence of persistent spatial structure in the resulting zones.

2.5 Simulation Framework

The simulation is initialized with a customer base of 500 locations, whose coordinates are sampled from a uniform distribution over a 200 × 200 km square centered at the depot. The depot is placed at fixed coordinates (100,100) and remains static throughout the simulation.

Customers are assigned to frequency classes according to the fixed proportions defined in Section 2.1. Their weekly ordering behavior follows the class-specific probabilities, and class assignments evolve over time according to a fixed transition matrix calibrated to maintain long-term distributional stability (Mardešić et al., 2024).

In our experiments, the initial customer base is distributed as follows: 15% very stable, 35% stable, 35% occasional, and 15% rare customers. Each frequency class is associated with a weekly expected demand rate of 0.25, 0.15, 0.08, and 0.03, respectively, which corresponds to an average of one order every 4, 6–7, 12, and 33 weeks. These values reflect empirical ordering behavior in last-mile delivery environments.

Customer behavior evolves over time through a discrete-time Markov process governing transitions between frequency classes. The transition matrix is calibrated to reflect realistic long-term customer dynamics while

preserving the overall class distribution in expectation. This structure enables the simulation of both short-term demand fluctuations and long-run structural shifts in the network (Mardešić et al., 2024).

Rows and columns correspond (in order) to the four frequency classes: very stable, stable, occasional, and rare. Each row defines the probability distribution of transitions for a customer currently in the corresponding class.

All customers are initially marked as **active** and unassigned - i.e., without a base route. During each reoptimization cycle, all currently active customers are used to construct the backbone route structure. Customers entering the system after the most recent reoptimization remain unassigned until they place their first order, at which point they are matched to an existing backbone route:

- by angular sector inclusion (Angular Heuristic), or
- by proximity to route geometry (OR-Tools Heuristic).

Each week, every active customer has an independent probability p_{exit} of deactivation. Once deactivated, a customer is permanently removed from the simulation and cannot place further orders. We vary p_{exit} across simulation scenarios to study the impact of network volatility on reoptimization effectiveness. To maintain realistic customer turnover dynamics, new customers are introduced weekly, with their number directly dependent on p_{exit}. Specifically, we define a mapping from p_{exit} to an interval $[n_{min}, n_{max}]$ from which the weekly number of new customers is sampled uniformly at random. This mapping ensures that higher exit probabilities are matched with proportionally higher inflow rates, preserving a roughly stable population size while enabling variable levels of customer churn. New customers are placed uniformly in the spatial domain and assigned to frequency classes according to the fixed proportions defined in Section 2.1.

Weekly demand is generated by simulating order realizations for active customers. Those with a valid base route are assigned to their respective routes. Customers without a base assignment—either newly added or previously unmatched—are assigned dynamically using the heuristic-specific rules described above.

For each backbone route, the weekly delivery group includes all currently assigned customers, both persistent and newly matched. A delivery tour is constructed for each group by solving a TSP instance using Google OR-Tools.

A strict limit of 1,000 kilometers is enforced in both heuristics at two levels:

- Backbone route construction: Both heuristics ensure that the base routes respect the 1,000 km limit during initial zone definition. The Angular Heuristic achieves this by rejecting sector placements whose computed delivery length, obtained by solving a TSP instance, exceeds the threshold (see Section 2.3). Similarly, the OR-Tools-based heuristic imposes the constraint directly in the VRP model as a hard upper bound.
- Weekly delivery routing: For each active zone in each week, a TSP instance is solved using Google OR-Tools with a hard constraint on total tour length. If no valid solution exists within the 1-second time limit that satisfies the 1,000 km constraint, the algorithm raises an exception and the assignment is rejected.

This two-level enforcement guarantees that no delivery route violates operational feasibility constraints at any point in the planning horizon. While previous work by Barros et al. (2020) focused on static customer clustering in the Consistent VRP (ConVRP), our setup examines dynamic demand and uniform spatial distribution under reoptimization policies.

2.6 Experimental Scenarios

We evaluate the effect of reoptimization under a grid of experimental configurations that vary in reoptimization frequency and network volatility. The full experimental design is summarized in Table 1.

Each configuration is evaluated over 50 independent simulation runs using different random seeds. All results are reported as averages with corresponding standard deviations.

As a baseline for comparison, we include a non-reoptimized variant in which the backbone structure is computed only once at the beginning of the planning horizon and remains fixed for all 50 weeks.

The current study focuses on spatially homogeneous customer distributions. A heterogeneous variant, based on Gaussian clusters, has been designed but is not included here due to space constraints. Its evaluation is left for future work.

parameter	values	description
Reoptimization interval K	1, 3, 5, 7, 10, 15, 20, 50	Number of weeks between backbone updates. K = 50 represents the static variant.
Network volatility	0.005, 0.01, 0.05, 0.1, 0.2	Weekly probability that a customer exits the network. Entry rate is matched.
Spatial distribution	Homogeneous	Spatial layout of customer locations.
Simulation horizon	50 weeks	Duration of each simulation run.
Number of runs	50	Independent simulation runs per scenario.
Number of vehicles/sectors	5	Set as a target value.
Initial customer count	500	Size of customer base at week 0.

Table 1 Experimental scenarios

3 Results and Discussion

This section presents the results of the simulation experiments conducted under a uniform spatial distribution of customer locations. We report findings separately for the two heuristics used for zone construction: the OR-Tools-based min-max clustering heuristic and the angular partitioning heuristic. In each case, we evaluate the effect of reoptimization frequency and customer network volatility on total distance savings. A comparative discussion follows at the end of the section.

Preliminary results discussed here pertain only to the homogeneous spatial setting. Experiments involving heterogeneous spatial distributions (Gaussian clusters) have been designed but are deferred to future work due to space limitations. We hypothesize that the benefits of reoptimization may be even more pronounced under spatial clustering, where outdated zone structures are more likely to misalign with emerging customer demand patterns.

Each data point in the following figures reflects the average relative distance savings over 50 complete simulation runs (each spanning 50 weeks), compared to the corresponding non-reoptimized baseline.

3.1 Results for the OR-Tools Heuristic

The experiments using the OR-Tools-based backbone heuristic demonstrate that route reoptimization leads to modest but consistent improvements in routing efficiency. Across all tested levels of customer turnover - including high-volatility scenarios with an exit probability of 0.2 - total distance traveled is reduced by up to 2% relative to the non-reoptimized baseline. While the absolute magnitude of savings remains moderate, the trend is robust across simulation runs and stable across reoptimization frequencies.

A closer examination of Figure 1 shows that the efficiency gains depend strongly on the reoptimization interval. The most frequent schedules (e.g., every 1 or 3 weeks) yield the highest savings across all volatility levels, with average reductions consistently above 1.6%. Reoptimization every 5 weeks also performs well, offering a good balance between efficiency and stability. In contrast, infrequent reoptimization (e.g., every 15 or 20 weeks) results in substantially lower savings - often below 1.0% - and converges toward the static baseline.

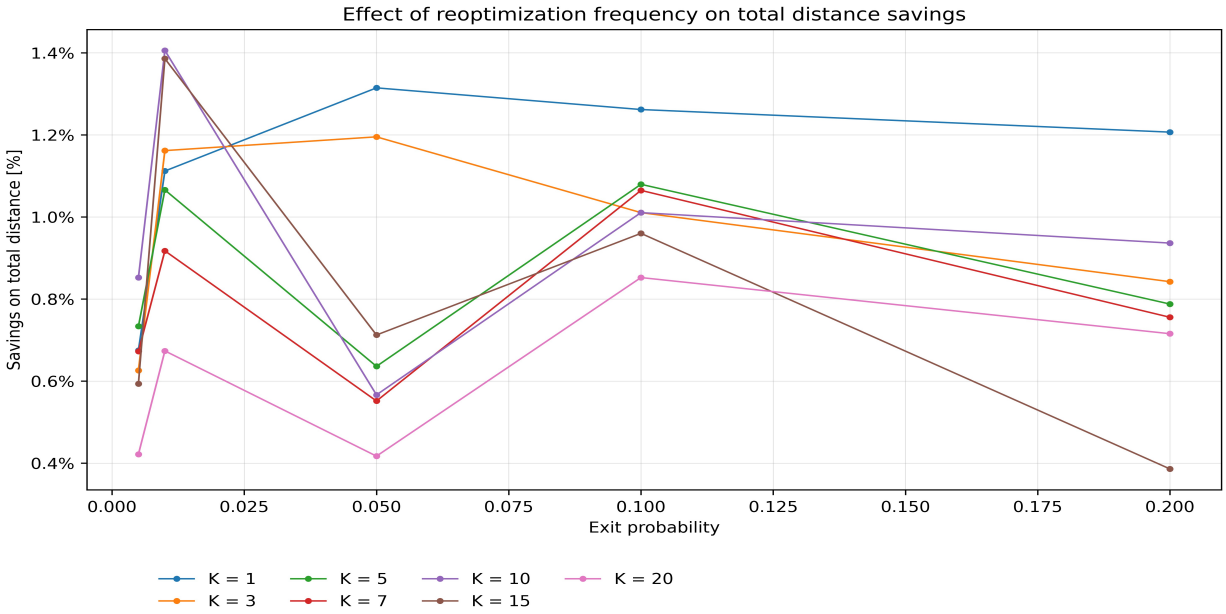


Figure 1 Total distance savings as a function of reoptimization interval and customer volatility (OR-Tools heuristic)

Interestingly, the benefit gap between reoptimization every 1 week and every 3 weeks is small, indicating diminishing returns for very frequent updates. These findings support the use of moderate reoptimization intervals (3 to 5 weeks) in practical settings that aim to improve routing efficiency while minimizing structural disruption.

3.2 Results for the Angular Heuristic

The angular partitioning heuristic exhibits a clear and quantifiable trade-off between efficiency and stability. Reoptimization leads to measurable savings in total distance, particularly in dynamic settings with high customer turnover. For example, with the most frequent reoptimization ($K = 1$), distance savings reach approximately 2.0% at the highest tested exit probability (0.2). As the interval between reoptimizations increases, the benefit diminishes. Values of $K = 10$ and higher yield less pronounced improvements, though still positive under moderate exit probabilities.

These results confirm that a simple geometric partitioning method, when coupled with a capacity-aware placement strategy, can respond effectively to structural changes in customer distribution. At the same time, route stability - as defined in Section 2.4 - consistently decreases with more frequent reoptimization. This behavior confirms the intuitive tension between dynamic responsiveness and customer-zone consistency: frequent reoptimization improves distance metrics but results in lower structural stability. Figure 2 illustrates this dual trend, showing both metrics side by side for each tested configuration.

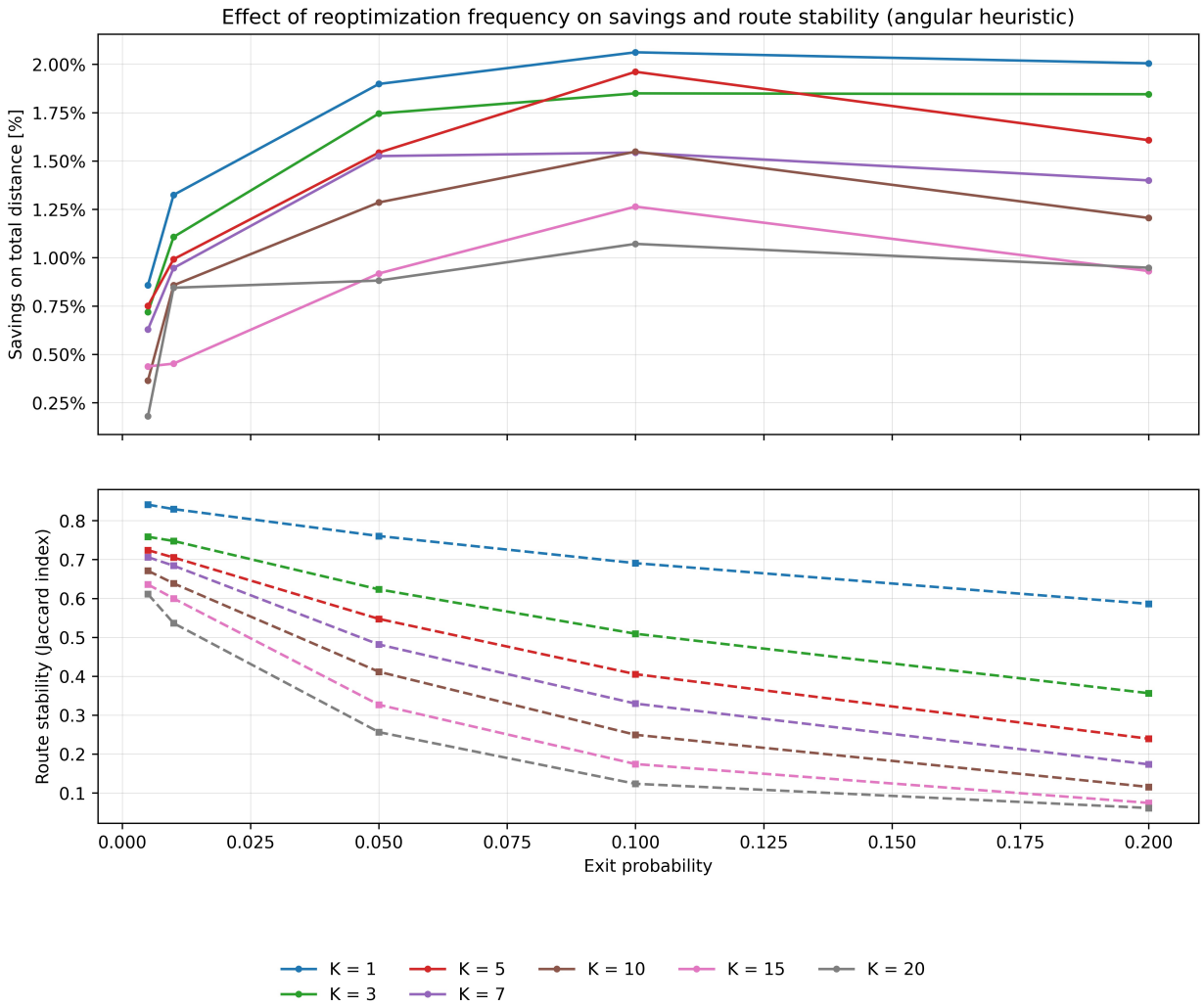


Figure 2 Effect of reoptimization frequency on total distance savings and route stability (angular heuristic)

4 Conclusion

Our findings demonstrate that route reoptimization improves delivery efficiency under dynamic demand, with varying effectiveness depending on the heuristic used. The OR-Tools heuristic shows modest gains (up to ~1.7%), especially under moderate volatility, but its performance tends to plateau or decline with higher customer churn. In contrast, the angular heuristic offers consistent improvements across all tested volatility levels, reaching up to 2.0% savings while enabling interpretable analysis of route stability.

Reoptimization intervals of 3–5 weeks generally provide the best balance between routing efficiency and structural stability. Frequent updates (e.g., weekly) yield diminishing returns, while infrequent updates risk outdated zone structures. These findings suggest that reoptimization frequency should be tuned to match the volatility of the operating environment and the importance of zone continuity.

While both heuristics respect operational constraints, they differ in transparency and robustness. The angular heuristic, despite its geometric simplicity, offers more predictable behavior and better scalability across volatility levels. These features make it particularly attractive for delivery systems requiring interpretable and adaptive planning.

This study focused on spatially homogeneous customer distributions to isolate the effects of reoptimization frequency and volatility. Future work will extend the analysis to heterogeneous spatial layouts, where the interaction between customer clustering and reoptimization strategies may lead to stronger performance differentiation. We also aim to investigate adaptive reoptimization triggers based on observed demand dynamics.

Acknowledgements

The work of Jana Klicnarová was supported by the Czech Science Foundation project GA24-10078S.

References

Barros L, Linfati R, Escobar, JW (2020) An exact approach for the consistent vehicle routing problem (Con-VRP). Advances in Production Engineering & Management, 15(3), 255–266. https://apem-journal.org/Archives/2020/APEM15-3_255-266.pdf

Bent R, Van Hentenryck P (2004) Scenario-based planning for partially dynamic vehicle routing with stochastic customers. Operations Research 52(6): 977–987. https://doi.org/10.1287/opre.1040.0144

Crainic TG, Mancini S, Perboli G, Tadei R (2011) Two-echelon vehicle routing problem: A satellite location analysis. Procedia - Social and Behavioral Sciences 20: 222–231. https://www.sciencedirect.com/science/article/pii/S1877042810010621

Liu F, Lu C, Gui L, Zhang Q, Tong X, Yuan M (2023) Heuristics for Vehicle Routing Problem: A Survey and Recent Advances. arXiv preprint arXiv:2303.04147. https://arxiv.org/abs/2303.04147

Mardešić N, Erdelić T, Carić T, Đurasević M (2024) Review of Stochastic Dynamic Vehicle Routing in the Evolving Urban Logistics Environment. Mathematics 12(1): 28. https://doi.org/10.3390/math12010028

Perron L, Furnon V (2023) OR-Tools: Google's Operations Research Tools. Available at: https://developers.google.com/optimization

Pillac V, Gendreau M, Guéret C, Medaglia AL (2013) A review of dynamic vehicle routing problems. European Journal of Operational Research 225(1): 1–11. https://doi.org/10.1016/j.ejor.2012.08.015

Sirait ZA, Lubis RS (2023) Completion of CVRP with a Combination of Sweep and Local Search Algorithms: Case Study PT AJW. Available at: https://www.researchgate.net/publication/377022749

Sluijk N, Florio AM, Kinable J, Dellaert N, Van Woensel T (2023) Two-echelon vehicle routing problems: A literature review. European Journal of Operational Research 304(3): 865–886. https://doi.org/10.1016/j.ejor.2023.01.015

Toth P, Vigo D (2014) Vehicle Routing: Problems, Methods, and Applications (2nd ed.). SIAM.

Zhang J, Luo K, Florio AM, Van Woensel T (2022) Solving Large-Scale Dynamic Vehicle Routing Problems with Stochastic Requests. arXiv preprint arXiv:2202.12983. https://arxiv.org/abs/2202.12983

Received: 2025-06-23 Last revision: 2025-07-17 Accepted: 2025-07-17

New Negotiation Mechanism in Solving Double Auctions

Petr Fiala[1], Renata Majovská[2] Lorena Caridad López del Río[3]

Abstract. Auctions are important market mechanisms for the allocation of goods and services. In forward auctions, a single seller sells resources to multiple buyers. In reverse auctions, a single buyer attempts to source resources from multiple suppliers, as is common in procurement. Auctions with multiple buyers and sellers are called double auctions and are becoming increasing popular in electronic commerce. It is well known that double auctions in which both sides submit demand or supply bids are much more efficient than several one-sided auctions combined. The paper presents new models and solutions for negotiation mechanism in double auctions.

Keywords: combinatorial auctions, double auctions, iterative approach

JEL Classification: C44
AMS Classification: 90C15

1 Introduction

Auctions are important market mechanisms for the allocation of goods and services (Fiala, 2012). Auction theory has caught tremendous interest from both the economic side as well as the Internet industry. Auctions have emerged as a particularly interesting tool for negotiations. An auction provides a mechanism for negotiation between buyers and sellers. We consider so called combinatorial auctions where bidders can place bids on combinations of items.

There is possible to formulate single-sided combinatorial auctions, forward auctions and reverse auctions. In forward auctions a single seller sells resources to multiple buyers. In reverse auctions, a single buyer attempts to source resources from multiple suppliers, as is common in procurement. Auctions with multiple buyers and sellers, so called double auctions, are becoming increasing popular in electronic commerce. It is well known that double auctions in which both sides submit demand or supply bids are much more efficient than several one-sided auctions combined. The aim of the paper is to propose models and solving approaches for double auctions

The paper presents a model for double auctions. Special cases of the double auction model are single-sided auctions. The formulated combinatorial double auction can be transformed to a combinatorial single-sided auction. The model can be solved by methods for single-sided combinatorial auctions. We propose to use an iterative approach for combinatorial auctions based on primal-dual algorithm.

The rest of the article is organized as follows. Section 2 introduces the elements of combinatorial auctions. Section 3 introduces the basic model of double auctions. A negotiation mechanism for solving double auctions is presented in Section 4. Section 5 presents the conclusions.

2 Combinatorial auctions

Combinatorial auctions (see Cramton et al., 2006; de Vries and Vohra, 2003) are those auctions in which bidders can place bids on combinations of items, so called bundles. The advantage of combinatorial auctions is that the bidder can more fully express his preferences. This is particular important when items are complements. The auction designer also derives value from combinatorial auctions. Allowing bidders more fully to express preferences often leads to improved economic efficiency and greater auction revenues. Most of the auctions studied in the literature are one-sided: either multiple buyers compete for commodities sold by one seller, or, multiple sellers compete for the right to sell to one buyer.

[1] Prague University of Economics and Business, Department of Econometrics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, pfiala@vse.cz.
[2] University of Finance and Administration, Prague, Department of Computer Science and Mathematics, Estonská 500, 101 00 Praha 10, Czech Republic, renata.majovska@mail.vsfs.cz.
[3] University of Seville, Department of Financial Economics and Operations Management, Avenida Ramón y Cajal 1, 41018 Seville, Spain, lcaridad@us.es.

The problem, called the winner determination problem, has received considerable attention in the literature. The problem is formulated as: Given a set of bids in a combinatorial auction, find an allocation of items to bidders that maximizes the seller's revenue. It was introduced many important ideas, such as the mathematical programming formulation of the winner determination problem, the connection between the winner determination problem and the set packing problem as well as the issue of complexity.

2.1 Forward auction

Many types of combinatorial auctions can be formulated as mathematical programming problems. From different types of combinatorial auctions, we present a forward auction of indivisible items with one seller and multiple buyers. Let us suppose that one seller S offers a set R of r items, $j = 1, 2, \dots, r$, to n potential buyers $B_1, B_2, \dots, B_n$ (see Figure 1).

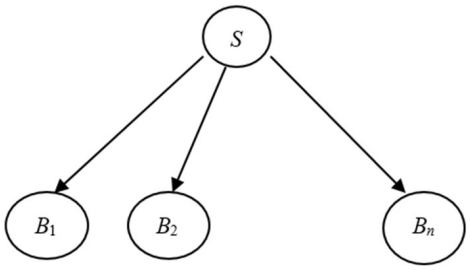


Figure 1 Forward auction

Items are available in single units. A bid made by buyer B_i, $i = 1, 2, \dots, n$, is defined as $b_i = \{C, p_i(C)\}$, where $C \subseteq R$, is a combination of items, $p_i(C)$ is the offered price by buyer B_i for the combination of items C.
The objective is to maximize the revenue of the seller given the bids made by buyers. Constraints establish that no single item is allocated to more than one buyer.
Binary variables are introduced for model formulation:
$x_i(C)$ is a binary variable specifying if the combination C is assigned to buyer B_i $(x_i(C) = 1)$.
The forward auction can be formulated as follows

$$\sum_{i=1}^{n} \sum_{C \subseteq R} p_i(C) x_i(C) \rightarrow max$$

$$\text{subject to} \tag{1}$$

$$\sum_{i=1}^{n} \sum_{C \subseteq R} x_i(C) \leq 1, \forall\, j \in R$$

$$x_i(C) \in \{0, 1\}, \forall\, C \subseteq R, \forall\, i, i = 1, 2, \dots, n$$

The objective function expresses the revenue. The constraints ensure that overlapping sets of items are never assigned.

2.2 Reverse auction

We present a reverse auction of indivisible items with one buyer and several sellers. This type of auction is important for supplier selection problem. Let us suppose that m potential sellers $S_1, S_2, \dots, S_m$ offer a set R of r items, $j = 1, 2, \dots, r$ to one buyer B (see Figure 2).

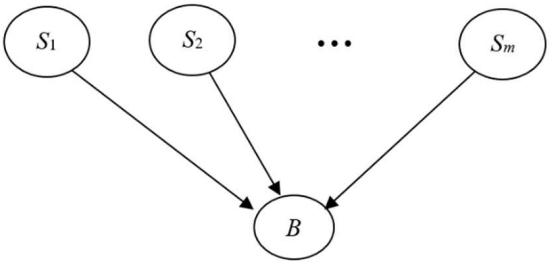


Figure 2 Reverse auction

A bid made by seller S_h, $h = 1, 2, \dots, m$, is defined as $b_h = \{C, c_h(C)\}$, where $C \subseteq R$, is a combination of items, $c_h(C)$ is the offered price by seller S_h for the combination of items C.

The objective is to minimize the cost of the buyer given the bids made by sellers. Constraints establish that the procurement provides at least set of all items.

Binary variables are introduced for model formulation:
$y_h(C)$ is a binary variable specifying if the combination C is bought from seller S_h $(y_h(C) = 1)$.

The reverse auction can be formulated as follows

$$\sum_{h=1}^{m} \sum_{C \subseteq R} c_h(C) y_h(C) \longrightarrow min$$

$$\text{subject to} \tag{2}$$

$$\sum_{h=1}^{m} \sum_{C \subseteq R} y_h(C) \geq 1, \forall j \in R$$

$$y_h(C) \in \{0,1\}, \forall C \subseteq R, \forall h, h = 1, 2, \dots, m$$

The objective function expresses the cost. The constraints ensure that the procurement provides at least set of all items.

3 Double auctions

Auctions with multiple buyers and multiple sellers are becoming increasing popular in electronic commerce. The numerous applications in electronic commerce, including stock exchanges, business-to-business commerce, bandwidth allocation, etc. have led to a great deal of interest in double auctions. Double auctions are not so often studied in the literature as single-sided auctions (see Xia et al., 2005).

For double auctions, the auctioneer is faced with the task of matching up a subset of the buyers with a subset of the sellers. The profit of the auctioneer is the difference between the prices paid by the buyers and the prices paid to the sellers. The objective is to maximize the profit of the auctioneer given the bids made by sellers and buyers. Constraints establish the same conditions as in single-sided auctions.

We present a double auction problem of indivisible items with multiple sellers and multiple buyers. Let us suppose that m potential sellers $S_1, S_2, \dots, S_m$ offer a set R of r items, $j = 1, 2, \dots, r$, to n potential buyers $B_1, B_2, \dots, B_n$ (see Figure 3).

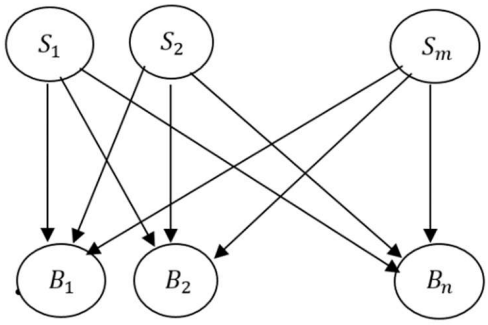


Figure 3 Double auction

A bid made by seller S_h, $h = 1, 2, \dots, m$, is defined as $b_h = \{C, c_h(C)\}$, a bid made by buyer B_i, $i = 1, 2, \dots, n$, is defined as $b_i = \{C, p_i(C)\}$ where $C \subseteq R$ is a combination of items, $c_h(C)$ is the offered price by seller S_h for the combination of items C, $p_i(C)$ is the offered price by buyer B_i for the combination of items C.

Binary variables are introduced for model formulation:

$x_i(C)$ is a binary variable specifying if the combination C is assigned to buyer $B_i (x_i(C) = 1)$,

$y_h(C)$ is a binary variable specifying if the combination C is bought from seller $S_h(y_h(C) = 1)$.

$$\sum_{i=1}^{n} \sum_{C \subseteq R} p_i(C) x_i(C) - \sum_{i=1}^{n} \sum_{C \subseteq R} c_h(C) y_h(C) \longrightarrow max$$

subject to

$$x_i(C) \leq y_h(C), \forall j \in R$$

$$x_i(C) \in \{0,1\}, \forall C \subseteq R, \forall i, i = 1, 2, \dots, n \tag{3}$$

$$y_h(C) \in \{0,1\}, \forall C \subseteq R, \forall h, h = 1, 2, \dots, m$$

The objective function expresses the profit of the auctioneer. The constraints ensure for buyers to purchase a required item and that the item must be offered by sellers. Special case of double auction for one seller is the forward auction and special case of double auction for one buyer is the reverse auction.
The formulated combinatorial double auction can be transformed to a combinatorial single-sided auction. Substituting $y_h(C),\ h = 1, 2, \dots, m,$ with $1 - x_i(C), i = n + 1, n + 2, \dots, n + m,$ and substituting $c_h(C),\ h = 1, 2, \dots, m,$ with $p_i(C), i = n + 1, n + 2, \dots, n + m,$ we get a model of a combinatorial single-sided auction.

$$\sum_{i=1}^{n+m} \sum_{C \subseteq R} p_i(C) x_i(C) - \sum_{i=n+1}^{n+m} \sum_{C \subseteq R} p_i(C) \longrightarrow max$$

subject to

$$\sum_{i=1}^{n+m} \sum_{C \subseteq R} x_i(C) \leq m, \forall j \in R$$

$$x_i(C) \leq y_h(C), \forall j \in R \tag{4}$$

$$x_i(C) \in \{0,1\}, \forall C \subseteq R, \forall i, i = 1, 2, \dots, n + m$$

4 Negotiation mechanism

The model (4) can be solved by methods for single-sided combinatorial auctions. Complexity is a fundamental question in combinatorial auction design. The algorithms proposed for solving combinatorial auctions are exact algorithms and approximate ones. Many researchers consider iterative auctions as an alternative.

One way of reducing some of the computational burden in solving the problem is to set up a fictitious market that will determine an allocation and prices in a decentralized way. We propose solving the problem using a negotiation mechanism, whereby sellers and buyers negotiate prices.

In the iterative approach, there are multiple rounds of bidding and allocation and the problem is solved in an iterative and incremental way. Iterative combinatorial auctions are attractive to bidders because they learn about their rivals' valuations through the bidding process, which could help them to adjust their own bids (see Parkes, 2001). The key challenge in the iterative combinatorial auctions design is to provide information feedback to the bidders after each iteration (see Pikovsky and Bichler, 2005).

There is a connection between efficient auctions for many items, and duality theory. The Vickrey auction can be taken as an efficient pricing equilibrium, which corresponds to the optimal solution of a particular linear programming problem and its dual. The simplex algorithm can be taken as static approach to determining the Vickrey outcome. Alternatively, the primal-dual algorithm can be taken as a decentralized and dynamic method to determine the pricing equilibrium. A primal-dual algorithm usually maintains a feasible dual solution and tries to compute a primal solution that is both feasible and satisfies the complementary slackness conditions. If such a solution is found, the algorithm terminates. Otherwise the dual solution is updated towards optimality and the algorithm continues with the next iteration. The fundamental work (Bikhchandani, Ostroy, 2002) demonstrates a strong interrelationship between the iterative auctions and the primal-dual linear programming algorithms. A primal-dual linear programming algorithm can be interpreted as an auction where the dual variables represent item prices. The algorithm maintains a feasible allocation and a price set, and it terminates as the efficient allocation and competitive equilibrium prices are found.

The primal-dual approach is used for solving. For the problem (4) we will formulate the LP relaxation and its dual. The scheme can be outlined as follows:

1. Choose initial prices for sellers and buyers.

2. Announce current prices and collect bids.

3. Compute the current dual solution by interpreting the prices as dual variables. Try to find a feasible allocation, an integer primal solution that satisfies the stopping rule. If such solution is found, stop and use it as the final allocation. Otherwise update prices and go back to 2.

To keep the record clear, we will illustrate the process of formulating the dual and interpreting the results using the winner determination problem.

For the winner determination problem, we will formulate the LP relaxation and its dual. Consider the LP relaxation of the winner determination problem (1):

$$\begin{gathered}\sum_{i=1}^{n} \sum_{C\subseteq R} v_i(C)x_i(C) \longrightarrow max \\ \text{subject to} \\ \sum_{C\subseteq R} x_i(C) \leq 1, \forall i, i = 1,2,\dots,n \\ \sum_{i=1}^{n} \sum_{C\subseteq R} x_i(C) \leq 1,\ \forall j \in M \\ x_i(S) \geq 0, \forall C \subseteq R, \forall i, i = 1,2,\dots,n\end{gathered} \tag{5}$$

The corresponding dual to problem (2)

$$\begin{gathered}\sum_{i=1}^{n} p(i) + \sum_{j\epsilon C} p(j) \longrightarrow min \\ \text{subject to} \\ p(i) + \sum_{j\in S} p(j) \geq v_i(C)\ \forall i, C \\ p(i), p(j) \geq 0, \forall i, j\end{gathered} \tag{6}$$

The dual variables $p(j)$ can be interpreted as anonymous linear prices of items, the term $\sum_{j\epsilon C} p(j)$ is then the price of the bundle C and $p(i) = \max_{C}\left[v_i(C) - \sum_{j\epsilon C} p(j)\right]$is the maximal utility for the bidder i at the prices $p(j)$.

Following two important properties can be proved for the problems (5) a (6):

1. The complementary-slackness conditions are satisfied if and only if the current allocation (primal solution) and the prices (dual solution) are in competitive equilibrium.

2. The formulation (5)-(6) is weak. For the optimal allocation there no always exist anonymous linear competitive equilibrium prices.

The formulation (5)-(6) can be strengthened. It proves the existence of discriminatory non-linear competitive equilibrium prices for the optimal allocation.

5 Conclusion

The paper is devoted to modeling and solving double auctions. The numerous applications in electronic commerce have led to a great deal of interest in double auctions. Modeling and solving of combinatorial double auctions are a promising subject for research and for practical exploitations. The paper presents a model of double auctions. The proposed combinatorial double auction gives us an opportunity to design modifications of the auction. The formulated model can be transformed to a combinatorial single-sided auction and solved by methods for single-

sided combinatorial auctions. We propose solving combinatorial double auctions using a negotiation mechanism. The primal-dual algorithm can be taken as a decentralized and dynamic method to determine equilibrium. The negotiation process may be modified and supplemented by other procedures.

Acknowledgements

This work was supported by the grant No. IGA F4/18/2024, Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague.

References

Cramton P, Shoham Y, Steinberg R (eds.) (2006) Combinatorial Auctions. MIT Press, Cambridge

Bikhchandani S, Ostroy JM (2002) The package assignment model. Journal of Economic Theory, 107(2), 377–406. DOI: 10.1006/jeth.2001.2957

de Vries S, Vohra R V (2003) Combinatorial auctions: A survey. INFORMS Journal of Computing 15: 284-309. DOI: 10.1287/ijoc.15.3.284.16077

Fiala P (2012) Aukce - teorie a praxe. Professional Publishing, Praha

Parkes D C (2001) Iterative Combinatorial Auctions: Achieving Economic and Computational Efficiency. PhD thesis, University of Pennsylvania

Pikovsky A, Bichler M (2005) Information Feedback in Iterative Combinatorial Auctions. In: Wirtschaftsinformatik 2005. Conference proceedings. Springer, Bamberg

Xia M, Stallaert J, Whinston A B (2005) Solving the combinatorial double auction problem. European Journal of Operational Research 164: 239–251. DOI: 10.1016/j.ejor.2003.11.018

Received: 2025-06-08 Last revision: 2025-08-10 Accepted: 2025-08-12

Dynamic Stochastic General Equilibrium Model with Heterogeneous Expectations for Small Open Economy

Igor Kohút[1]

Abstract. We propose a small open economy model which deviates from traditional DSGE frameworks by allowing agents to form expectations using both rational and subjective mechanisms, reflecting real-world deviations from full rationality. By incorporating heterogeneous expectation formation, the model should better represent observed macroeconomic fluctuations, particularly in response to external shocks. In the empirical section, we perform impulse response simulations, comparing the newly developed model with a standard DSGE model to assess the impact of usage of heterogeneous expectations. The findings suggest that incorporating heterogeneous expectations leads to prolonged adjustment dynamics with observable oscillations during adaptation period. This underscores the importance of expectation heterogeneity in macroeconomic modeling and policy design.

Keywords: DSGE models, heterogeneous expectations, small open economy, rational expectations

JEL Classification: E32, E52, D84, F41
AMS Classification: 91B62, 91B50, 91B84, 37M05

1 Introduction

Dynamic Stochastic General Equilibrium (DSGE) models are a class of macroeconomic models that describe the behavior of economies over the business cycle. Their core feature is microeconomic foundations: households are assumed to maximize expected lifetime utility and firms maximize profits, subject to specific constraints. A standard assumption is rational expectations—agents are presumed to form forecasts based on all available information, including the structure of the model itself.

However, recent crises have shown that the standard DSGE assumption of fully rational, homogeneous agents is unrealistic. Agents often use simplified rules due to cognitive limits (Massaro, 2013; Hommes, 2021; Cornea-Madeira et al., 2019 Recent contributions highlight different ways to relax the full-information rational expectations assumption. One example is a unified modeling framework that encompasses various approaches to expectation formation and learning, providing a basis for analyzing stability, convergence, and policy effects in environments with heterogeneous agents (Evans et al., 2025). Another example is survey-based evidence showing that households' inflation expectations are heterogeneous, with some agents extrapolating recent trends and others expecting mean reversion (Kostyshyna et al., 2025). Building on this literature, the goal of this paper is to develop a small open economy DSGE model with heterogeneous expectations and examine how the presence of boundedly rational agents affects macroeconomic dynamics.

2 DSGE model with heterogeneous expectations

In this section, we construct a custom DSGE model of a small open economy with price rigidities and heterogeneous expectations. This represents a deviation from the standard New Keynesian framework, which typically assumes fully rational agents. The following subsections introduce the model's assumptions and derive the IS and Phillips curves under heterogeneous expectations, leading to the full model specification.

2.1 Model assumptions

The model is based on a New Keynesian small open economy DSGE framework and includes several key assumptions to ensure internal consistency and empirical relevance. The domestic economy consists of a continuum households and a continuum of intermediate-goods-producing firms. Monetary policy is conducted via a Taylor-type rule, consistent with standard New Keynesian formulations (Woodford, 2003), while the economy interacts with the rest of the world through trade and financial flows.

[1] Prague University of Economics and Business, Department of Monetary Theory and Policy, kohi00@vse.cz.

Households maximize their lifetime utility subject to a budget constraint. Their utility depends on consumption and disutility from labor. Consumption is subject to external habit formation, introducing inertia into the model. The household with subjective expectations saves via domestic and foreign bonds and consumes both domestic and imported goods, which are imperfect substitutes, following the Armington aggregation. Firms produce differentiated goods under monopolistic competition and face nominal rigidities in price setting, modeled via Calvo pricing. Each period, only a subset of firms can reset their prices, introducing price stickiness and allowing for non-neutral effects of monetary policy. The production function is Cobb-Douglas with technology subject to exogenous shocks. Inputs include labor and intermediate goods. The model allows for imported intermediate inputs, reflecting the openness of the economy.

Expectations are heterogeneous: a fraction of agents forms rational expectations, using full knowledge of the model structure. The remainder have unspecified expectations other than rational. The term 'subjective expectations' is used throughout the model derivation as a general expression for the expectations held by an individual household, regardless of whether those expectations are rational or not. The exchange rate follows uncovered interest parity (UIP) with potential deviations due to risk premia. The model includes terms-of-trade dynamics and allows for imperfect risk sharing between countries. Overall, the model blends rational structure with behavioral heterogeneity to capture amplification effects and realistic responses to economic shocks in an open economy setting.

2.2 Households Behavior

A household with subjective expectations of type $e \in [0,1]$ maximizes the expected (discounted) present value of utility subject to its budget constraint, that is:

$$max\ \tilde{E}_{e,t} \sum_{s=t}^{\infty} \beta^{s-t}\, U(C_{e,s}, L_{e,s}) \tag{1}$$

$$s.t.\ \int_0^1 P_{D,t}(j) C_{e,D,t}(j) dj + \iint_0^1 P_{i,t}(j) C_{e,i,t}(j) dj\, di + Q_t B_{e,t} = B_{e,t-1} + W_t L_{e,t} + T_t \tag{2}$$

where: $\beta \leq 1$ is discount factor, $\tilde{E}_{e,t}$ is expectation operator based on subjective expectation type e at time t, $C_{e,s}$ is total consumption of the household with subjective expectations e, aggregated from domestic and foreign goods, $L_{e,s}$ is hours worked by the household with subjective expectations e, $U(C_{e,s}, L_{e,s})$ is utility function, assumed to be continuous and twice differentiable, $P_{D,t}(j)$ price of domestic good $j \in [0,1]$ at time t, $P_{i,t}(j)$ price of good j, produced in country $i \in [0,1]$ at time t, $C_{e,D,t}(j)$ consumption of domestically produced good j by a household with subjective expectation e at time t, $C_{e,i,t}(j)$ consumption of good j from country i by a household with subjective expectation e at time t, Q_t is price of a one-period zero-coupon bond, $B_{e,t}$ amount of one-period bonds purchased by the household with subjective expectation e at time t at price Q_t (the household receives the face value at $t+1$), where the transversality condition holds: $\lim_{s\to\infty} \tilde{E}_{e,t} B_{e,s+1} = 0$, W_t nominal hourly wage at time t, T_t amount of transfers received by the household with subjective expectation e at time t.

The first term in the budget constraint denotes expenditure on domestic goods, and the second term refers to expenditure on foreign goods. The specific variables in the integrals will be described and explained in detail later in the thesis. The condition $\lim_{s\to\infty} \tilde{E}_{e,t} B_{e,s+1} = 0$ implies that while the household is allowed to borrow to maximize its utility, it must not leave any debt outstanding at the end of its life. It is assumed that each household consists of a continuum of agents employed across firms and that income is pooled within the household. This allows for insurance against Calvo risk. Let us further assume that a household with subjective expectations e has the following utility function at time t, with parameters $\sigma, \varphi > 0$:

$$U(C_{e,t}, L_{e,t}) = \frac{C_{e,t}^{1-\sigma}}{1-\sigma} - \frac{L_{e,t}^{1+\varphi}}{1+\varphi} \tag{3}$$

The household with expectation type e thus faces two optimization problems: determining the optimal consumption structure and the optimal level of total consumption. First, we derive the optimal composition of consumption for the household with expectation e at time t. In this model, we do not assume a single representative household, but rather multiple representative households that share the same expectation type e. However, it is also possible that a given household does not share its expectations with any other, so the derivation is presented in a general form, allowing for heterogeneity in subjective expectations across households.

Let us assume that the total aggregated consumption of a household with subjective expectation e at time t, $C_{e,t}$, can be expressed as:

$$C_{e,t} = [(1-\alpha)^{\frac{1}{\eta}} C_{e,D,t}^{\frac{\eta-1}{\eta}} + \alpha^{\frac{1}{\eta}} C_{e,Z,t}^{\frac{\eta-1}{\eta}}]^{\frac{\eta}{\eta-1}} \tag{4}$$

Where α is degree of openness of the economy (share of foreign goods in domestic consumption), η is elasticity of substitution between domestic and foreign goods, $C_{e,D,t}$ consumption of domestic goods at time t by the household with subjective expectation e, $C_{e,Z,t}$ is consumption of foreign goods at time t by the household with subjective expectation e.

In the model, each household seeks to optimize its intertemporal utility given a budget constraint and a specific expectation rule. The process begins by decomposing total consumption into domestic and foreign components using a CES aggregator, incorporating the openness of the economy and the elasticity of substitution between goods. Next, the household determines the optimal consumption allocation between domestic and foreign goods by minimizing expenditure while maintaining a fixed utility level. This yields the standard demand functions based on relative prices and preference parameters. Subsequently, the household's intertemporal decision problem is solved. Through first-order conditions, we derive the Euler equation that governs optimal consumption over time. This equation captures the trade-off between present and future consumption, adjusted for interest rates, price levels, and subjective expectations.

By incorporating habit formation into the utility function and assuming heterogeneous expectations, the derived consumption rule accounts for behavioral persistence and cognitive differences among agents. This leads to a forward-looking consumption decision, influenced by expected future variables and shaped by the type of expectations (rational or heuristic) employed by the household. The final outcome is an individual consumption rule that links current consumption not only to observable variables but also to the expectations and perceived performance of different forecasting strategies. This rule plays a crucial role in driving the aggregate dynamics of the model. The consumption rule for a household with subjective expectation e at time t in log-linearized form is as follows:

$$c_{e,t} = -\frac{\beta}{\sigma}\tilde{E}_{e,t}\sum_{s=t}^{\infty}\beta^{s-t}\,(i_s - \pi_{s+1} - \rho) + (1-\beta)\left(\frac{\varphi+1}{\varphi+\sigma}\right)\tilde{E}_{e,t}\sum_{s=t}^{\infty}\beta^{s-t}\,(w_s - p_s) \tag{5}$$

Where i_s is short-term nominal interest rate, ρ is the discount rate reflecting households' willingness to postpone consumption into the future, π_t is inflation rate in time t, w_t is log-deviation of variable W_t from its steady state, p_t is log-deviation of variable P_t from its steady state.

After deriving the individual consumption rules for households with subjective expectations, the next step is to ensure consistency at the aggregate level. The resulting market-clearing equation links individual behaviors to macroeconomic outcomes, forming a key equilibrium condition of the model. In the end, we obtain the IS curve with heterogeneous expectations for a small open economy:

$$\begin{aligned} y_t = &-\frac{\beta}{\sigma}\tilde{E}_t\sum_{s=t}^{\infty}\beta^{s-t}\,(i_s - \pi_{s+1} - \rho) + (1-\beta)(\varphi+1)\tilde{E}_t\sum_{s=t}^{\infty}\beta^{s-t}\,y_s \\ &+\left(\frac{\alpha\omega}{\sigma}s_t - \alpha\omega(1-\beta)\left(\frac{\varphi+1}{\varphi+\sigma}\right)\tilde{E}_t\sum_{s=t}^{\infty}\beta^{s-t}\,s_s\right) - \varphi(1-\beta)\left(\frac{\varphi+1}{\varphi+\sigma}\right)\tilde{E}_t\sum_{s=t}^{\infty}\beta^{s-t}\,a_s \end{aligned} \tag{6}$$

Where ω is a constant, s_t is log-deviation of variable S_t, which represent terms of trade.

2.3 Firms Behavior

A core assumption of the model is that each domestic firm produces a differentiated good $j\in[0,1]$ under monopolistic competition. The model assumes a simplified Cobb-Douglas production function without capital, where output depends on the level of technology and labor. Firms face identical real marginal costs (in domestic prices), expressed in log-linearized form as:

$$mc_t = -\upsilon + w_t - p_{D,t} - a_t \tag{7}$$

where $\upsilon = -ln(1-\tau)$ represents government support per employee.

Due to price stickiness arising from long-term contracts, customer retention motives, and the costs of price changes, firms do not adjust prices instantly in response to market conditions. The Calvo pricing mechanism (Calvo, 1983) is used to capture this rigidity, assuming firms adjust prices probabilistically over time.

Each firm has subjective expectations of type $e\in[0,1]$ and employs labor supplied by households. We also assume that households ultimately control firms (as owners or through appointed CEOs), implying that firms discount future profits using the same subjective expectations as households. This allows the use of the stochastic discount factor Q_t in the firm's intertemporal profit maximization. Firms are assumed to maximize their expected discounted profits subject to a demand constraint and their production technology, which depends on labor and productivity shocks. The combination of monopolistic competition, Calvo-type price stickiness, and

heterogeneous expectations leads to a modified version of the New Keynesian Phillips Curve (NKPC). This Phillips curve incorporates expectation heterogeneity into the pricing decision, capturing how different belief systems among firms shape inflation dynamics. The final expression in log-linearized form - the NKPC under heterogeneous expectations - presented below plays a central role in the model's dynamic behavior.

$$\pi_t = \kappa \tilde{E}_t \sum_{s=t}^{\infty} (\theta\beta)^{s-t} \tilde{y}_s + \xi \tilde{E}_t \sum_{s=t}^{\infty} (\theta\beta)^{s-t} (\pi_{s+1} - \alpha \Delta s_{s+1}) + \alpha \Delta s_t \tag{8}$$

2.4 Deriving IS and NKPC with Heterogeneous Expectations

This final modeling section consolidates the behavioral foundations of households and firms into a coherent macroeconomic framework by deriving two key equilibrium conditions: the IS curve and the New Keynesian Phillips Curve (NKPC), both modified to incorporate heterogeneous expectations.

Starting from the consumption rules of heterogeneous households and the intertemporal optimality condition for bond holdings, we derive the IS curve that links output to expected future output and the real interest rate. Heterogeneous expectations are modeled as a combination of rational and boundedly rational forecasting rules. A fixed share of agents forms model-consistent (rational) expectations, while the remaining agents rely on backward-looking heuristic predictors, updating their forecasts as a weighted average of previous expectations and observed realizations. The weights evolve endogenously according to the relative accuracy of competing predictors, following the continuous-choice mechanism of Brock et al. (2005), Diks & van der Weide (2005), and Massaro (2013). As a result, the IS curve incorporates a weighted average of subjective expectations rather than a single rational forecast, capturing belief dispersion, gradual learning, and endogenous expectation switching.

Simultaneously, firm behavior under Calvo pricing yields the NKPC. Firms choose prices based on discounted expected future marginal costs, but due to expectation heterogeneity, each type of firm forms its own forecast about future inflation and marginal costs. This leads to an aggregate Phillips curve where inflation depends not only on current marginal costs but also on a weighted composition of different inflation expectations across firms.

Together, the IS and NKPC equations form the backbone of a small open economy DSGE model with behavioral heterogeneity. These equations capture how belief-driven dynamics—rather than purely rational expectations—can amplify or dampen responses to shocks. The full system of model equations in log-linearized form is presented below and provides the basis for the simulation analysis that follows.

$$\begin{aligned} \tilde{y}_t = {} & \Omega_{yit} i_t + \Omega_{y\pi t1}(E_t \pi_{D,t+1} - r_t^n) + \Omega_{yyt1} E_t \tilde{y}_{t+1} \\ & + \Omega_{y\pi t}(i_{t-1} + \pi_{D,t-1} - r_{t-1}^n) + \Omega_{yyt} \tilde{y}_{t-1} \end{aligned} \tag{9}$$

$$\pi_{D,t} = \Omega_{\pi yt} \tilde{y}_t + \Omega_{\pi\pi t1} E_t \pi_{D,t+1} + \Omega_{\pi y1} \tilde{y}_{t-1} + \Omega_{\pi\pi t} \pi_{D,t-1} \tag{10}$$

$$i_t = \Phi_y \tilde{y}_t + \Phi_\pi \pi_{D,t} + e_2 \tag{11}$$

$$r_t^n = \rho + \frac{1}{\varsigma_\pi} (\xi_k + \Omega_{ry*} y_t^* + \varsigma_{ws\Delta}^* E_t y_{t+1}^* + \Omega_{ra} a_t + \varsigma_{as\Delta} E_t a_{t+1}) \tag{12}$$

$$y_t^* = \rho_y y_{t-1}^* + e_1 \tag{13}$$

$$a_t = \rho_a a_{t-1} \tag{14}$$

Where $\pi_{D,t}$ is domestic inflation rate, $\tilde{y}_t$ is the output gap, y_t^* denotes foreign output, r_t^n is the natural interest rate. The variables e_1, e_2 represent exogenous stochastic shocks. Ω_{yit}, $\Omega_{y\pi t1}$, Ω_{yyt1}, $\Omega_{y\pi t}$, Ω_{yyt}, $\Omega_{\pi yt}$, $\Omega_{\pi\pi}$, $\Omega_{\pi y}$, $\Omega_{\pi\pi t}$, Φ_y, Φ_π, ξ_k, Ω_{ry*}, $\varsigma_{ws\Delta}^*$, Ω_{ra}, $\varsigma_{as\Delta}$, ρ_y, ρ_a are structural parameters or policy coefficients, calibrated in the simulation section

3 Simulation of the SOE DSGE model with heterogeneous expectations

In this section, we conduct an impulse response simulation using the model derived in equations (9) to (14). We consider three types of exogenous shocks: a foreign (world economy) shock e_1, and a monetary policy shock e_2. The analysis focuses on how these shocks affect the economy under different shares of rational agents. Several parameter values are calibrated in line with standard small open economy DSGE literature. Following the approaches of Massaro (2013), Galí and Gertler (1999), and Smets and Wouters (2003), we consider two alternative shares of rational agents - 35% and 100% - for each type of shock. The case with 100% of agents forming rational expectations corresponds to the standard New Keynesian small open economy model with price stickiness and

homogeneous rational expectations, which serves as the benchmark in our analysis. The simulations were carried out using code written in Dynare.

Table 1 presents the values of structural parameters used in the simulations for the case where 35% of agents form rational expectations. Similarly, *Table 2* shows the parameter values for the benchmark case, where all agents form rational expectations.

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
$\Omega_{\pi\pi t1}$	0,990	Ω_{yit}	-1,794	Ω_{yyt}	0	Ω_{ry*}	-0,433
$\Omega_{\pi y1}$	0	$\Omega_{y\pi t1}$	0,308	$\Omega_{\pi yt}$	1,804	$\varsigma^*_{ws\Delta}$	0,433
$\Omega_{\pi\pi t}$	0	ρ_y	0,975	Ω_{yyt1}	1,028	Ω_{ra}	0,126
Φ_y	0,950	ρ_a	0,950	$\Omega_{y\pi t}$	0	$\varsigma_{as\Delta}$	0,154
Φ_π	0,975	ξ_k	-0,004				

Table 1 Values of structural parameters where 35% of agents form rational expectations

Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
$\Omega_{\pi\pi t1}$	0,347	Ω_{yit}	0,084	Ω_{yyt}	-3,027	Ω_{ry*}	-0,433
$\Omega_{\pi y1}$	1,161	$\Omega_{y\pi t1}$	-0,004	$\Omega_{\pi yt}$	0,632	$\varsigma^*_{ws\Delta}$	0,433
$\Omega_{\pi\pi t}$	0,644	ρ_y	0,975	Ω_{yyt1}	0,014	Ω_{ra}	0,126
Φ_y	0,950	ρ_a	0,950	$\Omega_{y\pi t}$	18,874	$\varsigma_{as\Delta}$	0,154
Φ_π	0,975	ξ_k	-0,004				

Table 2 Values of structural parameters where 100% of agents form rational expectations

3.1 Impulse Response to a Foreign Economy shock

Figure 1 displays the impulse response functions to a foreign economy shock for two cases: when the share of agents forming rational expectations is 35% and when it is 100% (the benchmark). In the benchmark scenario, a positive foreign shock leads to an increase in both actual and potential output, a narrowing of the output gap, and a decline in inflation. In response, the central bank reduces the nominal interest rate to stabilize inflation. All endogenous variables gradually converge to their steady-state values. The situation changes when the model incorporates boundedly rational agents. In this case, the foreign shock leads to a decline in domestic output and an increase in inflation, with a noticeable divergence from the steady state. Consequently, monetary policy reacts in the opposite direction, with the nominal interest rate increasing. As the share of boundedly rational agents rises, we observe oscillatory dynamics in the endogenous variables, which reflect the backward-looking nature of these agents' expectation formation. While all variables eventually converge to their steady-state values, the adjustment process is significantly slower compared to the purely rational expectations case.

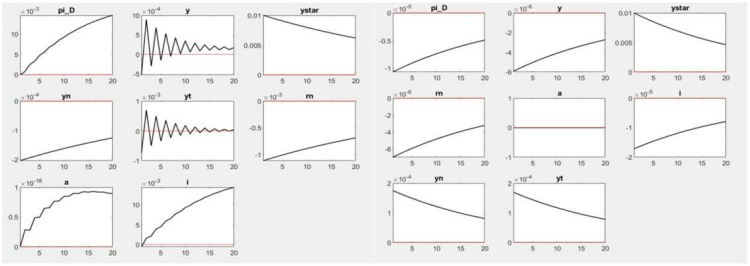


Figure 1 Response of the endogenous variables to a foreign economy shock

3.2 Impulse Response to a Monetary Policy Shock

Figure 2 illustrates the impulse response functions to a monetary policy shock, governed by the Taylor rule as defined in equation (11), again for two cases 35% and 100% agents forming rational expectations. In the benchmark case, a positive monetary policy shock leads to a sharp decline in the other endogenous variables, which, however, return to their steady-state values extremely quickly. The dynamics differ in the simulations with boundedly rational agents. As the share of such agents increases, we observe growing oscillations in the responses of endogenous variables. This behavior reflects the backward-looking nature of forecast formation among these agents. Eventually, all variables once again converge to their steady-state values, but the adjustment is significantly delayed compared to the scenario with purely rational expectations.

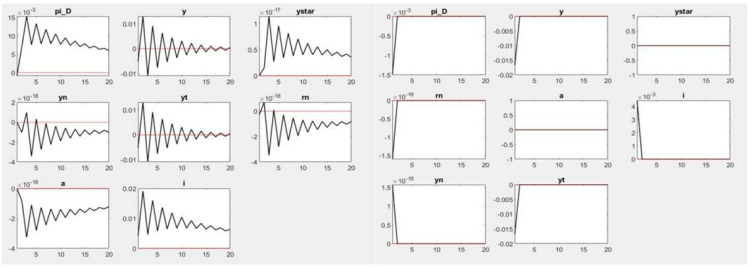


Figure 2 Response of the endogenous variables to a Monetary Policy Shock

4 Conclusion

This paper developed and analyzed a small open economy DSGE model that incorporates heterogeneous expectations through a mixture of rational and boundedly rational agents. The model builds upon standard New Keynesian foundations but introduces behavioral heterogeneity to better reflect real-world deviations from full rationality. The simulations focused on impulse responses to two types of shocks: foreign demand, and monetary policy. The results demonstrate that the presence of boundedly rational agents significantly alters the dynamic adjustment of the economy. Compared to the rational benchmark, the model with heterogeneous expectations exhibits delayed convergence, more pronounced oscillatory behavior, and in some cases even reversed short-term responses.

The backward-looking nature of subjective expectations contributes to inertia and volatility, with important implications for the design and calibration of monetary policy. The results have direct implications for monetary policy in small open economies. When a significant share of agents forms expectations adaptively, standard Taylor-type interest rate rules may lead to prolonged oscillations or even short-term destabilization following shocks. Policymakers should account for slower expectation adjustment by applying more gradual policy responses, incorporating forward guidance, or using macroprudential tools to dampen volatility. Furthermore, exchange rate management in such environments may require greater attention, as heterogeneous expectations can amplify terms-of-trade fluctuations and cross-border spillovers.

In line with recent findings that behavioral deviations such as diagnostic expectations can amplify policy effects and call for more gradual interest rate adjustments (Magzhanov, 2025), the results here suggest that policies optimized under full rationality assumptions may underestimate the complexity of real-world adjustment dynamics. Overall, the findings highlight the relevance of modeling expectation heterogeneity in open economy frameworks. They support the inclusion of behavioral heterogeneity in macroeconomic models, as advocated by De Grauwe (2012). Further research may extend this framework to include learning or regime switching.

References

Calvo GA (1993) Staggered prices in a utility-maximizing framework. Journal of monetary Economics. 2(3), 383-98.

Cornea-Madeira A, Hommes C, Massaro D (2019) Behavioral heterogeneity in US inflation dynamics. Journal of Business & Economic Statistics, 7(2), 288-300.

De Grauwe P (2012) Booms and busts in economic activity: A behavioral explanation. Journal of Economic Behavior & Organization, 83(3), 484-501.

Evans GW, Gibbs CG, McGough B (2024) A unified model of learning to forecast. American Economic Journal: Macroeconomics, forthcoming.

Galı J, Gertler M (1999) Inflation dynamics: A structural econometric analysis. Journal of monetary Economics, 44(2),195-222.

Hommes C (2021) Behavioral and experimental macroeconomics and policy analysis: A complex systems approach. Journal of Economic Literature, 59(1), 149-219.

Kostyshyna O, Salle I, Truong H (2025) Anchored Inflation Expectations: What Recent Data Reveal. Bank of Canada.

Magzhanov T (2025) Monetary Policy under Diagnostic Expectations. Available at SSRN 5266320.

Smets F, Wouters R (2003) An estimated dynamic stochastic general equilibrium model of the euro area. Journal of the European Economic Association, 1(5),1123-75.

Woodford M, Walsh CE (2005) Interest and prices: Foundations of a theory of monetary policy. Macroeconomic Dynamics, 9(3), 462-8.

Received: 2025-08-12 Last revision: 2025-08-18 Accepted: 2025-08-19

Integrating Market Attention into Portfolio Optimization

Jialei Xiong[1], Diana Barro[2], Martina Nardon[3]

Abstract. Behavioral finance research has increasingly examined how market attention influences financial market dynamics. In recent years, the development of big data technology has facilitated the use of search engine data, particularly Google Trends, to measure market attention. However, existing studies mostly focus on verifying the statistical correlation between Google Trends and market variables (e.g., returns, trading volume, or volatility). Few investigations examine the systematic embedding of attention indicators, like Google Trends data, in portfolio construction and asset allocation strategies. Building on the existing literature, this paper applies Google Trends search data as a proxy variable for market attention and explores its practical application in stock investment strategies. This research concentrates on integrating Google Trends search data as an attention factor into an actionable portfolio optimization framework. The study also compares the performances with a traditional optimization model that does not incorporate attention factors. The results indicate that market attention has significant practical value in portfolio optimization.

Keywords: Market attention, Google Trends, Portfolio optimization, Behavioral finance

JEL Classification: G11, G41
AMS Classification: 90A09, 91B28, 91G10

1 Introduction

Behavioral finance addresses gaps in traditional models by explaining empirical trends in financial behavior, capital management, and stock market bubbles. Traditional asset pricing models assume that investors are rational and that market information is efficient (Fama, 1970). However, behavioral finance fills the gaps left by traditional models in explaining empirical trends in financial behavior, capital management, and stock market bubbles (Ritter, 2003). Behavioral finance delves into the behavioral aspects of individual and institutional investors in financial markets, playing a crucial role in investment management strategies (Seth et al., 2020). As Ofek et al. (2004) emphasize, investors' emotions and biases significantly influence market outcomes. Investor behavior is often driven by cognitive biases, emotional reactions, and social influences (Barberis and Thaler, 2003; De Long et al., 1990). Building on the role of investor behavior, a key yet under-explored aspect is how market attention influences asset allocation and portfolio performance.

The rise of the internet and big data has enhanced the ability to analyze investor behavior, offering new insights into financial decision-making processes. Especially in the financial markets, large amounts of data can be utilized to gain a deeper understanding of the decision-making processes of market participants. Recent advances in big data technology have enabled researchers to use digital footprints (e.g., online search queries) to quantify market attention. Among these, Google Trends has become a widely adopted metric for measuring real-time changes in investor sentiment and information-seeking behavior (Da et al., 2011; de Castro et al., 2023). Research indicates that search interests for specific stocks or market indices can predict short-term price movements, volatility (Hamid and Heiden, 2015), and trading volume (de Castro et al., 2023). Existing literature primarily focuses on the statistical correlation between attention metrics and market variables (e.g., returns, volatility, trading volume), with further research exploring the use of Google Trends search interest to construct profitable trading strategies (Preis et al., 2013). While existing research has explored integrating investor sentiment into portfolio optimization frameworks to enhance investment strategy performance (Banholzer et al., 2019), systematically embedding attention metrics into portfolio construction frameworks remains underdeveloped.

This study proposes a novel method that integrates Google Trends based attention factors into portfolio optimization, thereby filling a research gap. We define market attention factors based on changes in Google Trends search interest and apply them to a diversification strategy (Kristoufek, 2013). This strategy applies the diversity of market

[1] VŠB - Technical University of Ostrava, Department of Finance, 17. listopadu 2172/15, 708 00 Ostrava-Poruba, Czech Republic, jialei.xiong@vsb.cz.
[2] Ca' Foscari University of Venice, Department of Economics, San Giobbe, Cannaregio 873, 30121 Venice, Italy, d.barro@unive.it.
[3] Ca' Foscari University of Venice, Department of Economics, San Giobbe, Cannaregio 873, 30121 Venice, Italy, mnardon@unive.it.

attention to distributing risk and applies a rebalancing method to capture dynamic changes in search interest over time. To support our approach, we draw on two key observations from the literature.

Attention as a behavioral risk factor: High levels of market attention are typically associated with overvaluation and subsequent price adjustments (Stambaugh and Yuan, 2017). By incorporating attention signals, portfolios can avoid overinvesting in assets experiencing speculative bubbles.

Emerging markets as a testing ground: Previous research has primarily focused on developed markets (e.g., U.S. or European stock markets), but due to lower information transparency and higher retail investor participation (Barber and Odean, 2000), attention-driven inefficiencies may be more pronounced in emerging markets. Our case study using the Hong Kong Stock Exchange fills this empirical gap. Google Trends accounted for 91.21% of Hong Kong's search market share as of May 2025 (StatCounter, 2025), ensuring the representativeness of the local investor behavior sample.

Our comparative analysis of attention-driven portfolios versus traditional models reveals that incorporating market attention significantly enhances risk-adjusted returns. This finding not only bridges the gap between behavioral finance and asset allocation theory but also provides practical insights for investors seeking to mitigate irrational market pressures. Future research should therefore concentrate on the development of hybrid models combining attention signals with alternative data sources, such as social media sentiment. Furthermore, the existing framework should be extended to other asset classes, such as cryptocurrencies.

The structure of the following paper is as follows. The second section is the methodologies employed to extract attention-based insights. Section 3 describes the data. Section 4 presents the empirical results. Section 5 discusses the implications and concludes.

2 Methodology

We propose a novel approach to integrating market attention into the existing portfolio optimization framework, aiming to improve investment performance beyond that of traditional strategies. A substantial amount of recent empirical evidence has demonstrated that Google Trends data can be effectively normalized on a scale from 0 to 100, where 100 represents the highest search volume within the specified period and region. It is worth noting that reported search interest fluctuates depending on the selected time frame, geographic coverage, and other settings; therefore, proper adjustment and preprocessing are essential for empirical research.

This study analyzes Google Trends' search interest data to determine whether fluctuations in market attention can improve portfolio performance within a traditional optimization framework. Since search interest is a relative measure, we quantify the week-to-week change in market attention by calculating the relative change in search interest ($\Delta GSV_{t,i}$). The computation is defined as follows:

$$\Delta GSV_{t,i} = \frac{|GSV_{t,i} - Med(GSV_{t-1,i}, \ldots, GSV_{t-3,i})|}{Med(GSV_{t-1,i}, \ldots, GSV_{t-3,i})}. \quad (1)$$

Where $GSV_{t,i}$ denotes the Google Trends search interest for stock *i* in week *t*, and $Med(GSV_{t-1,i}, \ldots, GSV_{t-3,i})$ represents the median value over the past three weeks. This method enables the capture of short-term fluctuations in search interest, thereby providing a dynamic and quantifiable measure of market behavior.

Our methodology is based on the diversification strategy proposed by Kristoufek (2013), one of the first attempts to incorporate market attention into portfolio allocation. In Kristoufek's "Google Trends Strategy", investor attention is represented by the search volume of stock-related keywords. This assumes that increased online search activity reflects greater market concern or uncertainty, which translates into higher stock risk. Within this framework, stocks attracting greater search intensity are given lower portfolio weights, while those receiving relatively less attention are given higher weights.

While this approach provides an interesting link between online search behavior and financial decision-making, it relies on basic Google Trends figures, which may not fully reflect the constantly changing nature of investor attention. Search volumes are relative and subject to fluctuations over time, often influenced by external events that can affect their comparability over time or between assets. To address these limitations, the present study introduces an adjusted measure of search activity: $\Delta GSV_{t,i}$. This reflects the week-to-week change in Google Search Volume for stock *i* at time *t*, focusing on changes rather than absolute levels. This emphasizes shifts in investor attention that are more closely aligned with market dynamics and potential reallocation behavior.

Accordingly, in our optimization framework, the portfolio weight of stock *i* in week *t* is determined by incorporating $\Delta GSV_{t,i}$ as a key input, ensuring that the allocation not only responds to persistently high levels of attention but also to sudden changes in market focus. This modification refines Kristoufek's original idea, allowing for a more sensitive and dynamic integration of investor attention into portfolio construction. The practical relevance of this refinement will be explored further in Section 4, where we will compare the performance of our adjusted strategy with that of the standard approach, the portfolio weight of stock *i* in week *t* is determined as follows:

$$w_{t,\mathrm{i}} = \frac{\Delta GSV_{t,i}{}^{-\beta}}{\sum_{j=1}^{n} \Delta GSV_{t,i}{}^{-\beta}}. \tag{2}$$

The β (power law parameter) measures how sensitive the weights are to search frequency: when $\beta > 0$, the weights for stocks with higher search frequencies are reduced; when $\beta < 0$, the weights for stocks with higher search frequencies are increased; when $\beta = 0$, all stocks have equal weights.

3 Data description

The first stage of this study involved defining the scope of tradable assets to ensure both representativeness and data availability. During the research period, the Hang Seng Index (HSI) comprised 82 constituent stocks. However, due to time constraints and limitations in retrieving Google Trends data, we restricted our focus to the top 30 constituents ranked by current market capitalization. Stocks for which valid search interest data could not be obtained during the designated period were excluded. This was to maintain data quality and analytical robustness.

Weekly search interest data for the selected 30 stocks were manually collected from the Google Trends platform, covering the period from January 2020 to December 2024. This five-year period represents the longest timeframe for which Google Trends provides weekly data. For analytical purposes, transaction costs were assumed to be insignificant, and the portfolio was rebalanced weekly to align asset weights with the latest market information.

The purpose of using Google Trends in this study is primarily to closely monitor investor search activities related to the selected stocks. Consistent with prior literature (Kristoufek, 2013; Preis et al., 2013), using stock tickers as search keywords is the most reliable method for obtaining highly relevant search results with minimal irrelevant information. Accordingly, each stock was entered into the search in the format "HKG: [ticker code]". To collect data for Tencent Holdings Limited, for example, the keyword "HKG:0700" was used. Furthermore, to improve the accuracy and comparability of results, the "Topic" category in Google Trends was selected instead of "Search Term," ensuring that the data included searches broadly related to the stock entity, rather than only exact keyword matches.

The advantage of this data collection method is that it minimizes noise introduced by differences in natural language expression or ambiguity. It improves the comparability of different stock data. The resulting Google Trends dataset provides a strong foundation for quantifying market attention, portfolio construction, and performance analysis.

4 Results

The Google Trends strategy framework of this study introduces the power-law parameter β as a key factor in adjusting weight distribution. This parameter controls the direction and strength of the influence of market attention, as measured by search interest, on stock weights. We set the range of β values to be between −1 and 1, with a step size of 0.1. We can then obtain the stock weight distribution under different power-law conditions by substituting different β values into *Equation* (2).

From a theoretical perspective, positive β values correlate with the principles of risk diversification strategies — when the search interest of a particular stock increases, its weight is reduced. Accordingly, to mitigate the potential risks associated with concentration in stocks with high market attention. Conversely, negative β values indicate a chasing-attention investment approach, where the strategy increases the investment weights of stocks with high search interest.

To demonstrate the effect of various β specifications, we have chosen 4 representative values from the 21 possible options for visual and comparative analysis. Furthermore, to examine the potential influence of changes in market attention on investment strategies, we designed three different strategies to portfolio construction:

- Strategy A1: Regardless of whether the weekly change in search interest for stock *i* in week *t* is positive or negative, it is directly incorporated into *Equation (1)*.

- Strategy B1: The weekly change in search interest for stock *i* in week *t* is incorporated into *Equation (1)* only when it is negative (i.e., search interest decreases); otherwise, it is set to zero.
- Strategy C1: The weekly change in search interest for stock *i* in week *t* is incorporated into *Equation (1)* only when it is positive (i.e., search interest increases); otherwise, it is set to zero.

These three strategies focus specifically on incorporating market attention directly into portfolio weight adjustments. To further explore the combination of market attention and a traditional portfolio optimization framework, we apply the minimum-variance (min-variance) optimization approach to each of the above strategies, resulting in three additional strategies:

- Strategy A2: Apply the min-variance method to the set of stocks identified by Strategy A1.
- Strategy B2: Apply the min-variance method to the set of stocks identified by Strategy B1.
- Strategy C2: Apply the min-variance method to the set of stocks identified by Strategy C1.

The Hang Seng Index (HSI) is adopted as the benchmark portfolio for performance evaluation. Each strategy is compared against the benchmark portfolio across multiple β values in terms of return, risk, and the Sharpe ratio. This comparative analysis provides a thorough evaluation of the effectiveness of market attention when applied independently or incorporated into a conventional portfolio optimization framework.

	E(Ri)	***STD***	***Sharpe Ratio***	***MDD***
A1(*β=-1)*	34.804%	31.608%%	1.02830	37.113%
A1(*β=-0.5)*	36.332%	30.821%	1.10413	33.899%
A1(*β=0)*	12.076%	19.985%	0.48909	27.888%
A1(*β=0.5)*	9.629%	19.290%	0.37798	25.610%
A1(*β=1)*	7.968%	19.299%	0.29210	24.677%
B1(*β=-1)*	-6.186%	20.243%	-0.41927	51.139%
B1(*β=-0.5)*	-6.632%	19.535%	-0.45726	49.128%
B1(*β=0)*	-6.812%	19.353%	-0.47090	46.664%
B1(*β=0.5)*	-6.886%	19.277%	-0.47661	46.489%
B1(*β=1)*	-6.840%	19.783%	-0.46208	47.111%
C1(*β=-1)*	51.577%	34.719%	1.42929	46.723%
C1(*β=-0.5)*	52.072%	32.993%	1.50854	41.399%
C1(*β=0)*	27.319%	23.727%	1.05443	34.375%
C1(*β=0.5)*	21.811%	22.436%	0.86954	24.393%
C1(*β=1)*	17.028%	22.722%	0.64813	24.966%
A2	-1.023%	14.815%	-0.22435	37.960%
B2	-4.972%	15.682%	-0.43300	41.758%
C2	-0.454%	19.149%	-0.14259	44.184%
Min-var	-1.198%	14.916%	-0.23460	40.121%
HSI	-6.553%	23.365%	-0.37898	51.498%

Table 1 Comparison of performance under different strategies

Table 1 reports the performance of different investment strategies in terms of annualized return (***E(Ri)***), annualized volatility (***STD***), Sharpe ratio, and maximum drawdown (***MDD***). The comparison includes six strategies based on Google Trends data (A1, B1, C1 and their minimum-variance-adjusted versions A2, B2, C2), the traditional minimum-variance portfolio (Min-var), and the Hang Seng Index (HSI) benchmark.

Among strategies without minimum-variance adjustment, C1 performs best when β is negative. In particular, C1 (β = -0.5) achieves an annualized return of 52.072% and a Sharpe ratio of 1.50854, significantly outperforming other strategies and the benchmark. Although its maximum drawdown is higher than some low-risk portfolios, it remains acceptable given the high returns. A1 also performs well with a negative β, such as A1 (β = -0.5) with a 36.332% return and a 1.10413 Sharpe ratio, though still below C1. B1 consistently shows negative returns and low Sharpe ratios (ranging from -0.41927 to -0.47661), with maximum drawdowns exceeding 46%.

With minimum-variance adjustment, A2, B2, and C2 show much lower volatility (all below 20%). Although returns are lower than A1, B1, C1 the A2, B2, C2 group outperforms the HSI benchmark in volatility and drawdown

control, and demonstrates stronger defensive characteristics in volatile markets, making it suitable for risk-averse investors. The traditional Min-var portfolio performs similarly to A2, with no improvement over the benchmark in risk-adjusted terms.

Overall, incorporating market attention indicators—especially in C1-type strategies with negative β—can significantly enhance returns and risk-adjusted performance. Minimum-variance optimization effectively reduces risk but may limit return potential. Future research could optimize parameter selection, test strategy adaptability in different markets， and incorporate additional behavioral finance variables (e.g., investor sentiment or search data) to improve predictive power and portfolio allocation.

5 Conclusion

This study examines the performance of investment strategies that incorporate market attention indicators derived from Google Trends data. The findings show that adding market attention to portfolio construction can greatly improve investment results, especially when β is negative. Among the strategies evaluated, the C1 strategy achieves superior annualized returns and Sharpe ratios compared to alternative strategies and the market benchmark. This highlights the value of market attention as a behavioral finance variable in asset allocation.

At the same time, the incorporation of market attention is accompanied by increased portfolio volatility and maximum drawdown, reflecting the risks of uncertainty and sentiment associated with attention-driven signals. To reduce these risks, the study combines minimum variance optimization with attention-based strategies (A2, B2, and C2), which effectively reduce risk measures and enhance defensive characteristics. However, this risk reduction comes at the cost of lower returns, highlighting the trade-off between risk and reward.

Furthermore, the impact of market attention on portfolio performance shows non-linear and asymmetric patterns. When the attention parameter is positive, risk diversification mechanisms lower volatility but also constrain returns. Conversely, negative attention values are associated with opportunities for higher excess returns. These results emphasize the importance of considering the direction and strength of market attention signals when dynamically adjusting investment strategies.

In summary, the findings provide empirical support for incorporating behavioral variables, particularly market attention, into portfolio optimization frameworks. Using Google Trends as a dynamic input introduces a new perspective to traditional models and has practical implications for asset management. Future research should focus on adaptive parameter adjustment, evaluating strategy performance in various market conditions, and integrating additional behavioral indicators to improve model robustness and forecasting accuracy.

Acknowledgements

The research was financially supported by the SGS project no. SP2025/043 of VSB – Technical University of Ostrava.

References

Banholzer N, Heiden S, Schneller D (2019) Exploiting investor sentiment for portfolio optimization. Business Research, 12, pp.671-702. https://doi.org/10.1007/s40685-018-0062-6

Barber B M, Odean T (2000) Trading is hazardous to your wealth: The common stock investment performance of individual investors. The journal of Finance, 55(2), pp.773-806. https://doi.org/10.1111/0022-1082.00226

Barberis N, Thaler R (2003) A survey of behavioral finance. Handbook of the Economics of Finance, 1, pp.1053-1128. https://doi.org/10.1016/S1574-0102(03)01027-6

Da Z, Engelberg J, Gao P (2011) In search of attention. The journal of finance, 66(5), 1461-1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

de Castro J, Piccoli P (2023) Do online searches actually measure future retail investor trades?. International Review of Financial Analysis, 86, 102552. https://doi.org/10.1016/j.irfa.2023.102552

De Long J B, Shleifer A, Summers L H, Waldmann R J (1990) Noise trader risk in financial markets. Journal of political Economy, 98(4), pp.703-738. https://doi.org/10.1086/261703

Fama E F (1970) Efficient capital markets. Journal of finance, 25(2), pp.383-417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x

Hamid A, Heiden M (2015) Forecasting volatility with empirical similarity and Google Trends. Journal of economic behavior & organization, 117, 62-81. https://doi.org/10.1016/j.jebo.2015.06.005

Kristoufek L (2013) Can Google Trends search queries contribute to risk diversification?. Scientific reports, 3(1), p.2713. https://doi.org/10.1038/srep02713

Ofek E, Richardson M, Whitelaw R F (2004) Limited arbitrage and short sales restrictions: Evidence from the options markets. Journal of Financial Economics, 74(2), 305-342. https://doi.org/10.1016/j.jfineco.2003.05.008

Preis T, Moat H S, Stanley H E (2013) Quantifying trading behavior in financial markets using Google Trends. Scientific reports, 3(1), pp.1-6. https://doi.org/10.1038/srep01684

Ritter J R (2003). Behavioral finance. Pacific-Basin finance journal, 11(4), pp.429-437. https://doi.org/10.1016/S0927-538X(03)00048-9

Seth V, Kumar S (2020) A study of effect of behavioural biases on investment decisions. Durgadevi Saraf Institute of Management Studies (DSIMS) The Management Quest, 3(1), pp.21-50.

Stambaugh R F, Yuan Y (2017) Mispricing factors. The review of financial studies, 30(4), pp.1270-1315. https://doi.org/10.1093/rfs/hhw107

Website:

StatCounter (2025) https://gs.statcounter.com/search-engine-market-share/all/hong-kong

Received: 2025-06-08 Last revision: 2025-08-03 Accepted: 2025-08-04

Coordination of Expectations: A Case Study in Multi-Asset Markets

Michaela Sedláková[1]

Abstract. We present the results of a multi-asset Learning-to-Forecast experiment. Participants were asked to predict the prices of three risky assets, with their key information being the initial price history of these assets. Our research investigates how participants coordinate their expectations in a multi-asset market environment. To analyze this coordination, we decompose individual forecast errors into two components: dispersion error (DE) and common error (CE). The results reveal two key insights. First, in stable markets, a relatively small CE suggests that individual forecast errors tend to cancel out in the aggregate, consistent with the notion of rational expectations with error. Second, in markets characterized by moderately large bubbles, there is a clear dominance of smaller DE, indicating a high degree of coordination among participants, reflecting a spontaneous alignment of beliefs – possibly driven by similar interpretations of price trends, despite the absence of direct communication. This study is classified as a case study due to the limited number of sessions. While exploratory, the findings offer valuable insights into how initial conditions shape expectation coordination.

Keywords: experimental economics, expectations, asset pricing

JEL Classification: C92, D84, G12, G41
AMS Classification: 91A90, 91G30

1 Introduction

The way market participants form expectations has a profound impact on asset pricing, trading patterns, and the emergence of speculative bubbles. When individuals develop similar beliefs about future prices – often based on shared interpretations of past trends – this alignment can amplify market movements, sometimes leading to the formation of speculative bubbles. These bubbles occur when asset prices deviate significantly from their fundamental values, driven not by intrinsic worth but by collective optimism and trend extrapolation. Historical episodes such as the dot-com bubble, the U.S. housing crisis, and the Bitcoin surge illustrate how synchronized expectations can inflate prices beyond sustainable levels, ultimately resulting in sharp corrections. Gaining insights into expectation formation is therefore essential for understanding market behavior.

One effective way to study expectation formation is through controlled experiments, particularly Learning-to-Forecast (LtF) setups. In these experiments, participants act as financial forecasters and repeatedly predict future asset prices based on available information (see, for example, Hommes et al., 2005; Hommes et al., 2008; or Bao et al., 2021 for a review). Their predictions are aggregated to determine market prices, allowing researchers to observe how expectations evolve, how they respond to initial conditions, and how they influence market outcomes. Unlike survey-based methods, experimental designs offer full control over the information environment, enabling a clearer analysis of the mechanisms behind expectation formation.

Previous LtF experiments have shown that positive expectation feedback often drives asset prices away from their fundamental values, leading to bubbles and crashes (Hommes et al., 2008; Heemeijer et al., 2009; Hommes et al., 2021). These studies, mostly conducted in single-asset settings, also reveal that participants are usually highly coordinated in their forecasts. Extending this analysis to multi-asset environments allows us to examine how initial price history affects the degree of coordination in more complex market settings.

The aim of this paper is to study the coordination of expectations in a multi-asset market and to examine how initial price history influences this coordination. Using a LtF experimental framework, we extend the design of Anufriev et al. (2022) by introducing three risky assets, each with a distinct initial price history. This setup allows us to analyze how participants align their forecasts across assets and how different historical price patterns affect the degree of expectation alignment. To assess coordination, we analyze the interquartile range (IQR) of individual

[1] VSB – Technical University of Ostrava, Department of Finance, 17. listopadu 2172/15, 708 00 Ostrava–Poruba, michaela.sedlakova@vsb.cz.

forecasts over time and decompose individual forecast errors into two components: dispersion error (DE), which measures the variability of individual forecasts around the group average, and common error (CE), which captures the deviation of the average forecast from the realized market price. The sum of these two components constitutes the individual error (IE), representing overall forecasting inaccuracy.

Our experimental results reveal two key insights into expectation coordination in multi-asset markets. In stable markets, participants' forecasts become more coordinated after an initial learning phase, resulting in consistently low IQR. From the perspective of forecast error decomposition, a relatively small CE is associated with stable markets, suggesting that individual forecasting errors tend to cancel out – consistent with the concept of rational expectations with error. Second, in markets characterized by moderately large bubbles, we observe a dominance of low DE, indicating a high degree of coordination among participants. This coordination appears to emerge spontaneously, likely driven by shared interpretations of price trends, even in the absence of communication. Furthermore, an asset with a more stable initial price history tends to foster higher coordination.

The decomposition of IE is a common practice in the LtF literature. For example, Hommes et al. (2008) observed the emergence of significant bubbles in most of their experiments and found that only a small portion of IE could be attributed to DE. This suggests that participants were coordinated in their prediction strategies, challenging the notion of 'rational expectations with error' as a fully accurate description of expectation formation. Hanaki et al. (2019) investigated the impact of framing on asset price dynamics in an LtF experiment by varying participants' tasks (forecasting prices or returns) and the presentation of historical data (prices or returns). Surprisingly, they found no significant differences between treatments in terms of the fraction of error explained by CE. On average, about two-thirds of IE was accounted for by CE. In another study, Hommes et al. (2021) focused on the emergence of bubbles in large groups. Their IE decomposition revealed substantial variation across markets, with both CE and DE playing significant roles. Notably, stable markets tended to exhibit lower CE, indicating that individual errors often canceled each other out. Our findings align with this pattern: stable markets generally show low CE, while markets with moderately large bubbles exhibit higher levels of coordination.

The remainder of the paper is organized as follows: Section 2 introduces the experimental design. Section 3 presents the results. The final section concludes.

2 Experimental design

We have obtained data from six LtF experiments that were conducted at VSB – Technical University of Ostrava with a total of 74 students from Faculty of Economics. Each session involved a different group, with flexible sizes[2] ranging from 9 to 17 participants. The experiments took place in a computer classroom, where students interacted within a shared market environment. Printed instructions and a task briefing were provided at the start. Communication was not allowed. After the experiment, participants were ranked by average absolute prediction error and paid between 50 and 700 CZK.

The experiment follows a standard LtF setup and a present value asset pricing model. Participants act as mean-variance investors allocating wealth between risk-free and three risky assets. Risk-free asset yields a constant gross return $R = 1 + r > 1$, while risky assets pay IID dividends with mean $\bar{y}$ each period. The market-clearing price, following Anufriev et al. (2022), is defined as:

$$p_{t,a} = p_a^f + \frac{1}{(1+r)}\left(\bar{p}_{t+1,a}^e - p_a^f - \varepsilon_{t,a}\right). \tag{1}$$

Here, $p_{t,a}$ is a market-clearing price of risky asset a at time t, with $a \in \{1, 2, 3\}$ and $t \in \{11, \ldots, 60\}$ [3]. p_a^f is the asset's a fundamental value, defined as $p_a^f = \bar{y}/r$, where $\bar{y}$ is IID dividend mean and r is discount factor. $\bar{p}_{t+1,a}^e$ is the average expected price for the next period $t+1$ for the asset a, and $\varepsilon_{t,a}$ is a small random supply shock. According to Equation (1), the realized market price reflects a weighted average of fundamental price and average expectations of participants. Anticipated price increases can drive current demand, creating positive feedback. The rational equilibrium corresponds to the asset's fundamental value.

[2] The sessions were designed to include groups of approximately 15 participants, based on the availability of up to 20 computers. To ensure sessions could proceed even with lower attendance, the group size remained flexible. A session would commence if at least 8 participants arrived on time. In total, six sessions were conducted with the following attendance numbers: 11, 9, 15, 9, 13, and 17 participants in sessions EXP1 through EXP6. No sessions were cancelled.

[3] Periods 1 to 10 represent the initial price history of the risky assets that was provided to participants.

Participants received the following introduction to the experiment. Participants assume the role of pension fund analysts tasked with allocating a one-period investment. Investment options include one risk-free and three risky assets. Participants are informed that the risky assets are long-term uncorrelated. The risk-free investment is a government bond with interest rate 5%. Alternatively, the fund can invest in indefinitely lived risky assets, which involve uncertainty in future prices and dividends. These dividends are IID with a mean of $10 per period. With interest rate and dividends provided, participants have enough information to estimate the fundamental value of risky assets, which is set at $200 for consistency and ease of comparison (see Figure 1). This uniform fundamental value allows us to isolate and study the effect of different initial price histories. We set distinct initial price histories – one increasing, one decreasing, and one stable – to examine how different trend patterns influence participants' coordination in experimental multi-asset markets.

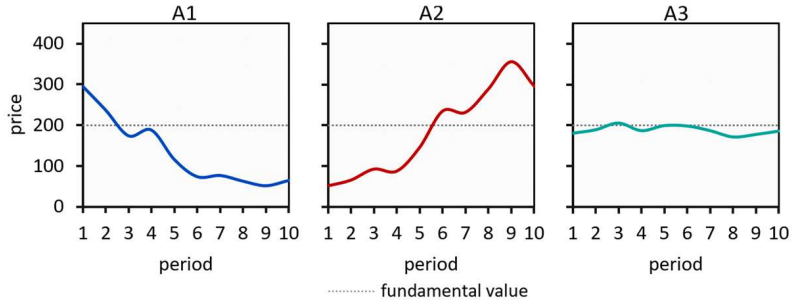


Figure 1 Initial price history for all risky assets
Source: own work

Participants are instructed to predict future asset prices, which the pension fund uses to make trading decisions. In line with standard LtF experiments, no explicit pricing formula is provided. However, key market features are described – for example, higher price forecasts increase asset demand, and multiple funds contribute to total demand. As in Anufriev et al. (2022), participants receive initial price history at the start. This study focuses on how initial price history influences expectation coordination, extending the analysis to a multi-asset setting.

In each period, participants make price forecasts using the following available information – past realized prices up to period $t-1$, participant's previous price predictions up to period t, participant's total average error as well as average errors for particular assets. Once all predictions from all participants for period $t+1$ were received, the realized price of assets for the current period t was determined according to Eq. (1) and this was repeated for all 50[4].consecutive periods. In EXP4, we obtained predictions for 40 periods instead of the full 50 due to technical issues.

3 Coordination of expectations

As previously discussed, the LtF literature shows strong evidence of expectation coordination. We now explore this in a multi-asset setting, focusing on how expectations align and how the initial price history of assets influences this coordination. Specifically, we ask: Does a stable price history foster stronger coordination among participants? To address this question, we analyze data from six experimental sessions in which participants predicted prices for three risky assets over multiple periods.

Based on observed price patterns (see Figures 2 and 3), we identify two types of market behavior:

1. Stable Markets – All assets remain close to their fundamental values, with average relative deviations under 20% (EXP1, EXP2, EXP3);
2. Moderately Bubble Markets – Some assets show moderate bubbles, peaking at 1.5 to 2 times their fundamental values (EXP4, EXP5, EXP6).

This categorization will serve as the basis for our subsequent analysis of our experimental data.

To analyze how coordination of expectations changes over time, we use the interquartile range (IQR) of individual predictions for each period. Figures 2 and 3 show the IQR alongside the realized price for each asset. A lower IQR indicates stronger coordination among participants, while a higher IQR suggests greater disagreement. The figures reveal that in most cases, the first 10 forecast periods show higher prediction variability – likely due to participants learning how to forecast accurately. We will examine the IQR patterns in more detail for both types of markets.

[4] Following previous LtF experiments, this duration is typical. For a broader discussion, including experiments with longer horizons, see Bao et al. (2021).

We begin by examining expectation coordination in stable experimental markets. As shown in Figure 2, the initial forecast periods display greater heterogeneity, which quickly stabilizes. Beyond this initial phase, the IQR remains consistently low, indicating strong alignment among participants – even without knowing others' predictions. When considering the impact of initial price history, differences in coordination are minimal after the initial periods. The main exception is EXP3, where asset A3 shows notably higher coordination, likely due to its more stable initial price history compared to other assets. Overall, participants demonstrate a high level of coordination in stable markets.

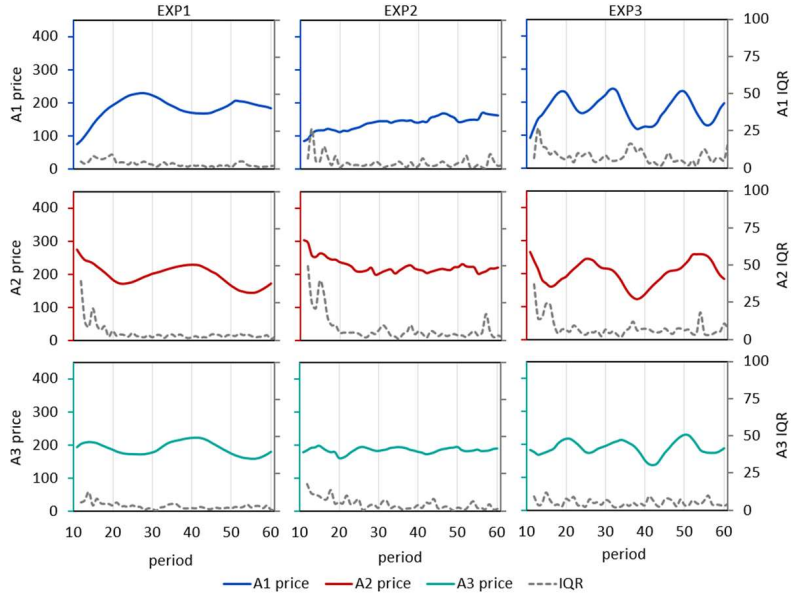


Figure 2 Realized market prices and IQR of individual price predictions in stable markets (EXP1–EXP3)
Source: own work

Next, we analyze experimental markets with moderate bubbles. As shown in Figure 3, participant predictions are more heterogeneous compared to stable markets. The data suggest that coordination tends to increase during periods of rising prices but often breaks down sharply following sudden price drops. Typically, heterogeneity reaches its maximum shortly after the bubble bursts, indicating increased uncertainty and misalignment in expectations. We also examine the role of initial price history. In EXP4 and EXP5, asset A3 consistently shows lower IQR values than other assets, suggesting stronger coordination – possibly due to its more stable initial history. Overall, these results highlight the dynamic nature of coordination in bubble markets, where participants adjust their expectations in response to price shifts, leading to fluctuating levels of alignment.

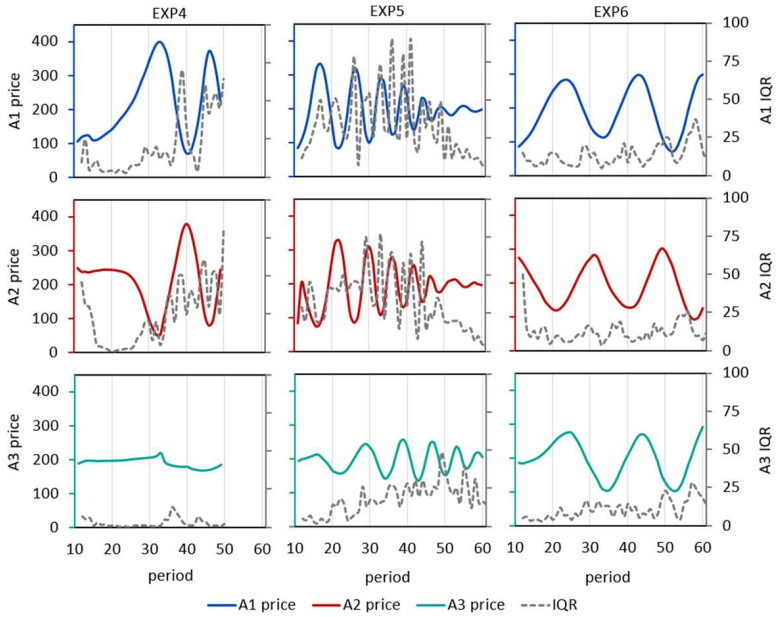


Figure 3 Realized market prices and IQR of individual price predictions in markets with moderately large bubbles (EXP4–EXP6)
Source: own work

We now test the effect of initial price history on participants' coordination using the paired Wilcoxon sign-rank test. Previous observations indicate that the first prediction periods are naturally associated with a higher degree of heterogeneity, as participants are more likely to learn to predict correctly during these times. Therefore, our testing approach involved calculating the average IQR of individual predictions, excluding the first ten prediction periods. We then compared asset A3, which has a stable historical development, to assets A1 and A2, which exhibit more volatile initial price histories. The alternative hypothesis for this test is that asset A3 has a lower IQR value compared to asset A1 or asset A2. The p-value for the comparison with asset A1 is 0.047, and for A2, it is 0.016. Both test results are statistically significant at the 5% level, suggesting that there is evidence to reject the null hypothesis. We conclude that a stable initial price history can potentially lead to higher coordination among participants.

To further investigate the coordination between participants, we use the following decomposition of the average individual quadratic error (IE) for particular assets in experimental markets, as introduced by Hommes et al. (2005):

$$IE = DE + CE,$$

$$\frac{1}{I \cdot T}\sum_{i}^{I}\sum_{t}^{T}\left(p_{i,t,a}^{e} - p_{t,a}\right)^2 = \frac{1}{I \cdot T}\sum_{i}^{I}\sum_{t}^{T}\left(p_{i,t,a}^{e} - \overline{p}_{t,a}^{e}\right)^2 + \frac{1}{T}\sum_{t}^{T}\left(\overline{p}_{t,a}^{e} - p_{t,a}\right)^2, \tag{2}$$

where I is number of participants in the particular experiment, T is number of predicted periods, $p_{i,t,a}^{e}$ is expectation about price of an asset a in the period t of participant i, $\overline{p}_{i,t,a}^{e}$ are average expectations about price of an asset a in a period t.

The left-hand side of Eq. (2) expresses the individual error (IE), calculated as the sum of all individual quadratic forecast errors in prices, averaged over time and individuals in the group. This error measures how well participants predict the market price. The first term on the right-hand side of Eq. (2) represents the dispersion error (DE), calculated as the average of the quadratic difference between individual forecasts and the average forecast for the particular period, averaged over time and individuals in the group. The second term on the right-hand side of Eq. (2) measures the common error (CE), calculated as the average quadratic distance between the mean forecast and the realized price over time.

Experiment	Asset	IE	DE	DE (%)	CE	CE (%)
EXP1	A1	57	34	60%	22	40%
	A2	49	18	37%	31	63%
	A3	36	13	36%	23	64%
EXP2	A1	796	538	68%	258	32%
	A2	523	387	74%	136	26%
	A3	311	213	68%	98	32%
EXP3	A1	67	50	74%	17	26%
	A2	175	140	80%	35	20%
	A3	16	4	29%	11	71%
EXP4	A1	2 497	423	17%	2 075	83%
	A2	2 638	679	26%	1 959	74%
	A3	267	228	85%	39	15%
EXP5	A1	3 517	1 149	33%	2 368	67%
	A2	3 660	1 132	31%	2 528	69%
	A3	1 083	420	39%	663	61%
EXP6	A1	794	170	21%	625	79%
	A2	862	388	45%	474	55%
	A3	754	299	40%	456	60%

Table 1 Decomposition of the individual error for all experiments. Stable markets – EXP1, EXP2, EXP3 and markets with moderately large bubbles – EXP4, EXP5, EXP6.
Source: own work

It is clear from Eq. (2) that DE and CE are not easily interpreted in absolute terms but relative to IE. This is because IE tends to be smaller in a more stable market than in a bubbly market. DE measures how well participants are coordinated. If DE is zero, it means that all participants have the same common prediction strategy. This implies that a relatively small percentage of DE compared to IE represents high coordination on a common prediction strategy. On the other hand, a relatively small CE means that individual forecast errors cancel each other out in the aggregate. This aligns with Muth (1961), who stated that even when individuals make forecast errors, on average

they make rational decisions, as individual errors are likely to cancel out. In this situation, expectation formation can be expressed as rational expectations with error.

The results show substantial variation across experimental markets. In Table 1, the IE is divided into the specified two components for all assets in each experiment. We omitted the first 10 predicted periods in the calculation for the same reason as in the pairwise IQR testing. It can be observed that a relatively small CE is obtained in stable markets, indicating that individual forecast errors tend to cancel each other out in the aggregate, which can be expressed as rational expectations with error. In contrast, in markets with moderately large bubbles, there is a clear dominance of smaller DE. This implies that in these markets, characterized by the emergence of more pronounced bubbles, there is a high degree of coordination among participants. Overall, these findings highlight the varying dynamics of forecast errors, suggesting that coordination is highly affected by the overall market behavior.

4 Conclusion

This case study explores how individuals coordinate their expectations in six multi-asset markets, with particular attention to the influence of initial price history. Using a Learning-to-Forecast experimental framework, we extended the traditional single-asset setup to include three uncorrelated risky assets. This design allowed us to investigate not only how participants form expectations individually, but also how their beliefs align collectively across different assets.

This paper is based on a limited number of experimental sessions and therefore serves as a case study. While the findings offer useful insights, they should be interpreted with caution. Future research should include more sessions to enhance robustness and generalizability.

Our findings reveal that coordination of expectations plays a crucial role in shaping market behavior. In stable markets, the interquartile range (IQR) of forecasts narrowed quickly after an initial learning phase, indicating strong alignment among participants. This was further supported by a relatively low common error, suggesting that individual forecasting mistakes tended to cancel out in the aggregate – consistent with the notion of rational expectations with error. In contrast, markets with moderately large bubbles exhibited more volatile IQR patterns and a dominance of low dispersion error, indicating that participants often aligned their expectations, likely driven by shared interpretations of price trends despite the absence of communication. Furthermore, our results suggest that the initial price history of assets can lead to different levels of coordination. Forecasts for the asset with a more stable initial price path showed higher coordination among participants. These findings underscore the importance of expectation coordination in shaping market dynamics.

Acknowledgements

The research was financially supported by the SGS project no. SP2025/043 of VSB – Technical University of Ostrava.

References

Anufriev M, Chernulich A, Tuinstra J (2022) Asset price volatility and investment horizons: An experimental investigation. J Econ Behav Organ 193: 19–48. https://doi.org/10.1016/j.jebo.2021.11.019

Bao T, Hommes C, Pei J (2021) Expectation formation in finance and macroeconomics: A review of new experimental evidence. J Behav Exp Finance 32: 100591. https://doi.org/10.1016/j.jbef.2021.100591

Hanaki N, Hommes C, Kopányi D, Tuinstra J (2019) The effect of framing on the emergence of bubbles and crashes in a Learning to Forecast experiment. CeNDEF Working paper University of Amsterdam.

Heemeijer P, Hommes C, Sonnemans J, Tuinstra J (2009) Price stability and volatility in markets with positive and negative expectations feedback: An experimental investigation. J Econ Dyn Control 33(5): 1052–1072. https://doi.org/10.1016/j.jedc.2008.09.009

Hommes C, Kopányi-Peuker A, Sonnemans J (2021) Bubbles, crashes and information contagion in large-group asset market experiments. Exp Econ 24(2): 414–433. https://doi.org/10.1007/s10683-020-09664-w

Hommes C, Sonnemans J, Tuinstra J, van de Velden H (2005) Coordination of expectations in asset pricing experiments. Rev Financ Stud 18(3): 955–980. https://doi.org/10.1093/rfs/hhi003

Hommes C, Sonnemans J, Tuinstra J, Van de Velden H (2008) Expectations and bubbles in asset pricing experiments. J Econ Behav Organ 67(1): 116–133. https://doi.org/10.1016/j.jebo.2007.06.006

Muth, J F (1961) Rational expectations and the theory of price movements. Econometrica 29(3): 315–335. https://doi.org/10.2307/1909635

Received: 2025-06-08 Last revision: 2025-08-06 Accepted: 2025-08-08

The Impact of Social Media on Consumer Behaviour and Its Effect on GDP Growth in India

Anshu[1]

Abstract. This study explores the relationship between social media-driven consumer behavior and its impact on GDP growth in India. With the increasing penetration of digital platforms, social media has emerged as a significant factor influencing consumer preferences, spending habits, and potentially overall economic growth. By analyzing secondary data, this research aims to identify the association between social media engagement and consumer spending, examine the role of platforms like Facebook, Instagram, and WhatsApp in shaping purchasing decisions, and investigate the correlation between increased social media use and GDP growth. Although the study finds strong positive correlations between these variables, it acknowledges that correlation does not imply causation. Consequently, the findings presented should be interpreted with caution, as the data analysis used here does not statistically confirm causal relationships. Nevertheless, evidence from industry trends and consumer surveys indicates that social media is closely linked with economic activities such as e-commerce growth and consumer engagement in India, thereby suggesting its important role in the nation's digital transformation and economic landscape.

Keywords: GDP, Social Media Usage, Consumer Behaviour, Social Media

JEL Classification: C44
AMS Classification: 90C15

1 Introduction

India has experienced rapid growth in social media usage, driven by increasing internet penetration, affordable smartphones, and cost-effective data plans. The rise of platforms like Facebook, Instagram, WhatsApp, and Twitter has revolutionized communication, e-commerce, and consumer engagement. With over 500 million active users, social media has become a crucial tool for businesses, enabling digital marketing, brand promotion, and direct customer interactions (Vitenu-Sackey, 2020). The introduction of Reliance Jio's affordable data services have further accelerated digital adoption, bringing millions of users online. Social media plays a transformative role in reshaping traditional business models, driving online shopping trends, and influencing consumer behavior. Small businesses and large corporations alike leverage these platforms for market expansion and customer outreach, contributing significantly to India's digital economy. The impact of social media extends beyond individual consumption, fueling growth in sectors like e-commerce, advertising, and fintech. However, it is crucial to emphasize that while social media usage and economic growth have occurred simultaneously, this does not necessarily imply causation. Correlation does not establish a direct causal relationship, and interpreting these associative patterns should be done cautiously. The widespread adoption of social media has also led to shifts in consumer preferences, making online reviews, influencer marketing, and personalized advertisements essential for purchasing decisions. As social media continues to evolve, its influence on India's economy is expected to grow, reinforcing its role as a key driver of GDP growth and digital transformation (Asongu et al. 2019).

1.1 Social Media Users

The number of users actively using social media worldwide has grown significantly. In the year 2020, an estimated 3.6 billion people were using different social media platforms worldwide. In 2021, this number climbed to 3.83 billion, showing an increase of 5% over the previous year. In 2022, the number of social media users was projected to rise to 4.01 billion, an estimated 4.7% increase from 2021. The growth trajectory continued with similar momentum for 2023 and 2024, predicting 4.19 billion and 4.33 billion users respectively. While this growth illustrates a global trend towards digital engagement, it should be noted that such data represents associative patterns, not direct evidence of economic causation.

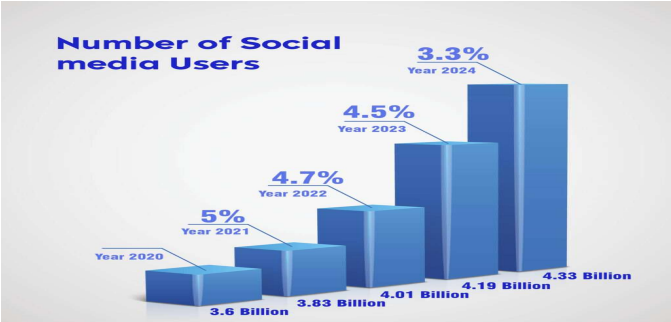


Figure 1 Number of social media users (Source: arihantwebtech.com)

1.2 Social Media Penetration

Social media penetration refers to the fraction of the population worldwide who use different social media applications. Here also the penetration percentage is set to rise. For social media penetration too, we have projected the percentage till the year 2024.To start with the base year 2020, 48.3% of people globally were estimated to be using one or the other social media applications. In 2021, the percentage was estimated to be around 50.3%. So, we saw in plain percentage terms a 2 point rise just in 2 years. The projected numbers for 2022, 2023 and 2024 also shows promising results. For the year 2022, the percentage population worldwide who will be using some social media applications is set to climb to 52.1 per cent, approximately a 1.8 percentage point rise from 2021 levels. Similarly, for the year 2023, the percentage is estimated to climb to 53.8 per cent and for the year 2024, the fraction of the population worldwide is estimated to rise to 55.3% all over. Again, it is essential to clarify that increases in penetration percentages demonstrate associative trends rather than direct causation related to economic outcomes.

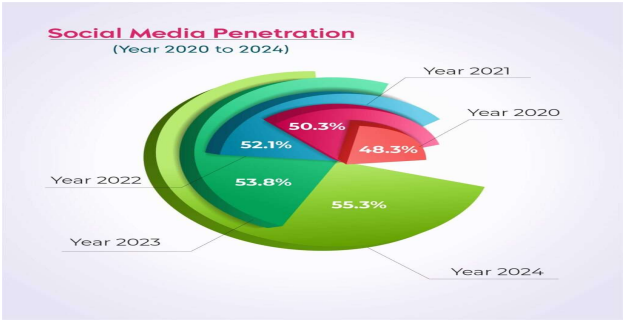


Figure. 2 Social media penetration (Source: arihantwebtech.com)

1.3 Social Media Platforms Users

Various social media platforms operate today, notably Facebook, Twitter, YouTube, WhatsApp, Instagram, Snapchat, Telegram, and TikTok, among others. Facebook, with approximately 2,898 million worldwide users, holds the largest user base, followed by YouTube (2,297 million), WhatsApp (2,021 million), and Instagram (1,312 million). India represents the biggest market for Facebook, with 351 million users. While the popularity and reach of these platforms are well-documented, such statistics do not inherently imply economic causality.

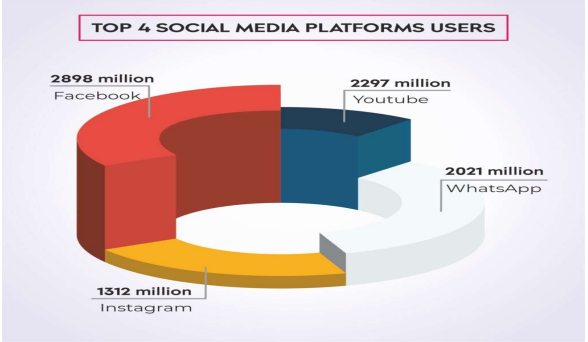


Fig. 3 Social media platforms users (Source: arihantwebtech.com)

1.4 Factors Driving Growth

- Cheaper data plans and increased internet accessibility have made social media more accessible to a wider audience.
- Smartphones have become the primary screen for Indian users, further driving social media usage.
- The Indian government's "Digital India" initiative has also contributed to the growth of internet and social media usage.
- The Covid-19 pandemic led to a surge in social media usage as people turned to platforms for information, expression, and interaction.

1.5 Significance of Social Media in Consumer Behaviour and Its Economic Implications

Social media has emerged as a transformative force in shaping consumer behaviour by influencing purchasing decisions, brand perceptions, and lifestyle preferences. Consumers increasingly depend on online reviews, influencer marketing, and targeted advertisements to make informed choices, while platforms like Instagram, Facebook, and YouTube facilitate deeper business-customer engagement, brand loyalty, and impulse buying through word-of-mouth marketing and live promotions. Although these behavioral shifts are significant, caution is warranted when interpreting their economic effects, as behavioral influence does not directly imply economic causation. Studying the broader impact of social media on GDP growth is therefore essential, especially as digital platforms now underpin e-commerce expansion, digital marketing, and advertising revenues key contributors to national economic activity. However, any observed relationship between social media usage and GDP growth should be interpreted as correlational, not causal, due to the limitations of secondary data analysis. With India's digital adoption accelerating, understanding how social media engagement influences consumer spending patterns and macroeconomic indicators remains an urgent research need. This study aims to explore that relationship by analyzing secondary data from credible sources to identify key social media platforms impacting consumer-driven economic trends. The research prioritizes establishing associative insights that can help policymakers and marketers design more effective, digitally informed strategies for inclusive and sustainable growth.

1.6 Research Question

With India's rapidly increasing digital adoption and the growing role of social media in shaping consumer engagement, this research seeks to answer a focused question: How does increased engagement with social media platforms correlate with GDP growth in India?

1.7 Hypothesis of the Research

H1: Increased social media engagement correlates with GDP growth, positively influencing India's economic development. This hypothesis reflects the core aim of the study to explore whether rising digital interactions on social media platforms such as Facebook, Instagram, WhatsApp, and YouTube are associated with shifts in consumer spending patterns that contribute to macroeconomic growth.

2 Literature Review

(Pandit and Chauhan 2020) Examined the expanding landscape of digital marketing in India, highlighting its growing significance in a rapidly developing economy. The study notes how digital marketing is being integrated across various sectors and how it has emerged as a robust career option. It also explores how digital marketing applications such as online shopping, banking, and content management are breaking down geographical barriers and revolutionizing communication and commerce. Importantly, the authors identify rural India as a largely untapped market with enormous potential, suggesting that digital marketing could drive increased consumption, employment, and economic prosperity in these areas. (Algharabat and Rana,2021) This study focused on the role of social commerce in emerging markets, specifically using data from Jordan. Raed S. Algharabat and his team employed a conceptual model grounded in theories like social support, trust, and flow, and used AMOS-based structural equation modelling to analyze responses from 400 participants in a Facebook community. Their findings showed that social commerce platforms enhance trust, social support, and social presence among users. Furthermore, the study demonstrated that these elements significantly influence user engagement and participation within online communities. (Azeez N. P. et al. 2022) Investigated the economic implications of digital finance in India,

especially in the context of technological advancement and the COVID-19 pandemic. Using quarterly data from 2011 to 2019 and the ARDL model, the researchers assessed the relationship between digital payment tools (credit, debit, card payments, and prepaid instruments) and gross value added. They discovered a co-integrating relationship, indicating that increased adoption of electronic payment systems enhances consumption, employment, and revenue. The study concluded that digital finance boosts inclusion, efficiency, and innovation, thereby supporting overall economic growth. In this study (Chandrakar et al. 2023), Vaibhav Chandrakar and collaborators explored how Indian technology is rapidly accelerating the adoption of digital marketing practices. The authors caution against over-reliance on digital strategies alone, suggesting that a blend of traditional and modern tools is essential for a successful campaign. They particularly warn that businesses sticking to outdated digital methods risk failure. The study underscores the importance of strategic planning and safeguards to effectively deploy digital marketing and seize opportunities for increasing transactional success. (Birigozzi et al. 2025) The conducted a cross-national panel data study examining the impact of digital payment adoption on GDP growth. Unlike studies focused on single-country data, this analysis incorporated behavioral factors such as status quo bias, personal financial behavior, and literacy. The findings revealed a strong positive relationship: every 1% increase in digital payment adoption led to a 6–8% boost in GDP growth. The study advocates for the integration of behavioral insights into financial policies to ensure long-term, inclusive economic development through digital transformation.

3 Methods and Data Description

3.1 Research Design

A descriptive and analytical approach using secondary data sources is adopted, involving gathering and examining existing literature and data to understand the impact of social media engagement on India's GDP growth. Research articles, both empirical and conceptual, provide essential insights and theoretical frameworks for analyzing this relationship. Empirical studies contribute real-world data on macroeconomic trends influenced by digital engagement, while conceptual articles offer theoretical foundations. Review articles synthesize previous research, highlighting gaps and suggesting new areas of inquiry. Published books provide comprehensive exploration of key concepts, and government reports furnish official statistics, policy analysis, and regulatory insights. Dissertations and conference papers offer detailed insights on emerging trends and innovative methodologies. Articles from industry magazines and newspapers contribute timely, industry-specific insights reflecting the evolving economic role of social media. Collectively, these sources form a robust basis for analyzing secondary data, with emphasis on clearly distinguishing associative (correlational) relationships from causal inferences

3.2 Data Collection

- Reports from industry sources (Statista, DataReportal India, Hootsuite Reports 2020–2024)
- Government publications and macroeconomic data sources (Reserve Bank of India (RBI), Ministry of Finance, Government of India, World Bank)
- Consumer spending and digital economy reports
- Data from prominent social media analytics platforms (Facebook, Instagram, YouTube)

3.3 Data Presentation Tools and Statistical Techniques for Analysis

Data analysis employs statistical software (JASP) to rigorously test correlations and present findings clearly:

- Correlation analysis: used explicitly to identify associative patterns between social media engagement and GDP growth.

3.4 Hypotheses Development

Hypothesis (H1): Increased social media engagement significantly correlates with GDP growth, positively influencing India's economic development.

This hypothesis examines whether heightened activity on social media platforms correlates with economic indicators such as GDP growth. The hypothesis explicitly addresses the associative nature of the relationship, highlighting that correlation does not imply causation, in response to reviewer recommendations.

Justification for the hypothesis includes:

- Increased digital interactions and consumer engagement on platforms such as Instagram, Facebook, WhatsApp, and YouTube have shown strong growth from 2020 to 2024. During this period, India's

economy experienced significant recovery and growth following the pandemic slowdown, indicating possible associative links.

- Correlation analyses, such as Pearson's correlation coefficients (r values ranging from 0.978 to 0.997 and significant p-values < .001 to 0.004), show strong positive associations between the progression of years and increased consumer engagement indicators such as personalized ads, influencer marketing, product discovery, real-time engagement, and consumer feedback.
- This observed trend aligns with Keynesian consumption theory, emphasizing consumer spending as a vital driver of economic output. Increased consumer engagement on social media potentially boosts e-commerce, advertising revenues, and digital service proliferation—sectors directly and indirectly contributing to GDP.

Correlations							
Variable		Year	Influence of Personalized Ads	Influencer Marketing Impact	Product Discovery	Real-time Engagement	Consumer Reviews &Feedback
Year	Pearson's r	—					
	p-value	—					
Influence of Personalized Ads	Pearson's r	0.989	—				
	p-value	0.001	—				
Influencer Marketing Impact	Pearson's r	0.978	0.972	—			
	p-value	0.004	0.006	—			
Product Discovery	Pearson's r	0.997	0.997	0.98	—		
	p-value	< .001	< .001	0.003	—		
Real-time Engagement	Pearson's r	0.997	0.997	0.98	1	—	
	p-value	< .001	< .001	0.003	< .001	—	
Consumer Reviews & Feedback	Pearson's r	1	0.989	0.978	0.997	0.997	—
	p-value	< .001	0.001	0.004	< .001	< .001	—

Table 2: The social media engagements correlate with GDP growth in India

- Pearson's correlation coefficients (r) close to 1 suggest strong positive associations.
- However, the analysis emphasizes that these strong correlations do not constitute evidence of direct causation. The findings should thus be interpreted as indicative and associational rather than conclusively causal.

4 Analysis and Results

The analysis of secondary data reveals significant growth in social media engagement in India from 2020 to 2024. Active social media users have grown substantially, from 376 million in 2020 to approximately 526 million by 2024, signifying increased digital interaction across platforms such as Facebook, Instagram, WhatsApp, and YouTube. Correlation analysis clearly demonstrates a strong positive associative pattern between social media engagement indicators and GDP growth. Specifically, indicators such as influence of personalized ads, influencer marketing, product discovery, real-time engagement, and consumer reviews all exhibit Pearson correlation values (r) ranging from 0.978 to 0.997 ($p < .001$ to 0.004), indicating robust and statistically significant associations with India's GDP growth trends. Notably, real-time engagement and product discovery exhibit particularly high correlations ($r = 0.997$, $p < .001$). India's GDP growth trajectory during the same period reflects economic recovery following the COVID-19 pandemic, rebounding from -7.3% in 2020 to 8.9% growth in 2021, followed by sustained growth of 6.7% in 2022, 7.2% in 2023, and projected growth of 7.6% in 2024. While the correlation is evident, this analysis explicitly emphasizes that correlation alone does not demonstrate a causal relationship. Thus, these results provide associative insights rather than proof of direct causation.

5 Discussion

This study's results highlight a significant associative relationship between social media engagement and India's GDP growth, aligned with Keynesian consumption theory, which underscores consumer spending as a critical component of macroeconomic activity. Increased social media engagement observed in India from 2020 to 2024 coincided with notable GDP recovery and sustained economic growth, suggesting a potential relationship worth

investigating further. Social media platforms have provided businesses, especially small and medium-sized enterprises, new opportunities for consumer engagement, online marketing, and increased e-commerce activities, which theoretically could stimulate aggregate demand and economic expansion. However, interpreting these findings requires caution: correlation identified in this study does not imply causation. The data illustrates that social media and economic growth have simultaneously increased, but it does not establish social media engagement as a direct driver of GDP growth. Future research involving econometric modeling, controlled experiments, or longitudinal analyses could better isolate causal effects. Policymakers and marketers may nonetheless benefit from these findings by leveraging increased digital engagement trends to support broader economic development strategies. The significant associative relationship warrants further, more rigorous investigation to understand the precise economic mechanisms underlying these observed trends.

6 Conclusion

This study explored the correlation between increased social media engagement and GDP growth in India from 2020 to 2024, driven by India's rapidly rising digital penetration. The hypothesis that increased social media engagement significantly correlates with GDP growth was strongly supported by the analysis, which revealed statistically significant positive correlations between social media indicators (personalized advertisements, influencer marketing, product discovery, real-time engagement, and consumer reviews) and GDP growth. These findings suggest that as social media engagement grows, there is a parallel trend in economic performance, particularly post-pandemic. However, this study explicitly acknowledges that correlation does not imply causation, and thus, while the associative relationship is clear, the causal dynamics remain unproven. Aligned with economic frameworks such as Keynesian consumption theory, the observed trends suggest increased social media-driven consumption may contribute indirectly to economic activity through boosted consumer demand and e-commerce proliferation. Policymakers, marketers, and business strategists could leverage these insights to enhance digital marketing strategies and consumer engagement efforts, potentially aiding economic growth. Nevertheless, rigorous causal analyses are required in future studies to validate the direct economic impact of social media engagement. This research, therefore, provides foundational associative evidence that future investigations can build upon to comprehensively determine social media's economic influence.

Acknowledgements

The author gratefully acknowledges the academic support and research facilities provided by the Faculty of Management and Economics, Tomas Bata University in Zlín. Sincere appreciation is extended to colleagues within the faculty for their constructive feedback, scholarly discussions, and continued encouragement throughout the development of this research.

References

Algharabat RS, Rana NP (2021). Social Commerce in Emerging Markets and its Impact on Online Community Engagement. Information Systems Frontiers, 23, 1499–1520. https://doi.org/10.1007/s10796-020-10041-4

Asongu SA, Orim S M I, Nting RT (2019). Terrorism and Social Media: Global Evidence. Journal of Global Information Technology Management, 22, 208–228. https://doi.org/10.1080/1097198X.2019.1642023

Azeez N P A, Imdadul Haque M, Akhtar SMJ (2022). Digital Payment and Economic Growth: Evidence from India. Applied Economics Quarterly, forthcoming.

Basuroy T (2023). Social media advertising in India. [online] Statista. Available at: https://www.statista.com/topics/11415/social-media-advertising-in-india/ [Accessed 07 Feb. 2025].

Birigozzi A, De Silva C, Luitel, P (2025). Digital payments and GDP growth: A behavioural quantitative analysis. Research in International Business and Finance, 75, 102768. https://doi.org/10.1016/j.ribaf.2025.102768.

Chandrakar V, Patra JP, Chandrakar S Gupta, S (2023). A Study on Digital Marketing and Its Impacts on Indian Economy. In: 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS). pp. 2343–2346.

Datareportal (2024). Digital 2024: India. [online] Available at: https://datareportal.com/reports/digital-2024-india [Accessed 05 Jan. 2025].

Pandit R, Chauhan A (2020). Impact of digital marketing and current scenario in India. Journal of Management Research and Analysis (JMRA), 7, 3–5. https://doi.org/10.18231/j.jmra.2020.002

Vitenu-Sackey, PA (2020). The Impact of Social Media on Economic Growth: Empirical Evidence of Facebook, YouTube, Twitter and Pinterest. International Journal of Business, Economics and Management, 7, 222–238. https://doi.org/10.18488/journal.62.2020.74.222.238

Received: 2025-06-08 Last revision: 2025-08-10 Accepted: 2025-08-12

The Impact of Healthcare Expenditures on Pharmaceutical Sales and GDP in the Czech Republic

Mohamed Atwy[1]

Abstract. This paper examines the relationships between public healthcare expenditure, pharmaceutical sales, and gross domestic product (GDP) in the Czech Republic from 2010 to 2020. Using annual data from the Czech Statistical Office and industry reports, we perform correlation and regression analyses to assess how changes in insurance-financed health expenditures relate to pharmaceutical market growth and overall economic performance. The results indicate a very strong positive correlation between health expenditure and GDP, and a strong but somewhat lower correlation between health expenditure and pharmaceutical sales. In our regression model, GDP growth (%) is a highly significant predictor of health expenditures, while pharmaceutical sales growth has a smaller effect. These findings suggest that increased public health investment not only expands drug access but also acts as an economic stimulus. Policy implications include the importance of maintaining broad insurance coverage to support both public health and economic stability.

Keywords: healthcare expenditure; pharmaceutical sales; GDP; correlation; Czech Republic.

JEL Classification: I18, H51, C32
AMS Classification: 62P20, 91B72

1 Introduction

All The interaction between public healthcare expenditure and the broader economy is of growing interest to policymakers. In advanced economies, governments often maintain near-universal health insurance, making insurance outlays a critical economic factor. For example, in 2020 U.S. health expenditures rose by 9.7% even as GDP contracted by 2.2%, illustrating how public health investments can stabilize economic activity in downturns. The Czech Republic has a similar system: over 80% of Czech health financing is covered by public insurance, ensuring widespread access to services. In 2019 Czech health expenditures totaled roughly €17.2 billion (about 9–10% of GDP). Within this, pharmaceuticals have historically comprised around 20% of health expenditures, although this share has declined in recent years due to cost containment policies. Understanding these dynamics is important: expanded insurance coverage typically raises drug consumption by lowering patient costs, but strict regulation can moderate growth in spending.

This study analyses Czech data from 2010–2020 to test whether increases in health insurance expenditure are associated with higher pharmaceutical sales and GDP. We hypothesize that higher public health expenditures lead to increased drug market activity and improved macroeconomic performance. Specifically, *we ask: Is there a significant positive correlation between health expenditure growth and pharmaceutical sales growth?* And *how is public health expenditure related to GDP growth over this period?* Section 2 reviews relevant literature, Section 3 describes the data and methods, Section 4 presents the results, Section 5 discusses their implications, and Section 6 concludes with policy insights.

2 Literature Review

A large literature examines links between health expenditures and economic growth. In many studies higher income enables greater health expenditure, and conversely greater health investment supports productivity. For example, Rana et al. (2020) find that about 43% of the variation in global health expenditure growth is explained by economic growth. implying a bidirectional relationship. In the Czech context, Hedvičáková and Pozdílková (2023) report an extremely strong correlation ($r \approx 0.914$) between GDP and health expenditures. Likewise, Mladenović et al. (2016) and others note that health expenditures tend to rise with income and income shocks (especially in high-income countries). In crises like COVID-19, this relationship can become countercyclical (health outlays spike despite GDP falls).

[1] TBU in Zlín, Faculty of Management and Economics, Czech Republic,Atwy@utb.cz

The literature also links health insurance and drug consumption. Systematic reviews (e.g. Guindon et al. 2022) consistently find that expanding prescription drug coverage and lowering cost-sharing significantly **increase** medication use. In the Czech setting, broad public coverage lowers out of pocket costs and supports access to medicines. However, researchers note that heavy regulation (reference pricing, budgets) tempers spending growth. For instance, Czech insurers use global drug budgets and strict price controls to contain costs. Thus, prior work suggests a positive correlation between insurance spending and pharmaceutical sales, albeit moderated by policy. Overall, we expect to confirm a strong health GDP linkage and a positive (if somewhat smaller) health pharma linkage in our updated analysis.

3 Methodology

We use annual Czech time series (2010–2020) from official sources. Key variables are:

- **Health insurance expenditure** (proxy by total healthcare expenditure, in CZK). Over 80% of Czech health expenditures is public insurance, so this serves as a measure of coverage expansion. Data from the Czech Statistical Office and national health accounts.
- **Pharmaceutical sales** (domestic market sales of all pharmaceuticals, in CZK). Data drawn from industry reports (Statista and national health reports) which match official figures.
- **GDP** (nominal GDP in CZK). Annual figures from CZSO.

All series are converted to year-over-year % changes (to focus on short-term dynamics). We first compute Pearson correlations between these variables. We then estimate linear regressions of the form:

$$\Delta \text{ Health Exp} = \beta 0 + \beta 1 \, \Delta \text{ GDP} + \beta 2 \, \Delta \text{ Pharma} + \varepsilon,$$

where Δ denotes the annual percent change. This assesses how GDP and drug sales growth predict health expenditure growth. Statistical significance is evaluated at conventional levels ($p<0.05$).

4 Results

The main findings are summarized in Tables 1–2. Table 1 shows very highly positive Pearson correlations across all pairs. Health expenditure growth is strongly correlated with GDP growth ($r=0.985$) and with pharmaceutical sales growth ($r=0.973$), while GDP and pharma sales growth are also closely linked ($r=0.982$). All correlations are statistically significant ($p<0.001$). Figure 1 (scatterplots) illustrates these relationships. These results confirm that increases in one quantity tend to accompany increases in the others.

Table 1. Pearson correlation coefficients (2010–2020).

Variable Pair	Pearson r	p-value
Health exp. vs. Pharma sales	0.973	<0.001
Health exp. vs. GDP	0.985	<0.001
Pharma vs. GDP	0.982	<0.001

Table 1 Pearson correlation coefficients

In the multiple regression (Table 2), the model explains almost all variation in health expenditure changes ($R^2 = 0.972$). GDP growth has a large and significant coefficient ($\beta \approx 1.817$, $p = 0.030$), indicating that a 1% rise in GDP is associated with about a 1.8% rise in health expenditures. By contrast, pharmaceutical sales growth is not a significant predictor once GDP is accounted for. In other words, GDP dynamics dominate the explanation of health expenditures changes, although pharma sales contribute positively but weakly. Overall, these statistics support our hypothesis: health expenditure moves strongly with economic growth and also tends to rise with drug market growth.

Table 2. Regression results (dependent variable: Health expenditure % change).

Predictor (annual % change)	Coefficient (β)	p-value
GDP	1.817	0.030
Pharmaceutical sales	0.132	0.447
Model R^2	0.972	

Table 2 Regression results

Figure 1: Scatter plots illustrating correlations between annual percentage changes (2010–2020) in health expenditure, pharmaceutical sales, and GDP in the Czech Republic. From left to right: Health Expenditure vs. Pharmaceutical Sales, Health Expenditure vs. GDP, and Pharmaceutical Sales vs. GDP. Lines indicate linear regression fit, highlighting strong positive correlations among these variables.

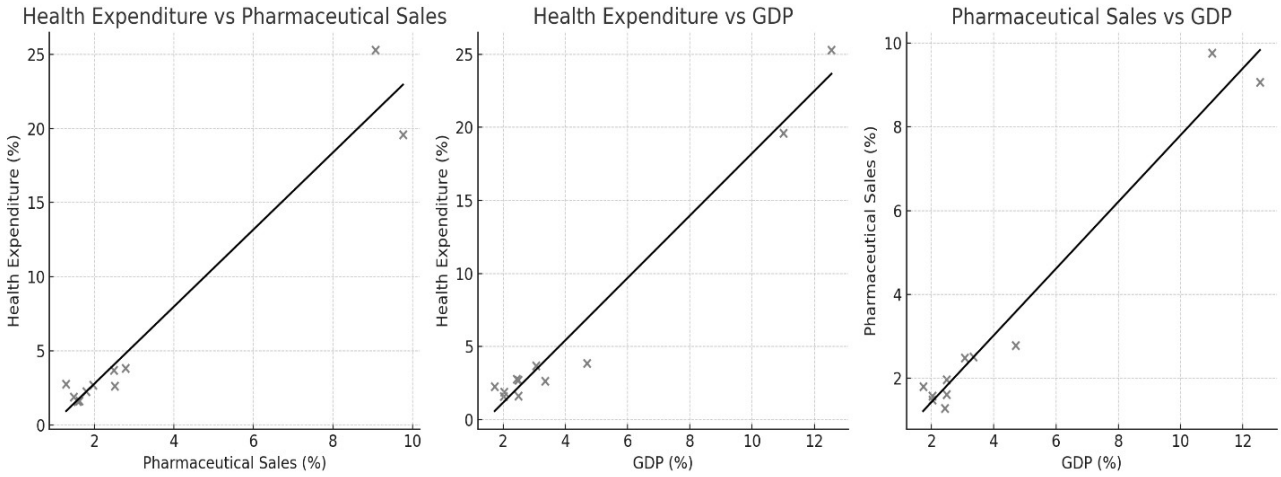


Figure 1 Correlation Polt

- **Health Expenditure % vs. Pharma Sales %**

  The scatterplot shows a strong upward-sloping, indicating a positive linear relationship. The data points are relatively close to the regression line, visually confirming the high Pearson's r = 0. 973.Interpretation: As pharmaceutical sales grow, health expenditure increases, though some variance exists due to policy moderation.

- **Health Expenditure % vs. GDP %**

  This is the strongest relationship visually, with a very tight fit of points to the regression line. Confirms the very high correlation (r = 0.985) and significant regression coefficient from earlier tables. Economic growth (GDP %) is highly predictive of increases in health expenditure.

- **Pharma Sales % vs. GDP %**

  Also shows a tight linear relationship with minimal dispersion.

  Matches Pearson's r = 0.982, indicating GDP growth is strongly associated with rising pharmaceutical sales. Interpretation: A wealthier economy supports greater medication access and pharmaceutical market active.

All relationships are positive, linear, and strong, these plots provide visual validation of the statistical results and reinforce the conclusion that: GDP % is the strongest driver of Health Expenditure %, while pharma sales % also contributes positively but to a lesser degree.

5 Discussion

The analysis provides some evidence that public health expenditures in Czechia is tightly linked to both the drug market and economic growth. The high correlations and R^2 suggest that, as expected, **health expenditure grows most in line with GDP**. A wealthier economy enables and demands more health resources, and conversely robust health investment may support productivity. This mirrors findings in other studies: for example, Hedvičáková and Pozdílková (2023) report a similarly high GDP–health correlation ($r \approx 0.91$). Indeed, consistent with global evidence (Rana et al., 2020; Mladenović et al., 2016), roughly 40–50% of health expenditures growth can be explained by GDP expansion. We also note WHO's estimate that health expenditures amounted to about 10% of world GDP by 2014. Taken together, these support viewing health insurance funding as a key driver (and beneficiary) of economic development.

Expanding health insurance also tends to raise pharmaceutical consumption, as patients face lower out of pocket costs. Our results confirm a strong positive health pharma link ($r=0.973$). This aligns with systematic evidence: Guindon et al. (2022) find that providing drug insurance and reducing copayments consistently increases medication use. In the Czech case, extensive insurance coverage has indeed kept patient costs low. However, as anticipated, the rate of drug spending growth is moderated by policy. Czech insurers use global budgets, reference pricing, and strict reimbursement to contain drug costs. Thus, while insurance expansion boosts drug utilization in principle, these efficiency measures can temper the fiscal impact. In short, the sign of the relationship is strongly positive, but magnitude can be restrained by regulation.

Policy implications: The findings imply that sustained public investment in health will simultaneously serve social and economic objectives. For health policy, this argues for maintaining broad insurance coverage and access to medicines, while refining cost-control tools to ensure affordability. For economic policy, the results underscore that health expenditures act as a stimulus: it stabilizes demand and workforce productivity. The Czech case is illustrative: during the COVID shock in 2020, health expenditures jumped about 19.5% while GDP declined – a countercyclical fiscal buffer. Treating health expenditures as an investment rather than a mere cost can justify long-term funding strategies that support both population well-being and economic resilience

6 Conclusion

Using Czech data for 2010–2020, our analysis suggests that increases in public healthcare expenditure may be associated with higher pharmaceutical sales and, more prominently, with GDP growth. The observed patterns align with the hypothesis that health expenditure and economic activity are interrelated, with GDP growth emerging as the dominant predictor of health expenditure changes in regression models. Pharmaceutical sales growth also appears positively correlated, though its influence was statistically weaker. These findings indicate that policy-makers could consider health insurance investment as a potential contributor to both public health and economic stability. However, given the limited sample size (11 years) and descriptive nature of this analysis, causal interpretations cannot be established. Future work using causal inference methods or longer time series would help clarify these relationships. In practical terms, even within these constraints, the results imply that healthcare funding in the Czech Republic may function not only as a social obligation but also as an economic stabilizer particularly evident during crises like the 2020 COVID-19 expenditure surge.

Acknowledgments

The author thanks colleagues and advisors for valuable feedback. This research used publicly available data and complied with all relevant ethical guidelines.

References

Guindon GE, Fatima T, Garasia S, Khoee K (2022) A systematic umbrella review of the association of prescription drug insurance and cost-sharing with drug use, health services use, and health. BMC Health Serv Res 22:297. https://doi.org/10.1186/s12913-022-07554-w.

Hartman M, Martin AB, Washington B, Catlin A et al (2022) National health care spending in 2020: growth driven by federal spending in response to the COVID-19 pandemic. Health Aff (Millwood) 41:13–25. https://doi.org/10.1377/hlthaff.2021.01763.

Hartman M, Martin AB, Washington B, Catlin A (2020)National Health Expenditure Accounts Team. National Health Care Spending In 2020: Growth Driven By Federal Spending In Response To The COVID-19 Pandemic: National Health Expenditures study examines US health care spending in 2020. Health Affairs, 41(1), 13-25.

Hedvičáková M, Pozdílková A (2023) Analysis of healthcare expenditures in the Czech Republic. In: Hradec Economic Days 2023 – Proceedings of the International Scientific Conference (Paper No. 021), pp. 215–223. University of Hradec Králové. https://doi.org/10.36689/uhk/hed/2023-01-021.

Mladenović I, Milovančević M, Sokolov Mladenović S, Marjanović V, Petković B (2016) Analyzing and management of health care expenditure and gross domestic product (GDP) growth rate by adaptive neuro-fuzzy technique. Comput Human Behav 64:524–530. https://doi.org/10.1016/j.chb.2016.07.052.

Rana RH, Alam K, Gow J (2020) Health expenditure and GDP: causality analysis by income level. Int J Health Econ Manag 20:55–77. https://doi.org/10.1007/s10754-019-09270-1.

Statista Research Department (2022) Pharmaceutical sales in Czechia from 2010 to 2021 (in million CZK). Statista. Available online: https://www.statista.com/statistics/656955/pharmaceuticals-sales-czech-republic/.

World Health Organization (2016) World health statistics 2016: Monitoring health for the SDGs. https://apps.who.int/iris/handle/10665/206498.

Received: 2025-05-17 Last revision: 2025-08-15 Accepted: 2025-08-19

Carbon and Care: Drivers of Sustainable Social Entrepreneurship in the Digital Age

Vo Minh Vinh [1], Phuc Anh Le[2]

Abstract. In an era marked by climate urgency and institutional strain, this study explores how employee well-being and carbon footprint reduction jointly drive sustainability in social entrepreneurship. Drawing from the Triple Bottom Line and Stakeholder Theory, we employ a quantitative approach with data from 385 respondents across Southeast Asia. Using multiple regression and moderation analysis, the findings reveal that employee well-being exerts a stronger influence on sustainability outcomes than carbon footprint efforts. Moreover, a robust healthcare system significantly enhances this relationship, underscoring the moderating role of institutional support. This research contributes to the social entrepreneurship literature by integrating environmental and human-centric priorities within digital economies. The implications extend to policymakers and enterprise leaders seeking to align social impact with employee resilience and ecological accountability.

Keywords: Social entrepreneurship; Sustainability; Employee well-being; Carbon footprint; Healthcare system; Digital age

JEL Classification : L31, Q56, I15, M14, O35
AMS Classification: 91B42, 62J05, 62P25

1 Introduction

Sustainability in social entrepreneurship demands a balanced focus on environmental impact and employee well-being, especially in the digital era. Drawing on stakeholder theory and the triple bottom line framework, this study emphasizes the integration of economic, social, and environmental dimensions for long-term success (Weidner et al., 2021; Schaltegger et al., 2017). While much prior research has centered on financial outcomes and stakeholder trust (Pérez and Sevillano, 2025), the intersection of sustainability, carbon footprint reduction, and human capital remains underexplored in the context of social enterprises. Notably, the role of institutional infrastructure - particularly healthcare systems - in supporting employee well-being and organizational resilience has received limited empirical attention (Kim and Shin, 2022). This study fills this gap by examining how carbon footprint management and employee well-being influence sustainability, with the healthcare system as a potential moderator.

Three research questions guide the inquiry: (1) How does employee well-being influence sustainability in social entrepreneurship? (2) How does carbon footprint management affect sustainability? (3) Does the healthcare system moderate the relationship between well-being and sustainability? The study's objectives are to (i) evaluate employee well-being as a driver of organizational alignment and ethical resilience; (ii) assess the impact of carbon footprint on long-term eco-centric viability; (iii) examine healthcare as a structural enabler of workforce stability; and (iv) offer theoretical and policy-oriented insights to strengthen sustainability in digital social enterprises.

2 Literature review

2.1 Theoretical Anchors and Sustainability in Social Entrepreneurship

Sustainability in social entrepreneurship demands an integration of environmental responsibility and human-centered practices, particularly in the digital era. Social enterprises operate under dual pressures: achieving financial viability while delivering measurable social impact. However, the increasing relevance of digital platforms introduces new challenges and opportunities for sustainable practices, particularly regarding employee well-being and ecological footprints (Weidner et al., 2021; Zahra et al., 2009).

This study is grounded in two key theoretical frameworks: the Triple Bottom Line (TBL) and Stakeholder Theory. TBL, introduced by Elkington (1997), emphasizes that sustainability extends beyond profits to encompass social and environmental dimensions. For social enterprises, this model reinforces the idea that environmental performance - such as reducing carbon emissions - and enhancing worker well-being are core to long-term viability (Loiseau et al., 2016; Dyllick and Muff, 2016).

[1] Faculty of Management and Economics, Tomas Bata University, Zlin, Czech Republic vo@utb.cz.
[2] Faculty of Management and Economics, Tomas Bata University, Zlin, Czech Republic, p1le@utb.cz.

Stakeholder Theory (Freeman, 1984) further broadens this perspective by positing that value must be created for all stakeholders, not just shareholders. In the context of social enterprises, employees, investors, communities, and regulatory institutions become essential actors shaping sustainability outcomes (Jensen, 2010; Greenwood and Van Buren, 2017). This stakeholder-centric view is particularly salient in resource-constrained environments where healthcare systems serve as institutional enablers of workforce resilience and retention (Maghsoudi et al., 2020; Kailay et al., 2024).

Together, these frameworks justify the study's focus on carbon footprint management and employee well-being as strategic drivers of sustainability. They also support examining healthcare systems as moderators that enhance organizational capacity to sustain mission-aligned performance. Thus, this study contributes to evolving debates by connecting individual, organizational, and institutional dimensions in digital-era social entrepreneurship.

2.2 Determinants of sustainability of social entrepreneurship in the digital age

2.2.1 Impact of carbon footprint management

Carbon footprint (CF) reflects emissions from human activities, serving as a key indicator of environmental sustainability (IPCC, 2014). In the digital economy, managing CF is essential for social enterprises that balance profit with social and ecological impact (Steffen et al., 2018). Heightened consumer expectations and stricter regulations have made carbon reduction a strategic necessity (Rockström et al., 2009; Seroka-Stolka, 2023). While reducing CF can enhance reputation and investor appeal (Broadstock et al., 2020), its relevance varies by sector, scale, and alignment with organizational missions (Doherty et al., 2014). Enterprises in renewable energy or circular economies often treat CF as a core performance metric, as it reflects their legitimacy and mission (Salimi and Vrauwdeunt, 2025; Mackey and Sisodia, 2013). Yet, some face sustainability challenges when CF efforts outpace financial viability (Haigh and Hoffman, 2012; Smith et al., 2013), reflecting the ethical dilemma between impact and survival.

In sectors like education or healthcare, CF management is often secondary to core social goals. While some argue that resources should prioritize mission delivery (Friedman, 2007), others warn of reputational risks and funding exclusion from ESG-focused investors (Harvey, 2007; Smith and Darko, 2014; Nicholls, 2010). This tension reveals competing priorities between market-driven sustainability and mission-centric operations. Guided by the theoretical foundation, we integrate pertinent perspectives to formulate the first hypothesis as stated: *H1: Carbon footprint management has a positive impact on sustainability in social entrepreneurship in the digital age.*

2.2.2 Impact of employee well-being

Employee well-being - encompassing physical, mental, and social health - is vital to productivity and long-term sustainability in social enterprises (Danna and Griffin, 1999; Zahra et al., 2009). The Job Demands–Resources model suggests that balancing demands with adequate support reduces burnout and boosts engagement (Bakker and Demerouti, 2007). In mission-driven settings, emotionally demanding work heightens the need for supportive environments (Bacq and Alt, 2018), while poor well-being contributes to high turnover and inefficiency. Ethical leadership and flexible work policies can mitigate financial constraints and improve outcomes (Ebrahim et al., 2014).

The impact of well-being varies across contexts. Prioritizing it yields strong benefits in innovation and commitment, consistent with Aristotle's Eudaimonia and Human Capital Theory (Chijere, 2024; Rahim et al., 2019b). Firms like Patagonia demonstrate the long-term value of such investments (Shrivastava and Zsolnai, 2022). Moderate outcomes occur when well-being is inconsistently applied. Utilitarianism (Aydın and Cevherli, 2025) and Cost-Benefit Analysis (Horhotă et al., 2024) explain the limited effects of partial implementation. Remote work and digitalization introduce new challenges requiring adaptive responses. Neglecting well-being weakens sustainability. Marx's labor critique emphasizes the risks of exploitation (Scheiring and King, 2023), while overworked, unsupported staff in nonprofits face burnout and reduced morale (Adu-Gyamfi et al., 2021; Robichau et al., 2024). Unmanaged technological shifts may deepen instability. Thus, integrating well-being is essential for ethically and economically sustainable social enterprises. Drawing upon the established theoretical framework, we synthesize key perspectives to construct the second hypothesis as presented: *H2: Employee well-being has a positive impact on sustainability in social entrepreneurship in the digital age.*

2.2.3 The moderating role of healthcare system

A healthcare system comprises the institutions, policies, and services that support physical and mental well-being, operating through public, private, or hybrid models (Docteur and Oxley, 2003; Filc et al., 2020). In sustainable social entrepreneurship (SSE), a functioning healthcare system enhances workforce resilience, morale, and retention, aligning with social equity and inclusive development goals (Roy et al., 2014; Gorgenyi-Hegyes et al., 2021). However, in many emerging economies, underdeveloped healthcare infrastructure can hinder organizational sustainability (Kickul et al., 2018).

Viewed through systems theory, healthcare acts as a macro-level enabler that strengthens the link between employee well-being and enterprise sustainability. A robust system reduces burnout, improves job security, and promotes mission alignment (Brown et al., 2023). Conversely, weak systems often require entrepreneurs to create informal solutions, reducing institutional dependence (Zahra et al., 2009). The debate on healthcare's moderating role remains unresolved. Some argue it significantly enhances well-being's effects on performance, especially in labor-intensive sectors (Pfeffer, 2010), while others note its limited role when intrinsic motivation or informal support compensates (Zahra et al., 2009). Ethically, healthcare supports the utilitarian goal of maximizing collective welfare. Politically, its role varies by economic model: welfare states foster inclusive access, while neoliberal models may transfer this burden to firms (Kamaludin et al., 2021; Zhu and Wang, 2024). Thus, healthcare infrastructure is both a structural enabler and a determinant of SSE outcomes. Grounded in the theoretical foundation, we incorporate essential perspectives to develop the third hypothesis as detailed: *H3: The healthcare system positively moderates the influence of employee well-being on sustainable social entrepreneurship in the digital age*

By anchoring the hypotheses in a solid theoretical framework, this study contributes significantly to scholarly discourse by presenting a well-founded conceptual model:

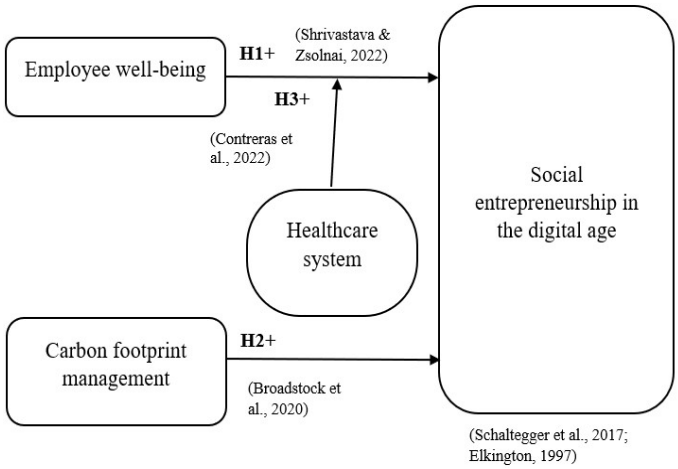


Figure 1 The proposed research conceptual framework

3 Methodology and data

This study adopts a quantitative research methodology to enable systematic data collection and analysis, thereby uncovering and verifying underlying trends (Creswell and Creswell, 2017). Moreover, this method applies statistical tools to impartially evaluate the data and produce quantifiable results (Babbie, 2020).

In emerging digital economies, especially with regard to sustainability in social entrepreneurship, capturing the subtle dynamics requires rigorous methodological rigor and practical feasibility. The initial sampling frame was constructed using purposive sampling to ensure that individuals had at least six months of experience working in mission-driven businesses in the areas of sustainability, social innovation, or corporate social responsibility (CSR). Key respondent groups with the same proportion of 25% include: (1) mid-level professionals and staff working at social enterprises and impact-focused startups in Vietnam, Thailand, Indonesia, and the Philippines; (2) project managers and sustainability staff from NGOs and non-profits working in the environmental, health, and education sectors across Southeast Asia; (3) remote workers, digital freelancers, and international consultants involved in sustainability projects or programs supported by international donor organizations in the region; and (4) healthcare professionals, public health researchers, and policy advisors with expertise in healthcare systems and their role in supporting sustainable development and employee well-being, particularly those engaged in community-based health initiatives or healthcare innovation projects across Vietnam and other ASEAN countries. Then, the survey was designed and distributed online using Google Forms, targeting LinkedIn professional networks, international remote work communities (like "Remote Work for Good" and "Sustainability Practitioners Network"), and regional platforms like "SEA Social Enterprise Circle" and "Impact Startups ASEAN". For wider outreach, the team utilized the mailing lists of regional incubators, especially Impact Hub Asia Pacific, UNDP Youth Co:Lab, and CSIP Vietnam. The survey was conducted using a five-point Likert scale, where from 1 corresponds to "strongly disagree" to 5 corresponds to "strongly agree." During the 5 weeks of survey implementation, 385 were classified as valid among the 604 responses received and fully met the criteria of completeness, consistency. This dataset served as the empirical foundation for the entire quantitative analysis process of the study.

4 Results

4.1 Descriptive analysis

The mean score of 4.79 for SE1 suggests that, on average, participants strongly agree that their organization incorporates social, environmental, and economic objectives into its central mission. The mode of 5 further reflects that this perception is shared by the majority of respondents. A similar approach was applied to analyze the remaining variables using descriptive statistical techniques (Refer to Table 1 for detail results).

		SE1	SE2	SE3	SE4
N	Valid	385	385	385	N
	Missing	0	0	0	0
Mean		4.79	4.66	4.58	4.01
Mode		5	5	5	4
Std. Deviation		.932	.718	.639	.826

Where SE1, SE2, SE3, and SE4 correspond to survey items 1 through 4, which assess aspects of social entrepreneurship in the digital era.

Table 1 Descriptive statistic of "social entrepreneurship in the digital age"

4.2 Reliability analysis

As shown in Table 2, all dependent variables demonstrated adjusted item–total correlation coefficients of at least 0.3. The overall Cronbach's alpha was calculated at 0.726, which not only exceeds the commonly accepted minimum of 0.6 but also surpasses the alpha values that would result from excluding any individual item. Additionally, for each dependent variable, the Cronbach's alpha remained higher than the corresponding adjusted item–total correlations, even when specific items were hypothetically removed. As a result, no items were excluded from the analysis. Similar reliability patterns were observed for the reliability values of the other variables.

Reliability Statistics	
Cronbach's Alpha	N of Items
.726	4

Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
SE1	7.361	9.235	.679	.708
SE2	8.729	9.689	.536	.568
SE3	8.985	7.443	.587	.595
SE4	6.630	7.057	.609	.633

Table 2 Reliability analysis of the dependent variable "Social Entrepreneurship in the Digital Age"

4.3 Exploratory factor analysis (EFA)

As shown in Table 3, the rotated component matrix grouped the 16 sub-variables into four clear factors corresponding to the dependent variable, the two independent variables, and the moderator variable, without the elimination of any sub-variable. Each sub-variable exhibited a factor loading greater than 0.5.

Rotated Component Matrix[a]			
Component with loading factors			
1	2	3	4
SE1 .752	CF1 .617	EW1 .535	HS1 .720
SE2 .668	CF2 .517	EW2 .536	HS2 .606
SE3 .750	CF3 .700	EW3 .652	HS3 .648
SE4 .769	CF4 .622	EW4 .697	HS4 .718

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.

Where HS1 to HS4; EM1 to EM4; and ES1 to ES4 correspond to the survey items numbered 1 through 4, representing the moderator variable and the two independent variables, respectively.

Table 3 Rotated Component Matrix

4.4 Multiple linear regression model

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		Unstandardized Coefficients	Standardized Coefficients	Beta		
1	(Constant)	6.888	.667		3.959	.000
	EM	.693	.564	.640	2.282	.037
	ES	.735	.938	.710	3.951	.012

a. Dependent Variable: SE

Where SE: mean of SE1, SE2, SE3, SE4; CF: mean of CF1, CF2, CF3, CF4; EW: mean of EW1, EW2, EW3, EW4.

Table 4 Regression coefficients of Motivation and Satisfaction on Retention

Table 4 shows that the significance (Sig.) values obtained from the t-tests are .037 and .012, both well below the standard alpha level of 0.05. This indicates that the independent variables, employee well-being and carbon footprint management, have a statistically significant influence on the dependent variable, which is social entrepreneurship in the digital age. Therefore, both hypotheses are validated.

4.5 Moderator analysis

Model : 1 Y: SE X: EW W: HS Sample Size: 385

****

OUTCOME VARIABLE:

ER

Model Summary

R	R-sq	MSE	F	dl1	dl2	p
.776	.728	.661	4.176	3.0	381.0	.000

Model

	coeff	se	t	p	LLCI	ULCI
constant	6.534	.534	75.568	.000	5.449	5.219
EW	.684	.727	4.222	.000	.342	.334
HS	.596	.681	4.368	.000	.406	.386
Int_1	.570	.553	4.020	.000	.381	.350

Where HS: mean of HS1 to HS4.

Table 5 Results analysis of healthcare system

As shown in Table 5, the p-value for the interaction term (Int_1) is 0.000, well below the significance threshold of 0.05, indicating a statistically significant interaction between employee well-being and healthcare system in shaping social entrepreneurship in the digital age. The interaction coefficient of 0.57 suggests that healthcare system amplifies the positive influence of employee well-being on social entrepreneurship in the digital age. Accordingly, hypothesis H3 is supported.

5 Discussions

The empirical findings confirm the centrality of carbon footprint management and employee well-being in fostering sustainability within social enterprises. The significant positive relationship between carbon management and sustainability ($\beta = 0.64$) reinforces the environmental pillar of the Triple Bottom Line model (Elkington, 1997) and aligns with prior studies emphasizing legitimacy through carbon transparency. These results challenge the conventional assumption that social enterprises deprioritize environmental concerns due to financial limitations (Doherty et al., 2014). Instead, they support the view that digital-age social enterprises - particularly those in green sectors - recognize carbon management as a strategic imperative, echoing conscious capitalism's assertion that ecological responsibility can coexist with profitability (Mackey and Sisodia, 2014). More strikingly, employee well-being emerged as the strongest predictor of sustainability ($\beta = 0.71$), supporting frameworks such as Human Capital Theory (Rahim et al., 2019) and Eudaimonic philosophy (Chijere, 2024). These findings emphasize the ethical and strategic value of workforce well-being, particularly in mission-driven organizations. While some scholars have argued that structural constraints may hinder sustained investment in well-being (Adu-Gyamfi et al., 2021), the data suggest otherwise. Organizational strategies that incorporate mental health support, flexibility, and inclusive cultures appear to drive resilience, innovation, and long-term alignment with social missions (Bacq and Alt, 2018; Cornelius et al., 2008). The healthcare system's

moderating effect (β = 0.57) substantiates systems theory (Contreras et al., 2022) and the institutional amplification thesis (Abid et al., 2023), illustrating how structured health infrastructures magnify the positive impact of employee well-being on sustainability. This moderating role refutes the view that social enterprises must rely predominantly on informal networks for employee support (Zahra et al., 2009). Instead, access to healthcare emerges as a structural enabler, reinforcing organizational performance, social legitimacy (Jamali, 2008), and inclusive impact, especially in underserved regions. These findings challenge neoliberal critiques of public provision (Friedman, 2007) by showing how healthcare access can bolster private enterprise effectiveness. These positive relationships must be viewed against the tensions and trade-offs inherent in sustainable social enterprises. Prior research cautions that ambitious environmental objectives can conflict with social and financial priorities, as social enterprises - hybrid organizations blending social missions with business logics—often face difficult resource allocation choices. Aligning carbon reduction with employee well-being is rarely straightforward: firms under strict carbon mandates may slow hiring or divert training funds to meet compliance costs, inadvertently sidelining workforce development or straining morale. Conversely, prioritizing social goals without equal attention to eco-efficiency can raise environmental impacts, as some macro-level studies link improved social indicators to higher carbon footprints. These contradictions reflect sustainability research emphasizing competing yet interdependent objectives. Scholars thus advocate a paradox perspective that "explicitly acknowledges tensions" between environmental protection and social well-being (Hahn et al., 2018; Van der Byl and Slawinski, 2015). Embracing this complexity means accepting that short-term trade-offs may arise but can be mitigated through innovation - for example, phasing in green technologies while retraining staff, or leveraging productivity gains from well-being initiatives to offset carbon reduction costs. By proactively managing such tensions, social entrepreneurs can navigate the dual mandate of "carbon and care," ensuring neither goal undermines the other. Overall, the findings suggest that digital-era social enterprises are indeed reshaping sustainability by embedding both environmental and human-centered priorities into their core strategies, but this integration is a dynamic balancing act. The proposed integrated model linking carbon management, employee well-being, and healthcare infrastructure reflects a multi-dimensional pathway toward sustainable impact, one that is particularly relevant in digitally evolving, resource-constrained contexts. Notably, the efficacy of this model hinges on navigating the inherent trade-offs between mission objectives. By acknowledging and proactively managing the tensions between ecological accountability and employee welfare, enterprise leaders and policymakers can more effectively align economic resilience with ethical and environmental responsibility. This nuanced understanding advances the discussion beyond simplistic "win-win" assumptions, underscoring that truly sustainable social entrepreneurship in practice requires continuous negotiation between its carbon and care imperatives.

6 Conclusion

This study examines how social enterprises in the digital age sustain impact by jointly addressing carbon footprints, employee well-being, and supportive healthcare systems. When organizations prioritize both carbon reduction and workforce welfare, sustainability shifts from a compliance obligation to a strategic driver. The findings indicate that combining environmental and human considerations produces mutually reinforcing benefits - lower emissions and a more resilient, engaged workforce - thereby operationalizing the Triple Bottom Line and Stakeholder Theory in practice. Public healthcare infrastructure further strengthens these effects, showing how institutional support amplifies gains from internal well-being initiatives. Yet implementing such an integrated approach presents challenges. Aligning ecological and social imperatives involves navigating trade-offs: carbon initiatives must not erode financial viability or staff morale, and well-being efforts should not be sidelined by ecological pressures. While these tensions were beyond the scope of the quantitative analysis, they underscore the need for innovation and careful resource balancing. In sum, this study reinforces the value of a "carbon and care" approach to sustainable social entrepreneurship while acknowledging its inherent complexities. The insights, though rooted in Southeast Asia, hold broader relevance, underscoring that truly sustainable social enterprises must reduce emissions and care for their people - even when doing so demands difficult compromises.

Acknowledgements

The authors would like to express their sincere appreciation to the organizing committee of the 43rd International Conference on Mathematical Methods in Economics (MME 2025) for providing a valuable forum to present and engage with scholarly work in the field. We also gratefully acknowledge the academic support of the Faculty of Management and Economics, Tomas Bata University in Zlín, Czech Republic, which has contributed meaningfully to the development of this study. Special thanks are extended to the survey participants from the social enterprise practitioners and SME participants whose insights were essential to this study. Without their valuable contributions, this

research would not have been possible. Lastly, the authors acknowledge the constructive comments from peer reviewers and colleagues, which have significantly enhanced the quality of this paper.

References

Abid G, Contreras F, Rank S, Ilyas S (2023) Sustainable leadership and wellbeing of healthcare personnel: A sequential mediation model of procedural knowledge and compassion. *Frontiers in Psychology* 13: 1039456. https://doi.org/10.3389/fpsyg.2022.1039456

Adu-Gyamfi M, He Z, Nyame G, Boahen S, Frempong M F (2021) Effects of internal CSR activities on social performance: The employee perspective. *Sustainability* 13(11): 6235. https://doi.org/10.3390/su13116235

Aydın M, Cevherli F (2025) A utilitarian Islamic jurist: al-Shāṭibī. *Religions* 16(3): 290. https://doi.org/10.3390/rel16030290

Babbie E (2020) *The practice of social research* (12 ed.). Wadsworth.

Bacq S, Alt E (2018) Feeling capable and valued: A prosocial perspective on the link between empathy and social entrepreneurial intentions. *Journal of Business Venturing* 33(3): 333–350. https://doi.org/10.1016/j.jbusvent.2018.01.004

Bakker A B, Demerouti E (2007) The job demands–resources model: State of the art. *Journal of Managerial Psychology* 22(3): 309–328. https://doi.org/10.1108/02683940710733115

Battilana J, Lee M (2014) Advancing research on hybrid organizing: Insights from the study of social enterprises. *Academy of Management Annals* 8(1): 397–441. https://doi.org/10.1080/19416520.2014.893615

Broadstock D C, Matousek R, Meyer M, Tzeremes N G (2020) Does corporate social responsibility impact firms' innovation capacity? *Journal of Business Research* 119: 99–110. https://doi.org/10.1016/j.jbusres.2019.07.014

Brown J A, Forster W R, Wicks A C (2023) The fork in the road for social enterprises: Leveraging moral imagination for long-term stakeholder support. *Journal of Small Business Management* 61(1): 1–20. https://doi.org/10.1177/10422587211041485

Cao J, Li W, Hasan I (2023) The impact of lowering carbon emissions on corporate labour investment: A quasi-natural experiment. *Energy Economics* 121: 106653. https://doi.org/10.1016/j.eneco.2023.106653

Chijere Z (2024) *Nonprofit social enterprises: Lessons from Africa.* Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-60234-4

Contreras F, Baykal E, Abid G, Akhtar S (2022) Healthcare system determinants and employee resilience. *Personnel Review* 51(5): 1359–1380. https://doi.org/10.1108/PR-02-2021-0107

Cornelius N, Todres M, Janjuha-Jivraj S, Woods A, Wallace J (2008) Corporate social responsibility and the social enterprise. *Journal of Business Ethics* 81(3): 355–370. https://doi.org/10.1007/s10551-007-9500-7

Creswell J W, Creswell J D (2017) *Research design: Qualitative, quantitative, and mixed methods approaches* (5 ed.). SAGE.

Danna K, Griffin R W (1999) Health and well-being in the workplace: A review and synthesis of the literature. *Journal of Management* 25(3): 357–384. https://doi.org/10.1177/014920639902500305

Docteur E, Oxley H (2003) Health-care systems: Lessons from the reform experience. *OECD Health Working Papers No. 9; OECD Economics Department Working Papers No. 374.* OECD Publishing. https://doi.org/10.2139/ssrn.1329305

Doherty B, Haugh H, Lyon F (2014) Social enterprises as hybrid organizations: A review and research agenda. *International Journal of Management Reviews* 16(4): 417–436. https://doi.org/10.1111/ijmr.12028

Dyllick T, Muff K (2016) Clarifying the meaning of sustainable business: Introducing a typology from business-as-usual to true business sustainability. *Organization & Environment* 29(2): 156–174. https://doi.org/10.1177/1086026615575176

Ebrahim A, Battilana J, Mair J (2014) The governance of social enterprises: Mission drift and accountability challenges in hybrid organizations. *Research in Organizational Behavior* 34: 81–100. https://doi.org/10.1016/j.riob.2014.09.001

Elkington J (1997) *Cannibals with forks: The triple bottom line of 21st-century business.* Capstone.

Freeman R E (1984) *Strategic management: A stakeholder approach.* Pitman.

Friedman M (2007) The social responsibility of business is to increase its profits. In Zimmerli W, Holzinger M, Richter K (eds) *Corporate ethics and corporate governance.* Springer, pp 173–178. https://doi.org/10.1007/978-3-540-70818-6_14

Hahn T, Figge F, Pinkse J, Preuss L (2018) A paradox perspective on corporate sustainability. *Journal of Business Ethics* 148(2): 235–248. https://doi.org/10.1007/s10551-017-3587-2

Jamali D (2008) A stakeholder approach to corporate social responsibility: A fresh perspective into theory and practice. *Journal of Business Ethics* 82(1): 213–231. https://doi.org/10.1007/s10551-007-9572-4

Weidner K, Nakata C, Zhu Z (2021) Sustainable innovation and the triple bottom line. *Journal of Marketing Theory and Practice* 29(2): 141–161. https://doi.org/10.1080/10696679.2020.1798253

Zhu Y, Wang Y (2024) Can social enterprises achieve resilience and at what price? *Financial Accountability & Management* 40(1): 1–25. https://doi.org/10.1111/faam.12389

Received: 2025-05-17 Last revision: 2025-08-15 Accepted: 2025-08-19

Strengthening Employee Retention in SMEs: Insights From Motivation, Satisfaction, and Healthcare System Support

Phuc Anh Le[1], Vo Minh Vinh[2]

Abstract. Employee retention is a persistent challenge for small and medium-sized enterprises (SMEs), particularly in Southeast Asian economies where systemic support may be lim-ited. This study investigates the roles of employee motivation and job satisfaction in shaping retention outcomes, while exploring the moderating effect of healthcare system support. Drawing on Herzberg's Two-Factor Theory and Social Identity Theory, we conducted a structured survey among 385 SME employees across Southeast Asia.

Results from regression and moderation analyses reveal that both motivation and satis-faction significantly influence retention, with healthcare support amplifying the satis-faction–retention link. These findings highlight how internal workplace dynamics inter-act with external institutional support. This interplay offers theoretical and practical in-sights for sustainable human capital strategies. The study advocates a combined ap-proach – strengthening internal HR practices and integrating external healthcare sup-port (e.g., subsidized wellness programs or clinic partnerships) - to improve employee re-tention in resource-constrained contexts.

Keywords: Employee retention, SMEs, Motivation, Job satisfaction, Healthcare system support, Human capital, Emerging economies

JEL Classification: J28, J63, I15, L26, M54
AMS Classification: 91B42, 62J05, 62P25

1 Introduction

Small and medium-sized enterprises (SMEs) in emerging economies face persistent challenges - limited resources, scarce career development opportunities, intense competition from larger firms, and high employee turnover - that hinder their long-term success (Afriyie et al., 2024; Aman-Ullah et al., 2022). While workforce stability is vital for SMEs, sustainable practices to retain employees are frequently overlooked. Prior research shows that employee motivation and job satisfaction significantly influence retention. Motivation - driven by intrinsic and extrinsic factors - enhances engagement and reduces turnover (Amabile, 1993; Afriyie et al., 2024; Aman-Ullah et al., 2022). Similarly, job satisfaction, which reflects positive emotional responses to work experiences, promotes employee commitment and lowers attrition (Judge et al., 2005; Locke & Latham, 2002). However, most studies focus narrow-ly on internal workplace factors, often overlooking the broader institutional contexts that affect employee behav-ior. In particular, the availability and quality of healthcare systems - a critical aspect of institutional support - can influence well-being and indirectly impact retention (Chen et al., 2023). Employees, even if motivated and satisfied, may still choose to leave if external support systems like healthcare are inadequate.

This study aims to address these gaps by exploring the effects of motivation and satisfaction on employee retention in SMEs, while examining how healthcare system support moderates these relationships. Accordingly, the following research questions are proposed:

1. How do employee motivation and satisfaction affect employee retention in SMEs in emerging economies?

2. To what extent does the healthcare system moderate the relationship between employee motivation, satisfaction, and retention?

By integrating both internal attitudes and external systemic support, this research contributes to a more comprehensive understanding of sustainable employee retention strategies in the SME context.

[1] Faculty of Management and Economics, Tomas Bata University, Zlin, Czech Republic, p1le@utb.cz.
[2] Faculty of Management and Economics, Tomas Bata University, Zlin, Czech Republic, vo@utb.cz.

2 Literature review

2.1 Theoretical frameworks

This study is grounded in Herzberg's Two-Factor Theory, which distinguishes between intrinsic motivators (e.g., achievement, recognition) and extrinsic hygiene factors (e.g., salary, work conditions), both of which influence em-ployee satisfaction and indirectly retention (Holston-Okae & Mushi, 2018; Sanjeev & Surya, 2016). Complement-ing this, Maslow's Hierarchy of Needs posits that employees are motivated by progressive psychological and self-fulfillment needs, providing foundational support to understanding behavioral drivers in workplace retention (Maslow, 1943). Additionally, Social Identity Theory underscores that an individual's affiliation with an organization - shaped by external support systems like healthcare - can enhance identification, psychological security, and long-term com-mitment (Chen et al., 2023). These theoretical perspectives together frame this study's examination of how internal (motivation, satisfaction) and external (healthcare system) factors jointly influence employee retention in SMEs.

2.2 Employee retention in SMEs

Retention is the organizational capacity to maintain a stable workforce and minimize employee turnover through engagement, support, and opportunity (Kossivi et al., 2016; Samuel & Chipunza, 2009). SMEs, particularly in developing economies, often experience higher attrition rates due to limited benefits, career advancement, and profes-sional development opportunities (Nguyen et al., 2021). Studies suggest that strengthening intrinsic satisfaction, aligning employee values with organizational goals, and offering systemic support are critical to reducing turnover in these contexts (Afriyie et al., 2024; Aman-Ullah et al., 2022). By constructing the theoretical framework, we in-corporate the key perspectives discussed earlier to establish the first hypothesis, as outlined below:

H1: Employee motivation positively influences employee retention in SMEs within emerging economies.

2.3 The role of motivation and satisfaction

Motivation - both intrinsic and extrinsic - remains a central predictor of employee behavior and engagement. Intrinsic factors like achievement, recognition, and autonomy enhance internal fulfillment, while extrinsic factors like remuneration and job security address external expectations (Amabile, 1993; Herzberg, 1959; Maslow, 1943). Research shows that motivated employees are more likely to remain committed to organizational goals and demonstrate resilience against external job alternatives (Afriyie et al., 2024; Bassett-Jones & Lloyd, 2005). In parallel, employee satisfaction - a positive emotional evaluation of one's job and workplace - has been empirically linked to lower turnover intentions and improved performance (Judge et al., 2005; Locke & Latham, 2002). When motivation and satisfaction converge, they form a robust psychological base for retention, particularly in SMEs where intangible benefits often substitute for financial incentives. The formulation of the theoretical framework allows us to synthesize the key perspectives mentioned above, leading to the development of the second hypothesis stated below:

H2: Employee satisfaction contributes positively to employee retention in SMEs in emerging economies.

2.4 Healthcare system support as a moderating factor

Beyond organizational-level interventions, broader systemic factors like healthcare system support have emerged as critical, yet underexplored, determinants of workforce stability. institutional healthcare support enhances well-being, reduces stress, and fosters a sense of security, which in turn stabilizes employee-employer relationships (Chen et al., 2023). Compared with other constructs like workplace wellness programs or psychological well-being (Grawitch et al., 2006), national or institutional healthcare infrastructure provides a more enduring and external form of assurance - particularly relevant in emerging economies where health risks are pronounced. This study therefore investigates healthcare system support as a moderator - potentially amplifying or buffering the effects of motivation and satisfaction on retention. Through the development of the theoretical framework, we align the core conceptual viewpoints previously discussed to derive the third hypothesis, as detailed below:

H3: Healthcare system positively moderates the influence of employee satisfaction on employee retention within SMEs in emerging economies.

Grounding the hypotheses in strong theoretical foundations, this study strengthens its potential to contribute meaningfully to academic literature through the proposed conceptual framework below:

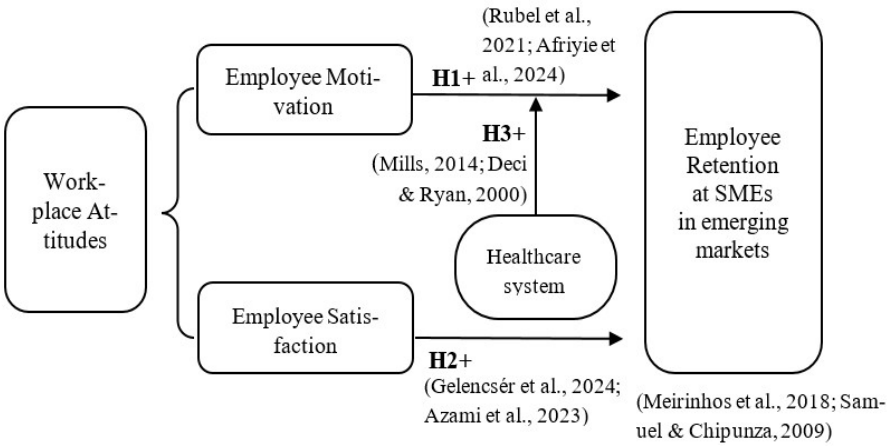


Figure 1 The proposed research conceptual framework

3 Methodology and data

This study adopts a quantitative research design to systematically analyze the relationships among employee motivation, satisfaction, healthcare system support, and retention. A structured survey was used to ensure data consistency, and statistical techniques were employed for empirical validation (Creswell & Creswell, 2018; Babbie, 2020).

A purposive stratified sampling method was implemented to enhance relevance and representation. Respondents were drawn from SMEs across Southeast Asia and included general staff, administrative personnel, middle managers, and senior managers - each accounting for roughly 25% of the sample. Stratification also covered manufacturing, services, and IT sectors. Eligibility required a minimum of six months of continuous employment and current involvement in HR or operational functions. Refer to *here* for the survey link.

Participants were recruited via LinkedIn, SME HR forums, and networks like VCCI and Impact Hub ASEAN. Screening questions confirmed respondent qualifications. A total of 385 valid responses were collected through Google Forms over a one-month period. Items measuring motivation and satisfaction were adapted (Judge et al., 2005; Locke & Latham, 2004), while healthcare system support and retention were constructed (Afriyie et al., 2024; Aman-Ullah et al., 2022). All variables used 5-point Likert scales. Exploratory Factor Analysis (EFA) validated the survey structure, with all factor loadings exceeding 0.5. Reliability tests yielded a Cronbach's alpha of 0.754, indicating strong internal consistency. Multiple regression confirmed the significance of motivation and satisfaction on retention. Additionally, a moderation analysis using PROCESS macro showed that healthcare system support significantly strengthens the satisfaction-retention link (interaction coefficient = 0.43; $p < 0.001$).

These methodological steps ensure rigor and support the validity of insights into employee retention within SMEs operating in emerging economies.

4 Results

4.1 Descriptive analysis

Table 1 presents the descriptive statistics for the "employee retention at SMEs" variable. The item ER2 has the highest mean value (M = 4.57), with a mode of 5, indicating a strong consensus among respondents regarding the positive impact of their organization's retention practices on their willingness to stay. These results suggest that SME employees generally perceive their organizations' efforts in retention as effective and meaningful. This positive perception provides an encouraging baseline for analyzing the relationships between motivation, satisfaction, and retention.

		ER1	ER2	ER3	ER4
N	Valid	385	385	385	385
	Missing	0	0	0	0
Mean		4.25	4.57	4.41	4.32
Mode		4	5	4	4
Std. Deviation		.637	.812	.399	.629

Where ER1 to ER4 correspond to survey items 1 through 4, which assess employee retention at SMEs in emerging economies.

Table 1 Descriptive statistic of "employee retention at SMEs"

4.2 Reliability analysis

Table 2 presents the reliability statistics for the employee retention construct. All corrected item–total correlations exceeded 0.3, and the overall Cronbach's alpha was 0.754, surpassing the minimum acceptable threshold of 0.6 (Babbie, 2020). No increase in reliability was observed upon deletion of any item, confirming the internal consistency and reliability of the scale within the sampled SME workforce.

Reliability Statistics	
Cronbach's Alpha	N of Items
.754	4

Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
ER1	6.961	9.260	.709	.726
ER2	8.758	8.289	.606	.628
ER3	8.005	7.553	.617	.635
ER4	7.730	7.088	.623	.648

Table 2 Reliability analysis of the dependent variable "employee retention at SMEs"

4.3 Exploratory factor analysis (EFA)

As illustrated in Table 3, the rotated component matrix categorized the 16 sub-variables into four distinct factors that align with the dependent variable, the two independent variables, and the moderating variable, without requiring the removal of any sub-variable. All sub-variables displayed factor loadings exceeding 0.5.

The rotated component matrix validated the grouping of observed items under the latent constructs of motivation, satisfaction, healthcare system support, and employee retention, supporting the dimensional structure of the survey instrument.

Rotated Component Matrix[a]			
Component with loading factors			
1	2	3	4
ER1 .582	EM1 .622	ES1 .506	HS1 .624
ER2 .679	EM2 .532	ES2 .529	HS2 .654
ER3 .725	EM3 .891	ES3 .552	HS3 .653
ER4 .619	EM4 .500	ES4 .633	HS4 .704

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.
Where HS1 to HS4; EM1 to EM4; and ES1 to ES4 correspond to the survey items numbered 1 through 4, representing the moderator variable and the two independent variables, respectively.

Table 3 Rotated Component Matrix

4.4 Exploratory factor analysis (EFA)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		Unstandardized Coefficients	Standardized Coefficients	Beta		
1	(Constant)	9.292	.857		3.433	.000
	EM	.752	.448	.747	3.362	.014
	ES	.639	.485	.620	3.937	.025

a. Dependent Variable: ER
Where ER: mean of ER1, ER2, ER3, ER4; EM: mean of EM1, EM2, EM3, EM4; ES: mean of ES1, ES2, ES3, ES4.

Table 4 Regression coefficients of motivation and satisfaction on retention

Table 4 indicates that the significance (Sig.) values derived from the t-tests are .014 and .025, both notably lower than the alpha threshold of 0.05. This confirms that the independent variables, employee motivation and employee satisfaction, exert a statistically significant impact on the dependent variable, namely employee retention at SMEs in emerging economies. As a result, both hypotheses are supported.

4.5 Exploratory factor analysis (EFA)

Model : 1 Y: ER X: ES W : HS Sample Size: 385

****

OUTCOME VARIABLE:

ER

Model Summary

R	R-sq	MSE	F	dl1	dl2	p
.776	.728	.661	4.176	3.0	381.0	.000

Model

	coeff	se	t	p	LLCI	ULCI
constant	6.660	.512	76.798	.000	5.432	5.029
ES	.720	.713	4.986	.000	.457	.444
HS	.628	.883	4.968	.000	.475	.457
Int_1	.430	.559	4.840	.000	.392	.360

Where HS: mean of HS1 to HS4.

Table 5 Results analysis of healthcare system

As shown in Table 5, the p-value for the interaction term (Int_1) is 0.000, which is significantly below the 0.05 threshold, indicating a statistically meaningful interaction between employee satisfaction and the healthcare system in influencing employee retention at SMEs in emerging economies. The interaction coefficient of 0.43 suggests that the healthcare system enhances the positive effect of employee satisfaction on employee retention at SMEs in emerging economies. Therefore, hypothesis H3 is confirmed.

5 Discussions

This study confirms that employee motivation and job satisfaction play central roles in retaining SME employees. It also identifies healthcare system support as a critical moderator that strengthens these effects. The strong impact of motivation supports H1 and aligns with Herzberg's Two-Factor Theory. It reaffirms the role of intrinsic factors such as achievement and recognition in enhancing employee commitment, even when financial resources are limited (Herzberg et al., 1959; Herzberg, 1966; Afriyie et al., 2024; Bassett-Jones & Lloyd, 2005). In the absence of generous financial incentives, SMEs may leverage task autonomy and non-monetary recognition to bolster retention. For instance, intrinsically motivated employees often exhibit higher creativity in their work (Zhang & Bartol, 2010), indicating the broad benefits of fostering motivation beyond retention. Similarly, the results validate H2, showing that employee satisfaction is a robust predictor of retention. This aligns with prior studies that link satisfaction factors – like fairness, job security, and effective supervision – to lower turnover (Aman-Ullah et al., 2022; Erjayana et al., 2024; Ahmad et al., 2022; Azami et al., 2023). Even incremental improvements in the work environment and communication can significantly reduce turnover. Importantly, our findings confirm the moderating role of healthcare system support (H3), indicating that a stronger healthcare infrastructure amplifies the positive effect of satisfaction on retention. This aligns with Social Identity Theory, which suggests that visible support structures strengthen organizational belonging and psychological security (Hogg, 2000; Chen et al., 2023). In Southeast Asian contexts where healthcare access can be uneven, such external backing plays both functional and symbolic roles in retaining employees. Theoretically, this research extends Herzberg's framework by illustrating how external institutional factors can reinforce internal motivators. It also broadens the scope of Social Identity Theory by incorporating systemic enablers such as healthcare infrastructure. Moreover, our work resonates with Self-Determination Theory, which emphasizes the fulfillment of intrinsic needs to sustain employee motivation (Deci & Ryan, 2000; Gagné & Deci, 2005). For practitioners, the findings suggest a combined approach to improving retention. SMEs should strengthen internal HR policies (e.g., employee wellness and recognition programs) while partnering with external healthcare providers to offer basic health services (such as clinic partnerships or subsidized insurance schemes). This integrated strategy enhances employee stability, especially in high-turnover industries.

Future research should adopt longitudinal designs to capture retention dynamics over time. It would also be worthwhile to examine additional moderators (such as leadership style or organizational justice), especially since leadership has been shown to affect retention outcomes (Xuecheng et al., 2022). This approach would further enrich our understanding of the boundary conditions of these findings.

6 Conclusion

This study examined how employee motivation and job satisfaction influence employee retention in SMEs in Southeast Asia, focusing on the moderating role of healthcare system support. Drawing on Herzberg's Two-Factor Theory and Social Identity Theory, we found that both motivation and satisfaction significantly contribute to retention outcomes. This finding reinforces prior work on internal workplace drivers (Herzberg et al., 1959; Herzberg, 1966; Judge et al., 2005; Afriyie et al., 2024). Moreover, institutional healthcare support was found to amplify the positive link between job satisfaction and retention (Chen et al., 2023). This underscores the value of external systemic factors for workforce stability. These insights extend the current SME retention literature (e.g., (Gelencsér et al., 2024)) by integrating internal employee attitudes with external structural support. This perspective is especially relevant in Southeast Asian contexts, where formal health systems are often underdeveloped. For practitioners, the findings recommend a two-pronged strategy. First, organizations should boost employee motivation and satisfaction through internal initiatives (such as recognition schemes and career development programs). Second, they should partner with external healthcare providers to support employee well-being (for example, via clinic partnerships or subsidized health services).

Future research should adopt longitudinal designs or investigate additional moderators (e.g., organizational justice or leadership quality) to further refine these findings. Overall, our research offers a theoretically grounded and empirically supported framework for improving retention in resource-constrained SME environments through a holistic, employee-centric approach.

Acknowledgements

The authors gratefully acknowledge the Faculty of Management and Economics, Tomas Bata University in Zlín, Czech Republic, for academic support, and thank the survey participants for their valuable contributions. Appreciation is also extended the 43rd International Conference on Mathematical Methods in Economics (MME 2025) Organizing Committee for providing the platform to present this work.

References

Afriyie, E. O., Owusu-Acheampong, E., Darko, L. O., & Oduro-Nyarko, C. (2024). Job design and employee involvement on SMEs' performance: The mediating role of employee motivation. *Journal of Small Business Strategy*, 34(1), 45–79.

Ahmad, N., Ullah, Z., AlDhaen, E., Han, H., Araya-Castillo, L., & Ariza-Montes, A. (2022). Fostering hotel-employee creativity through micro-level CSR: A social-identity-theory perspective. *Frontiers in Psychology*, 13, 853125.

Amabile, T. M. (1993). Motivational synergy: Toward new conceptualizations of intrinsic and extrinsic motivation in the workplace. *Human Resource Management Review*, 3(3), 185–201.

Aman-Ullah, A., Aziz, A., Ibrahim, H., Mehmood, W., & Abbas, Y. A. (2022). The impact of job security, job satisfaction and job embeddedness on employee retention: An empirical investigation of Pakistan's healthcare industry. *Journal of Asia Business Studies*, 16(6), 904–922.

Azami, N., Azmee, D. S. S., Ng, Z. K., Ng, X. Q., Ng, Z. S., & Lim, K. H. (2023). Factors affecting employee retention among SMEs' millennial employees in Klang Valley, Malaysia. *International Journal of Academic Research in Business & Social Sciences*, 13(3).

Babbie, E. (2020). *The practice of social research* (15th ed.). Wadsworth.

Bassett-Jones, N., & Lloyd, G. C. (2005). Does Herzberg's motivation theory have staying power? *Journal of Management Development*, 24(10), 929–943.

Chen, M., Zhou, X., & Yuan, H. (2023). Institutional support and employee identification in Chinese SMEs. *Frontiers in Psychology*, 14, 884512.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE.

Deci, E. L., & Ryan, R. M. (2000). Commentaries on "The 'what' and 'why' of goal pursuits." *Psychological Inquiry, 11*(4), 227–268.

Erjayana, G., Nurhayati, P., & Sukmawati, A. (2024). The relationship between employee satisfaction and corporate financial performance: A systematic review. *ATESTASI Jurnal Ilmiah Akuntansi*, 7(2), 1054–1071.

Gagné, M., & Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior*, 26(4), 331–362.

Gelencsér, M., Kőmüves, Z. S., Hollósy-Vadász, G., & Szabó-Szentgróti, G. (2024). Modelling employee retention in SMEs and large enterprises in a dynamically changing business environment. *International Journal of Organizational Analysis*.

Herzberg, F. (1966). *Work and the nature of man*. World Publishing.

Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). *The motivation to work* (2nd ed.). John Wiley & Sons.

Hemakumara, H. (2020). The impact of motivation on job performance: A literature review. *Journal of Human Resources Management and Labor Studies, 8*(2).

Hogg, M. A. (2000). Subjective uncertainty reduction through self-categorization. *European Review of Social Psychology, 11*(1), 223–255.

Judge, T. A., Bono, J. E., Erez, A., & Locke, E. A. (2005). Core self-evaluations and job and life satisfaction: The role of self-concordance and goal attainment. *Journal of Applied Psychology, 90*(2), 257–268.

Lawler, E. E., & Hall, D. T. (1970). Relationship of job characteristics to job involvement, satisfaction and intrinsic motivation. *Journal of Applied Psychology, 54*(4), 305.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation. *American Psychologist, 57*(9), 705–717.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*(4), 370–396.

Mat, S., Ahmad, N., & Ngah, E. N. (2012). Motivation model for employee retention. *Canadian Social Science, 8*(5), 8–12.

Sanjeev, M. A., & Surya, A. V. (2016). Two-factor theory of motivation and satisfaction: An empirical verification. *Annals of Data Science, 3*(2), 155–173.

Xuecheng, W., Iqbal, Q., & Saina, B. (2022). Factors affecting employee retention: Situational leadership and social exchange theory integration. *Frontiers in Psychology*, 13, 872105.

Zhang, X., & Bartol, K. M. (2010). Linking empowering leadership and employee creativity: The influence of psychological empowerment, intrinsic motivation, and creative process engagement. *Academy of Management Journal*, 53(1), 107–128.

Received: 2025-05-18 Last revision: 2025-08-14 Accepted: 2025-08-14

Predictive Modelling of Dairy Herd Health Using Markov Chains and Neural Networks

Jan Saro[1]

Abstract. Accurately predicting disease progression in dairy herds promotes effective health management and economic sustainability. In this context, Long Short-Term Memory (LSTM) networks outperform statistical models, such as Markov chains and naive prediction. However, advanced deep learning technologies requires major investments in infrastructure, which many smallholder dairy farms simply cannot afford. To assess the performance and suitability of these models for farms with varying technological and financial resources, we compared three predictive models for disease development, namely two statistical approaches and a machine learning model based on an LSTM neural network. For this purpose, we analyzed data from 36 diseases recorded in a dairy herd over a five-year period. The LSTM model showed the highest prediction accuracy, whereas the statistical models showed no difference in many scenarios. Nevertheless, the statistical models achieved satisfactory accuracy for diseases with low prevalence. As such, statistical models are suitable for less-equipped dairy farms seeking simpler and more cost-effective solutions. These findings support the adoption of tailored predictive approaches based on farm resources and disease profiles.

Keywords: prediction time series, neural network, Markov chains, naive prediction, machine learning, dairy farms, herd health

JEL Classification: C22, Q16
AMS Classification: 62M10

1 Introduction

Predicting disease progression enables dairy farmers to effectively manage herd health, fostering economic sustainability. By implementing preventive measures in a timely and accurate manner, dairy farmers can reduce the impact of diseases on both animal welfare and farm profitability. To this end, structured herd health management programs help farmers shift from treatment to preventive health management (Derks et al., 2013). This shift is crucial in modern dairy farming with increasing herd sizes and stringent quality standards, highlighting the importance of systematic approaches for herd health management (Hyndman and Athanasopoulos, 2018).

In recent years, both traditional statistical models and advanced machine learning techniques have been applied (University of Bristol, 2024) to disease prediction in dairy farming. Statistical models, such as Markov chains (Saro et al., 2024), have proved effective in modelling disease dynamics thanks to their simplicity and interpretability (Liu et al., 2011). In contrast, advanced deep learning approaches, particularly Long Short-Term Memory (LSTM) networks (Goodfellow et al., 2016), have shown superior accuracy in time-series forecasting in veterinary science (Zhang et al., 2022). However, advanced deep learning technologies require significant investments in infrastructure, and many smallholder dairy farms lack the funds needed to introduce such technologies to their daily operations. Concurrently, high data transmission costs and limited internet connectivity hinder the adoption of digital technologies in livestock farming (University of Bristol, 2024). Given these limitations, dairy farmers are often forced to use simpler and more cost-effective solutions.

Unlike our earlier studies, which focused primarily on the performance of Markov chain models (Saro et al., 2024), this paper offers a comparative perspective by incorporating an LSTM neural network and evaluating its applicability under varying farm conditions. Moreover, this study extends the scope by using a broader disease set and real farm data, aiming to assess practical implementation potential.

This study compares the performance of two statistical models with an LSTM-based deep learning model in predicting the progression of 36 diseases in a dairy herd over a five-year period. The main aim is to assess whether statistical models can serve as viable alternatives for less-equipped farms while also providing a bench-mark

[1] Czech University of Life Sciences Prague, Faculty of Economics and Management, Department of Systems Engineering, Kamycka 129,165 00, Praha - Suchdol, saroj@pef.czu.cz

comparison against state-of-the-art machine learning techniques. This research seeks to guide dairy farmers in selecting predictive approaches tailored to their specific needs and resources.

2 Data and Methods

2.1 Dataset

The data were continuously collected by the zootechnician of the dairy farm and subsequently processed using Python scripts. A dataset was collected for 750 dairy cows and 36 diseases in a single farm, in the Czech Republic, from January 1, 2018 to December 7, 2023. This dataset contained the total number of occurrences of each dairy cow disease monitored daily during the study period. These data were used for disease forecasting to help the zootechnician manage herd health.

2.2 Formula expression of dairy diseases

For each disease $i = 1, 2, \dots, D$ and day $t = 1,2, \dots, N$ and $d(i,t)$ is the number of dairy disease occurrences i on day t is expressed in the following formula:

$$d(i,t). \tag{1}$$

To calculate the total number of occurrences for disease i in each month m, add the total number of disease occurrences over all days t of that month:

$$S(i,m) = \sum_{t \in m} d(i,t) \tag{2}$$

, where:

- $S(i,m)$ is the total number of occurrences of disease i in month m
- $t \in m$ indicates that t belongs to month m.

2.3 Naive prediction

Using the naive prediction method (Hyndman and Athanasopoulos, 2018), the total number of occurrences of disease i forecast for the next month $m + 1$ is expressed as:

$$\hat{S}(i, m+1) = S(i,m) \tag{3}$$

, where:

- $\hat{S}(i, m+1)$: total number of occurrences of disease i predicted for the next month $m + 1$,
- $S(i,m)$: total number of occurrences of disease i in the current month m.

This approach leverages the assumption that disease occurrences in one month will persist into the next month without accounting for seasonal or external variations.

2.4 Neural Network (LSTM)

A Long Short-Term Memory (LSTM) neural network was applied to capture long-term relationships in time-series data. To predict the total occurrences of disease i in month $m + 1$, we define:

$$\hat{S}(i, m+1) = f(S(i,m), S(i, m-1), \dots, S(i, m-n+1)) \tag{4}$$

, where:

- $\hat{S}(i, m+1)$: total number of occurrences of disease i predicted for month $m + 1$,
- $f(°)$: Neural network function LSTM trained on historical monthly totals,
- $S(i, m-k)$: total number of occurrences of disease i in previous months $m - k$,
- n: The number of previous months used as input in the prediction.

The LSTM network consisted of one LSTM layer with 50 units followed by a dense output layer. The ReLU activation function was used. The model was trained using the Adam optimizer and MAE as the loss function, with 100 epochs and batch size of 16. Implementation was done in TensorFlow. Similar applications in veterinary science were described by (Zhang et al., 2022).

2.5 Homogenous Markov Chain

A Markov chain models the probabilities of states in a stochastic process, wherein the next state depends only on the current state and not on the preceding sequence of events (the "memoryless" property). The Markov-chain states are defined first, and then either one Transition matrix is calculated for the homogeneous model, or the Transition matrices are calculated for each step of the non-homogeneous model. After the predictions, the accuracy of the model is calculated by comparing the results with real data.
To clarify differences between the models, a Homogeneous Markov Chain (HMC) assumes that the transition probabilities remain constant, regardless of the quarter. This assumption simplifies the modeling process when disease occurrence patterns are relatively stable throughout the year. Conversely, in a Non-homogeneous Markov Chain (NHMC), transition probabilities vary with the quarter, capturing temporal or seasonal variations in disease dynamics. In the NHMC model, this flexibility is crucial for predicting the progression of diseases affected by seasonal factors, increasing the accuracy of the model. For this reason, the NHMC model is more suitable for data suggesting periodic changes in disease occurrence.

Step 1—Definition of the states of the Markov chain model

The states of the Markov chain model are defined as the number of dairy cows affected by disease per day. All states form the set $\{0,1,2,\dots,M\}$, where M is the number of dairy cows, 0 means that no dairy cow is affected by the disease in a chosen day, and M means that all dairy cows are affected by the disease in a chosen day. However, only a smaller number $K \ll M$ of cows is affected usually. Therefore, the real set of states of disease i is

$$\mathcal{S}(i) = \{s_0, s_1, \dots, s_{K-1}, s_K\} = \left\{0, 1, \dots, \max_{t=1,\dots,N-T}(d(i,t)) - 1, \max_{t=1,\dots,N-T}(d(i,t))\right\} \quad (5)$$

, where $d(i,t)$ is the number of occurrences of dairy cow disease i on day t.
If the probability of states referring to the highest number of disease occurrences per day is very low, the following subset $\mathcal{S}_R(i)$ of the set of states $\mathcal{S}(i)$ of the Markov chain model is used:

$$\mathcal{S}_R(i) \subseteq \mathcal{S}(i) \quad (6)$$

, where 0, 1, …, R are elements of $\mathcal{S}_R(i), R \leq K$, and state R aggregates all other $\{s_R, s_{R+1}, \dots, s_K\}$ states.

Step 2—Homogenous Markov Chain

Assuming the homogeneity of the process during the monitoring period, we first determine the transition matrices of all possible numbers of states R. For each disease i, the transition probabilities are calculated using the following formula:

$$\boldsymbol{P}_R(i) = \begin{pmatrix} p_R(i,1,1) & p_R(i,1,2) & \dots & p_R(i,1,R) \\ p_R(i,2,1) & \ddots & & \vdots \\ \vdots & & & \\ p_R(i,R,1) & \dots & & p_R(i,R,R) \end{pmatrix} \quad (7)$$

, where $p_R(i,a,b) = \frac{\sum_{t=1}^{N-T} c_{ab}(i,t)}{\sum_{b \in S_R(i)} \sum_{t=1}^{N-T} c_{ab}(i,t)} a, b = 1,2,\dots,R$, is the probability of the transition from state a sick dairy cows to state b sick dairy cows, N is the count of all days, T is the length of the predicted period, and $c_{ab}(i,t)$ is equal to either 1 if the transition from state a to state b occurs in time t or 0 otherwise.
At the end, the transition matrix is calculated for all reasonable R for which $\lfloor K/2 \rfloor \leq R \leq K$ because a smaller number of states would not describe the numbers of sick cows well enough.
State probabilities for each disease $i = 1, \dots, N$ are predicted as follows:

$$\hat{\boldsymbol{p}}_R^T(i) = \boldsymbol{p}_{init} \cdot \boldsymbol{P}_R^T(i) = \boldsymbol{p}_{init} \cdot \boldsymbol{P}_R(i) \cdot \boldsymbol{P}_R(i) \cdot \dots \cdot \boldsymbol{P}_R(i) \quad (8)$$

, where $\hat{\boldsymbol{p}}_R^T(i) = (\hat{p}_R(i,0), \hat{p}_R(i,1), \dots, \hat{p}_R(i,R))$ is the predicted distribution of states probabilities in predicted period T, $\boldsymbol{p_{init}}$ is vector of initial state probabilities with dimension R with all zeros, except the state describing the last count of occurrences of disease i, where its value is 1, $\boldsymbol{P}_R^T(i) = \boldsymbol{P}_R(i) \cdot \boldsymbol{P}_R(i) \cdot \dots \cdot \boldsymbol{P}_R(i)$ is the transition matrix from time $N - T + 1$ to N, i.e., T-th power of matrix $P_R(i)$, and T is the length of the predicted period.
The best value R^* is selected based on Chebyshev distance minimization to identify the best predictive accuracy:

$$R^* = \arg \min_{R=[K/2],\ldots,K} \left\{ \max_{j=1,\ldots,R} \left(\left| \hat{p}_r^T(i,j) - \frac{\sum_{t=N-T+1}^{N} c_j(i,t)}{T} \right| \right) \right\} \quad (9)$$

, where $\hat{p}_r^T(i,j)$ is a j-th element of the row vector $\hat{\boldsymbol{p}}_R^T(i)$, i.e., predicted probability of state j, and $c_j(i,t)$ is equal to either 1 if j dairy cows were sick with disease i in time t or 0 otherwise.
Based on predicted state probabilities the mean value of disease occurrences per day can be calculated as:

$$\hat{m}(i) = \hat{\boldsymbol{p}}_R^T(i) \cdot (0,1,2,\ldots,R^*)'. \quad (10)$$

The mean value of disease occurrences per day, $\hat{m}(i)$, is calculated as the scalar product of two vectors: the vector of predicted state probabilities $\hat{\boldsymbol{p}}_R^T(i)$ and the transposed vector of possible disease occurrence states $(0,1,2,\ldots,R^*)'$.

3 Results

3.1 Precision comparison of the models

Table 1 compares the performance of three predictive models, namely Markov chains, naive prediction, and neural networks, in forecasting monthly disease occurrences in dairy herds over two months. For diseases with zero occurrences, including abscess and acidosis, all models achieved perfect accuracy (MAE = 0.00). The neural network excelled in predicting sporadic diseases such as nerve damage and digestive troubles (MAE = 1.00) and outperformed other models for high-prevalence diseases such as Coli (MAE = 20.00 in month two). In contrast, Markov chains and naive prediction struggled with dynamic trends but performed similarly for consistent diseases such as retained bed.

The Mean Absolute Percentage Error (MAPE) further highlights differences among the models. In the first month, the neural network (LSTM), Markov chain (MCH), and naive prediction models achieved 29.60, 35.18 and 43.05% MAPE with 66.11, 66.36, and 66.61% standard deviation (SD), respectively. In the second month, the neural network also reached a higher accuracy (31.16% MAPE) than Markov chains (38.84%) and naive prediction (50.58%), with 53.41, 77.42, and 69.99% SD, respectively.

Disease	Real value per 1.month	MAE Markov chain 1.month	MAE Naive prediction 1.month	MAE Neural network 1.month	Real value per 2.month	MAE Markov chain 2.month	MAE Naive prediction 2.month	MAE Neural network 2.month
abscess	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
tympani	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
acidosis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prevention	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bleeding	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dermatitis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pneumonia	7.00	7.00	7.00	7.00	0.00	0.00	0.00	0.00
udder swelling	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
blood in milk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
postpartum sepsis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pain	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
nerve damage	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
jaw swelling	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mastitis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
digestive troubles	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00
damaged teat	0.00	0.00	0.00	0.00	4.00	4.00	4.00	4.00
torn during birth	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
spleen dislocation	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
high somatic	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00

peritonitis	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
eye injury	5.00	5.00	5.00	4.00	0.00	0.00	0.00	0.00
lying down after birth	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00
commercial vaccine	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
phlegmon	1.00	1.00	1.00	0.00	4.00	2.00	4.00	3.00
respiration	0.00	2.00	4.00	3.00	2.00	0.00	2.00	1.00
hoof	12.00	9.00	2.00	12.00	6.00	3.00	4.00	6.00
other	3.00	1.00	2.00	2.00	0.00	4.00	5.00	1.00
uterus	5.00	3.00	2.00	5.00	4.00	4.00	3.00	6.00
diarrhea	13.00	8.00	13.00	7.00	0.00	5.00	0.00	10.00
Coli	127.00	28.00	13.00	32.00	76.00	19.00	38.00	20.00
retained bed	8.00	8.00	6.00	2.00	8.00	8.00	6.00	2.00
endometritis	20.00	9.00	5.00	3.00	16.00	5.00	9.00	5.00
limb swelling	4.00	4.00	9.00	3.00	3.00	3.00	10.00	2.00
desiccation or drying up	60.00	26.00	14.00	13.00	37.00	4.00	9.00	10.00
temperature after calving	2.00	6.00	7.00	8.00	3.00	5.00	6.00	8.00
temperature	14.00	7.00	8.00	0.00	21.00	14.00	15.00	7.00
metabolic problem	22.00	9.00	7.00	4.00	36.00	24.00	7.00	17.00
necrobacillosis	24.00	11.00	13.00	19.00	28.00	15.00	17.00	14.00
mastitis PZ	13.00	25.00	4.00	2.00	30.00	7.00	13.00	18.00
mastitis PP	32.00	8.00	14.00	15.00	15.00	24.00	3.00	8.00
mastitis LZ	30.00	14.00	11.00	4.00	19.00	23.00	0.00	8.00
reproduction	358.00	6.00	61.00	67.00	310.00	31.00	13.00	2.00
mastitis LP	32.00	62.00	3.00	6.00	17.00	74.00	18.00	23.00

Table 1 Disease analysis over two predicted months by model. Columns include: Disease – the name of the disease; Real value per 1.month and 2.month – observed cases for each month; MAE Markov chain 1.mnoth and 2.mnoth – error of the Markov chain model; MAE Naive prediction 1.month and 2.month – error of the naive model; MAE Neural network 1.month and 2.month – error of the neural network model

4 Discussion

As shown by our comparative analysis, neural networks outperform statistical predictive models such as Markov chains and naive prediction in forecasting disease occurrences in dairy herds. Standing out as the most accurate model, neural networks are particularly more accurate for diseases with high variability or prevalence, such as E. coli- and reproduction-related conditions. These results corroborate the findings of other authors (Shine and Murphy, 2022), who have highlighted the extensive and transformative role of machine learning in dairy farming over the past two decades.

In turn, Markov chains and naive predictions performed similarly for stable or non-occurring diseases, such as mastitis and abscess, demonstrating that traditional statistical methods remain relevant for predictable disease trends. Neural networks may excel in dynamic environments for their ability to model complex, nonlinear patterns, though LSTM networks typically require more computational resources than statistical models, their application is feasible even on modest datasets with standard computing hardware. These demands may limit their adoption in resource-constrained settings. Therefore, our findings not only confirm that machine learning approaches surpass other models in complex scenarios but also demonstrate the feasibility and suitability of traditional methods in specific contexts, more specifically, in smallholder farms with limited resources and where the most prevalent diseases display predictable trends.

The findings of this study highlight the critical role of tailored predictive modelling in dairy herd health management, particularly in the context of varying technological and financial capabilities among farms. The superior

performance of the Long Short-Term Memory (LSTM) model in predicting disease progression underscores the potential of advanced machine learning techniques to enhance decision-making in agricultural health management (Zhang et al., 2022). However, the necessity for substantial infrastructure investment associated with LSTM models poses a significant barrier for smallholder dairy farms, which may lack the resources to implement such sophisticated systems (Vukasinovic et al., 2022). In contrast, traditional statistical models demonstrated satisfactory accuracy for diseases with low prevalence, suggesting that simpler approaches can still provide valuable insights for less-equipped farms (Li et al., 2020). Future research should focus on developing hybrid models that combine the strengths of both machine learning and statistical methods, ensuring that predictive analytics can be effectively utilized across diverse farming contexts (Reps et al., 2022).

5 Conclusion

Neural networks provide the most accurate predictions for complex and variable diseases, such as Coli and reproduction, whereas simpler models, such as Markov chains and naive prediction, perform similarly for stable or non-occurring diseases such as abscess and mastitis. Despite their accuracy, neural networks demand more resources, a requirement that may limit their feasibility in some farms. A tailored approach, matching predictive methods to disease profiles and farm resources, may help dairy farmers optimize herd health management as a function of the technological level of the farm.

References

Derks M, van Werven T, Hogeveen H, Kremer WDJ (2013) Veterinary herd health management programs on dairy farms in the Netherlands: Use, execution, and relations to farmer characteristics. Journal of Dairy Science 96:1623–1637.

Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press, Cambridge.

Hyndman RJ, Athanasopoulos G (2018) Forecasting: Principles and Practice (2nd ed.). OTexts, Melbourne.

Li Y, Sperrin M, Ashcroft D, Staa T (2020) Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. BMJ. https://doi.org/10.1136/bmj.m3919

Liu Q, Chen H, Zhou X, Shen Y (2011) A Markov chain model for predicting disease dynamics in livestock populations. PLoS ONE 6:e24833.

Reps J, Williams R, Schuemie M, Ryan P (2022) Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. BMC Medical Informatics and Decision Making 22. https://doi.org/10.1186/s12911-022-01879-6

Saro J, Ducháček J, Brožová H, Stádník L, Bláhová P, Horáková T, Hlavatý R (2024) Discrete homogeneous and non-homogeneous Markov chains enhance predictive modelling for dairy cow diseases. Animals 14:2542.

Shine P, Murphy MD (2022) Over 20 years of machine learning applications on dairy farms: A comprehensive mapping study. Sensors 22:52.

University of Bristol (2024) Challenges in implementing advanced technologies in livestock management: Insights from less-equipped farms. In: Population Health Sciences, Centre for Research. University of Bristol, Bristol.

Vukasinovic N, Gonzalez D, Przybyla C, Brooker J, Kulkarni A, Passafaro LT, McNeel AK (2022) Genetic control of wellness in dairy cattle. Animal Husbandry. https://doi.org/10.5772/intechopen.103819

Zhang Z, Chen Q, Xu P, Wang H (2022) Long short-term memory networks for time-series forecasting in veterinary science. ACM Transactions on Management Information Systems 13:1–23.

Received: 2025-05-26 Last revision: 2025-05-26 Accepted: 2025-07-29

Deriving Production Functions in Economics Through Data-Driven Dynamical Systems

Roman G. Smirnov

Abstract. In their seminal 1928 work, Charles Cobb and Paul Douglas empirically validated the Cobb-Douglas production function through statistical analysis of U.S. economic data from 1899 to 1923. While this established the function's theoretical foundation for growth models like Solow-Swan and its extensions, it simultaneously revealed a fundamental limitation: their methodology could not determine whether alternative production functions might equally explain the observed data.
This paper presents a novel dynamical systems approach to production function estimation. By modeling economic growth trajectories as dynamical systems, we derive production functions as time-independent invariants – a method that systematically generates all possible functional forms compatible with observed data.
Applying this framework to Cobb and Douglas's original dataset yields two key results: First, we demonstrate that the Cobb-Douglas form emerges naturally from exponential growth dynamics in labor, capital, and production. Second, we show how combining fundamental invariants of this exponential system generates the CES production function as a special case. Our methodology bridges statistical analysis with mathematical systems theory, providing both a verification mechanism for classical results and a tool for discovering new functional relationships.

Keywords: data-driven dynamical systems, production functions in economics, mathematical modeling, economic growth models

JEL Classification: E17, F43, F64, F68, O21, O41, O44, O47
AMS Classification: 37N40, 91B50, 91B55, 91B62

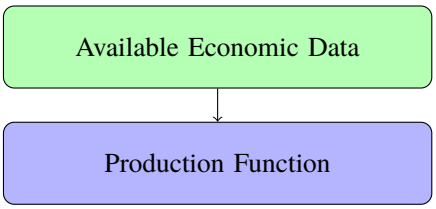
1 Introduction

Production functions constitute the foundational building blocks of economic modeling, providing the formal mathematical relationship between factor inputs (capital, labor, technology, land, and other productive resources) and economic output (firm-level production, aggregate GDP, and related measures). These functions are typically derived through three complementary approaches: mathematical tractability, empirical calibration, or theoretical consistency. The canonical Cobb-Douglas production function, whose early formalization in economic analysis is attributed to Knut Wicksell, Philip Wicksteed, and Léon Walras (see Humphrey, 1997 for more details and references), gained prominence due to its analytical convenience and theoretically appealing properties – most notably its constant returns to scale characteristic, which is an important property in economic contexts.

In their seminal 1928 paper (Cobb and Douglas, 1928; see also Douglas, 1976), Charles Cobb and Paul Douglas used this function that would subsequently become known as the Cobb-Douglas production function by empirically fitting it to data on U.S. economic growth from 1899 to 1922. Their work represented a pivotal advancement in growth theory, as it was the first to demonstrate that a specific form of the Cobb-Douglas production function could closely match observed patterns in production, capital accumulation, and labor inputs. This empirical validation established the Cobb-Douglas function as a foundational tool for subsequent theoretical and applied work in economics. Specifically, Cobb and Douglas plotted the available data on logarithmic scale, guessed what function could fit well the data, and then employed statistical tools to verify the accuracy of their assumptions about the form of the production function in question.

This approach, which ultimately legitimized the Cobb-Douglas production function, was considered revolutionary at the time. Unlike in physics – where fundamental concepts, differential equations, and conserved quantities are typically derived from first principles (e.g., Newton's second law) – economic models are often built on heuristic observations, simplifying assumptions, and mathematical convenience. Thus, the methodology employed by Cobb and Douglas in 1928, which relied on rigorous statistical analysis of empirical data, marked a significant breakthrough. It can be summarized as follows:

Dalhousie University, Department of Mathematics and Statistics, Halifax, 6297 Castine Way PO, BOX 15000 Canada B3H 4R2, Roman.Smirnov@dal.ca

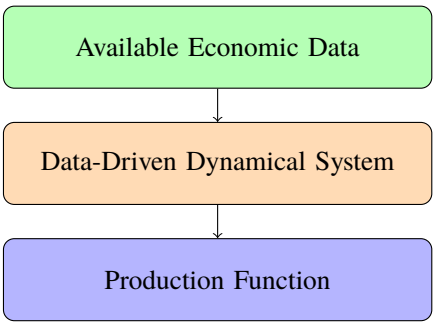


Despite its significance and its profound influence on the development of economic growth theory over the past century, this approach to deriving production functions leaves many critical questions unanswered. Most notably, even when a production function fits well to a given dataset, key uncertainties remain:

- Are there alternative production functions that are equally compatible with the data?
- What is the complete set of production functions that could describe the observed relationships?

These unresolved issues challenge the uniqueness and robustness of such models.

To address these questions, we have developed a novel approach for deriving production functions using dynamical systems theory. In a series of papers (Smirnov and Wang, 2020; Smirnov and Wang, 2021; Smirnov et al, 2022; Smirnov and Wang, 2024), we have introduced a new framework where production functions emerge as *time-independent invariants* of data-compatible dynamical systems. This method enables the derivation of new production functions that are not empirically derived from economic data *directly*, but rather emerge from dynamical systems determined either by data or mathematical assumptions (Smirnov and Wang, 2020; Smirnov and Wang, 2024). The process can be summarized in the following flow diagram:



In this paper, we review the method and apply it to derive a new production function using the dataset originally analyzed by Charles Cobb and Paul Douglas (Cobb and Douglas, 1928). This generalized production function has the well-known CES (Constant Elasticity of Substitution) production function (see Acemoglu, 2009 and the relevant references therein) as a special case. Furthermore, we establish necessary conditions for a dynamical system driven by exponential growth in labor, capital, and production – to admit the CES production function as a time-independent invariant.

This paper is organized as follows. Section 2 presents our novel dynamical systems approach to deriving production functions from economic data, using the exponential growth model that approximates the original Cobb-Douglas dataset Cobb and Douglas, 1928 as a key example. In Section 3, we establish conditions under which the Constant Elasticity of Substitution (CES) production function emerges from this exponential model. Concluding remarks and implications for economic modeling are discussed in Section 4.

2 From data to production functions via dynamical systems

Recall that in their seminal 1928 paper (Cobb and Douglas, 1928), Charles Cobb and Paul Douglas demonstrated that a production function of the form

$$Y = AL^{\alpha}K^{\beta}, \tag{1}$$

where $Y = f(L,K)$ represents total production output, L denotes labor input, K – capital input, A is total factor productivity (a measure of technological efficiency), and α and β are the output elasticities of labor and capital, respectively, provided an accurate empirical fit to US economic growth data from 1899 to 1922. Specifically, employing the methods of least squares estimations, the authors assumed constant returns to scale in (1), that is $\alpha+\beta = 1$ and obtained a good fit to the data by choosing $\alpha = 0.75$, which was later verified by the National Bureau

of Economic Research to be 0.741, and determined $A = 1.01$. The value of α was determined by plotting the time series of production (Y), labor (L), and capital (K) on a logarithmic scale and analyzing the relative distances between the curves (see Douglas, 1976 for details). Recent work by Smirnov and Wang, 2024 (see also Smirnov and Wang, 2021 and Smirnov et al, 2022) has demonstrated that more precise estimates can be obtained through data-driven dynamical systems approaches, which provide a more sophisticated framework for fitting the model to historical data. Specifically, rather than directly fitting the production function in (1) to the aforementioned dataset, the authors employed R programming to estimate parameters for the dynamical system:

$$\dot{L}(t) = b_1 L(t), \quad \dot{K}(t) = b_2 K(t), \quad \dot{Y}(t) = b_3 Y(t), \tag{2}$$

which models the exponential growth dynamics of capital, labor, and production. Our analysis, implemented in R, established that the solutions

$$L(t) = L_0 e^{b_1 t}, \quad K(t) = K_0 e^{b_2 t}, \quad Y(t) = Y_0 e^{b_3 t} \tag{3}$$

were consistent with the empirical results from Cobb and Douglas, 1928 (cf. Smirnov and Wang 2021) for the parameter values

$$\begin{aligned} b_1 &= 0.02549605, \quad \ln L_0 = 4.66953290 \quad \text{(labor)}, \\ b_2 &= 0.06472564, \quad \ln K_0 = 4.61213588 \quad \text{(capital)}, \\ b_3 &= 0.03592651, \quad \ln Y_0 = 4.66415363 \quad \text{(production)}. \end{aligned} \tag{4}$$

Recall, by combining the solutions given by (3) in a way that eliminates time t, we produced a one-parameter family of time-independent invariants of the Cobb-Douglas form under a specific linear condition satisfied by the parameters b_1, b_2, and b_3:

$$Y = AL^{\alpha} K^{\frac{b_3}{b_2} - \alpha \frac{b_1}{b_2}}, \quad 0 < \alpha < 1, \tag{5}$$

where α is a parameter and A is a constant.

We further showed that the constant-returns-to-scale Cobb-Douglas function ($\alpha + \beta = 1$) is a special case of (5) $b_1 < b_3 < b_2$. Notably, the parameter values b_1, b_2, and b_3 from (4), extracted from the dataset in Cobb and Douglas, 1928, satisfy this inequality.

Using the bi-Hamiltonian approach introduced in Smirnov and Wang, 2021, we obtained explicit formulas for the output elasticities of this particular family member:

$$\alpha = \frac{b_3 - b_2}{b_1 - b_2}, \quad \beta = \frac{b_3 - b_1}{b_2 - b_1}. \tag{6}$$

Using this formula (6) for the parameters b_1, b_2, and b_3 given by (4), we computed the corresponding α and β:

$$\alpha = 0.7341175376, \quad \beta = 0.2658824627. \tag{7}$$

Unsurprisingly, these values closely match the original estimates obtained in Cobb and Douglas, 1928 (see Smirnov and Wang, 2024 for details).

These findings demonstrate two key insights:

1. The Cobb-Douglas function (1) arises naturally from exponential growth in K, L, and Y which may explain its historical empirical fit compared to modern data (cf. Douglas, 1976).
2. Deriving production functions directly from data-driven dynamical systems provides a more comprehensive framework.

In the following section, we extend this approach to CES-like production functions, identifying parameter conditions under which a genuine CES function emerges from the dynamical system (2).

3 The CES production function revisited

Now, let us employ the techniques of the previous sections to link the Constatn Elasticity of Substitution (CES) production function to the same dynamical system (2) shown to be compatible with the data studied in Cobb and Douglas, 1928. Recall that the CES production function is given by

$$Y = A[\alpha K^p + (1 - \alpha) L^p]^{\frac{\nu}{p}}, \tag{8}$$

where, as before, Y represents production (output); A - the total factor productivity; $\alpha, 0 < \alpha < 1$ is the share parameter; K and L are the outputs (capital and labor respectively); $p = \frac{\sigma-1}{\sigma}$ is the substitution parameter; $\sigma = \frac{1}{1-p}$ is the elasticity of substitution; v is the degree of homogeneity of the production function (8): if $v = 1$ the function enjoys constant returns to scale, $v < 1$ it has decreasing returns to scale, and $v > 1$ indicates it has increasing returns to scale. For a comprehensive analysis of the elasticity of substitution σ between labor and capital and its relationship to contemporary economic trends, see Piketty, 2014.

The production function (8) was originally derived by Kenneth Arrow, Hollis Chenery, Bagicha Minhas, and Robert Solow in 1961 (see Acemoglu, 2009 and the relevant references therein) by fitting a specific linear combination of $\ln K$, $\ln L$, and $\ln Y$ to economic data. In what follows, we will rederive this function from the dynamical system (2) under some additional conditions.

First, we note that from the mathematical viewpoint the formulas (3) may be interpreted as the action of a one-parameter Lie group in the three-dimensional space represented by the economic variables K, L, and Y. Therefore, we should be able to obtain 3 (the dimension of the space) - 1 (the dimension of the Lie group acting in the space) = 2 fundamental invariants (see Smirnov and Wang, 2024 and the relevant references therein for more details). The Lie group action is so simple that we can easily derive two fundamental invariants via simple algebra. Indeed, combine first $L(t) = L_0 e^{b_1 t}$ and $Y(t) = Y_0 e^{b_3 t}$. Using the first formula, we first solve for t:

$$t = \frac{1}{b_1} \ln \frac{L}{L_0}.$$

Substituting this expression into $Y(t) = Y_0 e^{b_3 t}$, we arrive at the first fundamental time-independent invariant depending on Y and K given by

$$Y = \frac{Y_0}{L_0^{\frac{b_3}{b_1}}} L^{\frac{b_3}{b_1}} \quad (YL^{-\frac{b_3}{b_1}} = \text{const}). \tag{9}$$

Combining now the formulas $K(t) = K_0 e^{b_2 t}$ and $Y(t) = Y_0 e^{b_3 t}$ and repeating the same procedure by eliminating the parameter t, we arrive at the second fundamental time-invariant depending on Y and L given by

$$Y = \frac{Y_0}{K_0^{\frac{b_3}{b_2}}} K^{\frac{b_3}{b_2}} \quad (YK^{-\frac{b_3}{b_2}} = \text{const}). \tag{10}$$

The time-independent functions (9) and (10) are functionally independent and as fundamental invariants have the property that any other time-independent invariant of the Lie group action (3) is a function of (9) and (10). For example, the time-independent invariant (5) is a function of those. Indeed, it follows from (9) that

$$Y = BL^{\frac{b_3}{b_1}},$$

where B is a constant. Hence, $Y^\alpha = B^\alpha L^{\alpha \frac{b_3}{b_1}}$, $0 < \alpha < 1$. On the other hand, it follows from (10) that

$$Y = CK^{\frac{b_3}{b_2}},$$

where C is a constant and so $Y^\alpha = C^\alpha K^{\alpha \frac{b_3}{b_2}}$, $0 < \alpha < 1$. Therefore, we have

$$Y = CK^{\frac{b_3}{b_2}} = CK^{\frac{b_3}{b_2}}(1)^{\frac{b_1}{b_3}} = CK^{\frac{b_3}{b_2}} \left(\frac{Y^\alpha}{Y^\alpha}\right)^{\frac{b_1}{b_3}} = CK^{\frac{b_3}{b_2}} \left(\frac{B^\alpha L^{\alpha \frac{b_3}{b_1}}}{C^\alpha K^{\alpha \frac{b_3}{b_2}}}\right)^{\frac{b_1}{b_3}} = AL^\alpha K^{\frac{b_3}{b_2} - \alpha \frac{b_1}{b_2}}, \quad 0 < \alpha < 1,$$

as expected.

To derive the function (8) from the dynamical system (3) determined by the exponential growth in capital (K), labor (L), and production (Y), we will use the fundamental time-independent invariants (9) and (10) again, but this time we will combine them using addition and subtraction, rather than multiplication and division as in the case of derivation of the family of Cobb-Douglas functions (5) presented above. Indeed, let us consider the following algebraic expression:

$$Y^{\frac{1}{b_3}} = \alpha Y^{\frac{1}{b_3}} + Y^{\frac{1}{b_3}} - \alpha Y^{\frac{1}{b_3}}, \tag{11}$$

where $0 < \alpha < 1$ is some parameter. Next, using the time-independent invariants (9) and (10) and substituting these expressions instead of the terms of the RHS of (11), we obtain

$$Y^{\frac{1}{b_3}} = \alpha \frac{Y_0^{\frac{1}{b_3}}}{K_0^{\frac{1}{b_2}}} K^{\frac{1}{b_2}} + \frac{Y_0^{\frac{1}{b_3}}}{L_0^{\frac{1}{b_1}}} L^{\frac{1}{b_1}} - \alpha \frac{Y_0^{\frac{1}{b_3}}}{L_0^{\frac{1}{b_1}}} L^{\frac{1}{b_1}} = \alpha \frac{Y_0^{\frac{1}{b_3}}}{K_0^{\frac{1}{b_2}}} K^{\frac{1}{b_2}} + \frac{Y_0^{\frac{1}{b_3}}}{L_0^{\frac{1}{b_1}}} (1-\alpha) L^{\frac{1}{b_1}},$$

from which it follows

$$Y = \left[\alpha \frac{Y_0^{\frac{1}{b_3}}}{K_0^{\frac{1}{b_2}}} K^{\frac{1}{b_2}} + \frac{Y_0^{\frac{1}{b_3}}}{L_0^{\frac{1}{b_1}}} (1-\alpha) L^{\frac{1}{b_1}} \right]^{b_3}, \; 0 < \alpha < 1. \tag{12}$$

Therefore, we conclude that we have proven the following

Proposition 1. *The exponential model (2) admits the time-independent family of invariants (12) that are determined by the fundamental invariants (9) and (10).*

We note immediately that the family of functions (12) looks a lot like the CES production function (8) and under some additional assumptions reduces to it. Indeed, suppose $b_1 = b_2$ in (2), that is labor ($L = L(t)$) and capital ($K = K(t)$) enjoy an identical exponential growth. Also we note that without loss of generality we can assume that $L_0 = K_0 = Y_0$. Thus, we can always normalize the data so that the initial condition is equal to 100% – as it was done in Cobb and Douglas, 1928, for example. See also the values of $\ln L_0$ $\ln K_0$, and $\ln Y_0$ given by (4). Therefore, under these assumptions, the formula (12) reduces to (8) after factoring out the common factor in the RHS of (8) and identifying $\frac{1}{b_1} = \frac{1}{b_2} = p$ and $b_3 = \frac{\nu}{p}$. These observations put in evidence that we have the following

Corollary 1. *If $b_1 = b_2$ in the exponential model (2), it admits the time-independent family of invariants in the form of the CES production function given by (8).*

In other words, the CES production function (8), much like the Cobb-Douglas function, can also be defined as a consequence of the exponential growth in K, L, and L given by (2) under the additional condition that $b_1 = b_2$.

4 Conclusions

The Cobb-Douglas production function (1) was empirically validated by Cobb and Douglas (1928) using U.S. economic data from 1899-1922. However, their approach left open questions about the uniqueness and exhaustiveness of this functional form. In this paper we have briefly reviewed the fact that the Cobb-Douglas function emerges naturally as a time-independent invariant of a dynamical system modeling exponential growth in labor (L), capital (K), and the production (Y). We have shown that the system's parameters (b_1, b_2, b_3) derived from the dataset originally studied by Cobb and Douglas directly determine the elasticities α and β in (1) via the formula (6), yielding the values (7) that closely match the historical estimates obtained in Cobb and Douglas, 1928. The paper discusses a data-driven dynamical systems approach introduced recently to derive production functions, moving beyond ad hoc fitting. Furthermore, we have shown that the exponential model (2) compatible with the dataset studied in Cobb and Douglas, 1928 admits the one-parameter family of Cobb-Douglas production functions (5). Moreover, we have rederived this family of functions as a combination of the two fundamental invariants (9) and (10) associated with the group action (3). This method generalizes to other production functions (e.g., CES (8)) and provides a rigorous mathematical foundation for empirical calibration. Specifically, the Constant Elasticity of Substitution (CES) function (8) is derived in this paper as another invariant of the same dynamical system, but under the condition $b_1 = b_2$ by introducing a recombination of the fundamental invariants (9) and (10). This result links CES parameters (ν, p) to the dynamical system's growth rates (Proposition 1 and Corollary 1), offering new insights into its theoretical validity and empirical applicability.

Acknowledgements

The author thanks the organizers of the 43rd International Conference on Mathematical Methods in Economics (MME 2025) for the invitation to attend the conference and give a lecture.

References

Acemoglu D (2009) Introduction to Modern Economic Growth. Princeton University Press, Princeton

Cobb C, Douglas P (1928) A theory of production. American Economic Review 18 (Supplement):139–165

Douglas P (1976) The Cobb-Douglas production function once again: Its history, its testing, and some new empirical values. Journal of Political Economy 84(5):903–916. https://doi.org/10.1086/260489

Humphrey T (1997) Algebraic production functions and their uses before Cobb-Douglas. Federal Reserve Bank of Richmond Economic Quarterly 83:51–83

Piketty T (2014) Capital in the Twenty-First Century. Harvard University Press, Harvard

Smirnov R, Wang K (2020) In search of a new economic model determined by logistic growth. European Journal of Applied Mathematics 31:339–368. https://doi.org/10.1017/S0956792519000081

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. https://doi.org/10.3842/SIGMA.2024.045

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 Terzioğlu M.K.,. Springer International Publishing, Cham, pp 1–12

Received: 2025-05-29 Last revision: 2025-08-14 Accepted: 2025-08-19

A note on characterization of comonotonicity

Anton Hovana[1], Peter Trebuňa[2]

Abstract. The contribution is devoted to providing a full characterization of functions φ and ψ, where $\varphi, \psi\colon [0, \bar{y}] \to [0, \bar{y}]$, satisfying the inequality $\varphi(a \vartriangle b) \leq \varphi(a) + \varphi(b) - \psi(a \circ b)$ for any a, b, under fixed couple of operations $(\vartriangle, \circ)$, where $\vartriangle, \circ\colon [0, \bar{y}]^2 \to [0, \bar{y}]$, while providing another view on the class of comonotone functions. As it has been shown, convexity of the considered functions plays a key role.

Keywords: comonotone functions, convexity, duality, monotone measure

JEL Classification: C00
AMS Classification: 28E10

1 Introduction

The theory of non-additive measures and integrals emerged as a response to the growing need for mathematical tools capable of handling quantitative information in modern disciplines. Classical measure and integral theory is based on the principle of additivity, which mirrors fundamental laws in physics and geometry, where the total measure of a system is the sum of its parts. However, this additive framework often proves insufficient in contexts involving human behavior, decision-making, or mutual interactions. Non-additive measures provide a suitable alternative in such cases. Furthermore, in many applications, it is crucial to estimate one integral in terms of another. In this regard, integral inequalities become powerful and practical tools with applications across various fields of mathematics, including probability theory, optimization, and approximation theory.

A concept of comonotonicity of functions has an interesting history. Hardy, Littlewood and Pólya (see Hardy et al, 1934), included in their book the Chebyshev integral inequality for the Lebesgue integral with using comonotone functions, which they refer to as *similarly ordered functions*. The best result in that way is as follows, see Armstrong, 1993/94:

If $(X, \mathcal{A})$ is a measurable space, then two $\mathcal{A}$-measurable real functions f and g defined on X satisfy the inequality

$$\int f g \, \mathrm{d}\mathsf{P} \geq \int f \, \mathrm{d}\mathsf{P} \int g \, \mathrm{d}\mathsf{P}$$

for any probability measure P if and only if functions f, g are comonotone.

It is worth mentioning that in the paper of Dellacherie, 1971 there was the first time when the term "*comonotonicity*" appeared. In that paper, Dellacherie proved comonotone additivity of the Choquet integral, which was later improved by Schmeidler, 1986 providing a characterization of comonotone additive functionals representable by the Choquet integral (in fact, proving a necessary condition). Since that time the concept of comonotonicity has found its place in decision theory Yaari 1987, actuarian sciences Dhaene et al, 2002a, Dhaene et al, 2002b, Dybvig, 1988, and many others. It is well-known that the classical integral inequalities (including the above mentioned the Chebyshev one) need not hold in general when considering non-additive measure and integral setting. Thus, we have to require additional assumptions. In the last decade, a huge amount of papers appeared in the connection of inequalities and non-additive integrals where authors provided conditions under which this or that well–known inequality is true. The newest papers regarding to that problem (see Boczek et al, 2019a, Boczek et al, 2019b,Boczek et al, 2020) provide the necessary and sufficient conditions for validity the Chebyshev, Hölder and Minkowski inequalities in general versions. To do so, the authors had to consider different classes of functions which generalize the class of comonotone functions. Later on, in the paper Boczek et al 2025, there have been presented some connections among considered classes of functions, namely comonotone functions versus $\star_{\inf}$-associated functions and m-positively dependent functions. Moreover, the authors focused on the dual versions to these classes. Our aim is to extend one inequality stated there and provide a full characterization of the class of functions which fulfils this inequality.

[1] Technical University of Košice, Faculty of Mechanical Engineering, Department of Applied Mathematics and Informatics, Letná 1/9, 042 00 Košice-Sever, anton.hovana@tuke.sk
[2] Technical University of Košice, Faculty of Mechanical Engineering, Department of Industrial and Digital Engineering, Park Komenského 9, 042 00 Košice-Sever, peter.trebuna@tuke.sk

The structure of the paper is as follows. In the next section, we introduce basic notations used in the whole paper. The next section is devoted to our motivation, where we mention elementary definitions to be the paper consistent and provide our results. In Conclusion, we sum up our idea.

2 Preliminaries

Let (X, Σ) be a measurable space, where Σ is a σ-algebra of subsets of a nonempty set X while Σ^0 denotes $\Sigma \setminus \{\emptyset\}$. For the given measurable space (X, Σ) we denote by $\mathcal{F}$ the set of all Σ-measurable functions $f\colon X \to [0, \infty]$, and by $\mathcal{F}^{\bar{y}} = \{f \in \mathcal{F}\colon \sup_{x\in X} f(x) \leq \bar{y}\}$ the bounded functions with an upper bound $\bar{y}$, where $0 < \bar{y} < \infty$. Recall that function $f\colon X \to [0, \bar{y}]$ is *continuous* if $\lim\limits_{n\to\infty} x_n = x$ for all $x_n, x \in X$ such that $x_n \to x$.

For a fixed measurable space (X, Σ) we also consider the set $\mathcal{M}$ of all *monotone measures*, i.e., set functions $m\colon \Sigma \to [0, \infty]$ satisfying $m(A) \leq m(B)$ whenever $A \subseteq B$ and $A, B \in \Sigma$ with the boundary condition $m(\emptyset) = 0$ and $m(X) > 0$. If $m(X) = 1$, then m is called a *capacity* and $\mathcal{M}^1$ denotes the set of all capacities. Let $m(\Sigma \cap D) = \{m(A \cap D)\colon A \in \Sigma\}$ for a fixed $D \in \Sigma$ and $m \in \mathcal{M}$. To shorten the notation, in the case $D = X$, we denote by $m(\Sigma)$ the range of m.

Recall that the monotone measure is called

1. *subadditive* whenever for all $A, B \in \Sigma; m(A \cup B) \leq m(A) + m(B)$;
2. *superadditive* whenever for all $A, B \in \Sigma, A \cap B = \emptyset; m(A \cup B) \geq m(A) + m(B)$;
3. *submodular* whenever for all for all $A, B \in \Sigma; m(A \cup B) + m(A \cap B) \leq m(A) + m(B)$;
4. *supermodular* whenever for all for all $A, B \in \Sigma; m(A \cup B) + m(A \cap B) \geq m(A) + m(B)$;

A binary operation $\circ\colon [0, \bar{y}]^2 \to [0, \bar{y}]$ is called a *fusion function*. Pre-aggregation functions and aggregation functions are the most important examples of fusion functions (see Bustince et al, 2017). We say that a function $\circ\colon Y_1 \times Y_2 \to [0, \infty]$ is *non-decreasing* if $a_1 \circ a_2 \leq b_1 \circ b_2$ whenever $a_i \leq b_i$, where $a_i, b_i \in Y_i \subset [0, \infty]$ for $i = 1, 2$. A binary operation $\circ$ is *increasing* if $a < b$ implies $a \circ c < b \circ c$ and $c \circ a < c \circ b$ for each c. Binary operations defined on $[0, \infty]^2$ are: $\mathrm{M}(a, b) = a \wedge b$ and $\Pi(a, b) = ab$, where $a \wedge b = \min(a, b)$. A special type of such operations is *semicopula* (also called a *t-seminorm*) $\mathrm{S}\colon [0, 1]^2 \to [0, 1]$ is a non-decreasing function such that $\mathrm{S}(a, 1) = \mathrm{S}(1, a) = a$ (see Borzová-Molnárová et al, 2015, Ouyang et al, 2019). From it follows that $\mathrm{S}(x, 0) = 0 = \mathrm{S}(0, x)$ for all x and $\mathrm{S}(x, y) \leq x \wedge y$ for all x, y. Typical examples of semicopulas are the functions $\mathrm{M}|_{[0,1]^2}$, $\Pi|_{[0,1]^2}$ and $\mathrm{W}(a, b) = (a + b - 1) \vee 0$, where $a \vee b = \max(a, b)$ and $h|_A$ denote the restriction of h to a set A. Note that the function W is called the *Łukasiewicz semicopula*. The set of all semicopulas S will be denoted by $\mathfrak{S}$.

3 Results

As there is mentioned in Introduction, in the paper Boczek et al, 2025, the authors presented some connections among the class of comonotone functions and m-positively dependent functions.

We start with the basic definitions of the above mentioned classes of functions.

Definition 1. Functions $f, g\colon X \to [0, \infty)$ are called *comonotone* on a nonempty set $D \subseteq X$ if

$$(f(x) - f(y))(g(x) - g(y)) \geq 0$$

holds for all $x, y \in D$. In the case $D = X$, we simply say that f, g are *comonotone*.

The notion of positive dependence among random events was initially formalized by Lehmann in Lehmann 1966 within the framework of probability measures. Recall that two random events A, B are called positively dependent (with respect to the probability measure P) if $\mathsf{P}(A \cap B) \geq \mathsf{P}(A) \cdot \mathsf{P}(B)$. Following Lehmann, the concept of positive dependence of functions with respect to the operator $\vartriangle$ in the setting of monotone measures was introduced in Kaluszka et al 2014 as follows.

Definition 2. Let $f, g \in \mathcal{F}^{\bar{y}}$, $m \in \mathcal{M}$, and $A, B \in \Sigma^0$. Functions $f|_A$ and $g|_B$ are called *m-positively dependent with respect to an operator* $\vartriangle\colon m(\Sigma) \times m(\Sigma) \to m(\Sigma)$ (*m-$\vartriangle$-positively dependent*, for short) if

$$m\big(A \cap B \cap \{f \geq t\} \cap \{g \geq s\}\big) \geq m(A \cap \{f \geq t\}) \vartriangle m(B \cap \{g \geq s\}) \tag{1}$$

holds for all $s, t \in [0, \bar{y}]$. For $A = B = X$ we simply say that functions f, g are m-$\vartriangle$-positively dependent.

According to the definition of m-$\triangle$-positively dependent functions, there are two ways to establish relationships between the class of comonotone functions and the class of m-$\triangle$-positively dependent functions: either by providing statements under a fixed monotone measure m that enforces an additional condition on the operator $\triangle$, or vice versa. Unfortunately, both approaches yield only the sufficient condition for deriving m-$\triangle$-positive dependency from comonotonicity. We will follow idea mentioned in the paper Boczek et al, 2025.

Now, let $m \in \mathcal{M}$ be supermodular. In that paper, authors additionally assume

$$a \triangle b \leq a + b - a \circ b \quad \text{for any} \quad a, b \in m(\Sigma) \tag{2}$$

with $\circ\colon m(\Sigma) \times m(\Sigma) \to m(\Sigma)$ being a binary operation, and the condition (U) holds, where
(U) for each $C, D \in \Sigma$ we have $m(C \cup D) \leq m(C) \circ m(D)$.

Then immediately

$$\begin{aligned} m(C \cap D) &\geq m(C) + m(D) - m(C \cup D) \geq m(C) + m(D) - m(C) \circ m(D) \\ &\geq m(C) \triangle m(D). \end{aligned}$$

Putting $C = A \cap \{f \geq t\}$ and $D = B \cap \{g \geq s\}$ for any $s, t \in [0, \bar{y}]$, the following statement can be obtained.

Proposition 1. *Let $m \in \mathcal{M}$ be supermodular, $\circ\colon m(\Sigma) \times m(\Sigma) \to m(\Sigma)$ and conditions* (2) *and* (U) *hold. Then any functions $f|_A, g|_B \in \mathcal{F}^{\bar{y}}$ are m-$\triangle$-positively dependent.*

Example 1. Below we present several pairs of operations satisfying inequality (2):

$$(\triangle, \circ) = \{(\vee, \wedge), (\wedge, \vee), (\mathrm{T}, \mathrm{T}^*), (\mathrm{C}, \mathrm{C}^*), (G, G), (K_{\frac{1}{2}}, K_{\frac{1}{2}}), (K_1, K_0), (K_0, K_1)\},$$

where T is the Frank t-norm and T^* is the corresponding dual Frank t-conorm (see Klement et al, 2000, p. 110), C is a copula and C^* is the dual copula in the sense presented in Alsina et al, 2006, p. 21, and finally, $G(a, b) = \sqrt{ab}$ and $K_\lambda(a, b) = (1 - \lambda)a + \lambda b$ with $\lambda \in [0, 1]$, respectively.

It is obvious that the following statement is true.

Proposition 2. *The inequality* (2) *is satisfied for any pair $(\triangle, \circ) = (K_\lambda, K_{1-\lambda})$ where $K_\alpha(a, b) = (1 - \alpha)a + \alpha b$ with $\alpha \in [0, 1]$.*

Indeed, for arbitrary fixed $\lambda \in [0, 1]$ we have

$$K_\lambda(a, b) = (1 - \lambda)a + \lambda b = a - \lambda a + \lambda b \leq a + b - (\lambda a + (1 - \lambda)b) = a + b - K_{1-\lambda}(a, b).$$

Our attention is paid for the operation K_λ because of the following definition.

Definition 3. A function $f \in \mathcal{F}^{\bar{y}}$ is called convex if

$$f((1 - \lambda)u + \lambda v) \leq (1 - \lambda)f(u) + \lambda f(v)$$

for any $u, v \in X$ and $\lambda \in [0, 1]$. If $-f$ is convex, then f is called concave, i. e. the inequaity

$$f((1 - \lambda)u + \lambda v) \geq (1 - \lambda)f(u) + \lambda f(v)$$

holds for any $u, v \in X$ and $\lambda \in [0, 1]$.

It is necessary to mention that the original definition of convexity of a function, and concavity as well, is stated for real-valued function. We restrict our attention only for non-negative functions because the inequality (2) is connected with the range of monotone measure. The convexity of a function means geometrically that the points of the graph of $f|_{[u,v]}$ are under (or on) the chord joining the endpoints $[u, f(u)]$ and $[v, f(v)]$, for all $u, v \in X, u < v$; see Figure 1. Similarly, geometrical interpretation of concavity means that the points of the graph of $f|_{[u,v]}$ are above (or on) the chord joining the endpoints $[u, f(u)]$ and $[v, f(v)]$, for all $u, v \in X, u < v$.

Therefore, the inequality (2) can be provided in more general form

$$\varphi(a \triangle b) \leq \varphi(a) + \varphi(b) - \psi(a \circ b) \quad \text{for any} \quad f(a), f(b) \in m(\Sigma), \tag{3}$$

with $\circ\colon m(\Sigma) \times m(\Sigma) \to m(\Sigma)$ being a binary operation, and the condition (M) holds, where

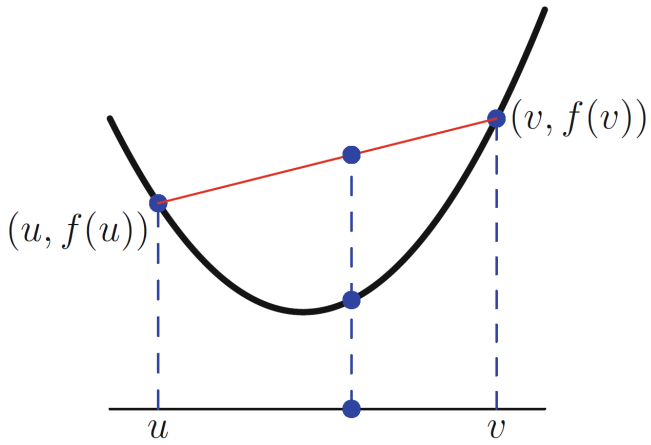


Figure 1 Convex function: the graph is under the chord.

(M)for each $C, D \in \Sigma$ we have $\psi(m(C \cup D)) \leq \psi(m(C) \circ m(D))$.

Then under the condition that $m \in \mathcal{M}$ is supermodular, we get

$$\begin{aligned}\varphi(m(C \cap D)) &\geq \varphi(m(C)) + \varphi(m(D)) - \psi(m(C \cup D)) \geq \varphi(m(C)) + \varphi(m(D)) - \psi(m(C \circ D)) \\ &\geq \varphi(m(C) \vartriangle m(D)).\end{aligned}$$

By similarity we can formulate the following.

Proposition 3. *Let $m \in \mathcal{M}$ be supermodular, $\circ\colon m(\Sigma) \times m(\Sigma) \to m(\Sigma)$ and conditions* (3) *and* (M) *hold. Then any functions $f|_A, g|_B \in \mathcal{F}^{\bar{y}}$ are m-$\vartriangle$-positively dependent.*

Based on Proposition 2 the previous statement we obtain.

Proposition 4. *The inequality* (3) *is satisfied for any pair $(\vartriangle, \circ) = (K_\lambda, K_{1-\lambda})$ where $K_\alpha(a, b) = (1-\alpha)a + \alpha b$ with $\lambda \in [0, 1]$ and $\varphi, \psi \in \mathcal{F}^{\bar{y}}$ being convex functions while $\psi \leq \varphi$.*

Indeed, for arbitrary fixed $\lambda \in [0, 1]$ we have

$$\begin{aligned}\varphi(K_\lambda(a, b)) = \varphi((1-\lambda)a + \lambda b) &\leq (1-\lambda)\varphi(a) + \lambda\varphi(b) = \varphi(a) + \varphi(b) - (\lambda\varphi(a) + (1-\lambda)\varphi(b)) \\ &\leq \varphi(a) + \varphi(b) - (\lambda\psi(a) + (1-\lambda)\psi(b)) \\ &\leq \varphi(a) + \varphi(b) - \psi(\lambda a + (1-\lambda)b) = \varphi(a) + \varphi(b) - \psi(K_{1-\lambda}(a, b)).\end{aligned}$$

It is worth mentioning that convex functions played a crucial role during the whole twentieth century, there was intense research activity and many significant results were obtained in geometric functional analysis, mathematical economics, convex analysis, and nonlinear optimization. Moreover, convexity of two non-negative functions imply comonotonicity of them, see Niculescu and Persson, 2018. Similarly, as there is provided in the paper Boczek et al, 2025, we can follow their considerations when using duality. The authors have shown that the dual class to the $m-\vartriangle-$positively dependent functions is the class of $m-\triangledown-$subadditive functions. Under fixed supmodular monotone measure $m \in \mathcal{M}$ the general version of considered inequality has the exepression

$$\varphi(a) + \varphi(b) - \psi(a \circ b) \leq \varphi(a \triangledown b) \quad \text{for any} \quad f(a), f(b) \in m(\Sigma)$$

and similar characterization of functions φ and ψ can be provided. In what follows by Proposition 4, $\varphi, \psi \in \mathcal{F}^{\bar{y}}$ are concave functions.

4 Conclusion

In this paper, the full characterization of functions $\varphi, \psi \in \mathcal{F}^{\bar{y}}$ satisfying the inequality

$$\varphi(a \triangle b) \leq \varphi(a) + \varphi(b) - \psi(a \circ b)$$

is provided. This problem is a partial solution to the inequality (2) stated in the recent paper Boczek et al, 2025 under fixed couple of operations $(\triangle, \circ) = (K_\lambda, K_{1-\lambda})$ where $K_\alpha(a, b) = (1 - \alpha)a + \alpha b$ and $\alpha \in [0, 1]$. As it has been shown, this operation refers to the definition of convex functions, which finds usage in different areas in mathematics. It is well-known fact that convex functions guarantee that every local minimum is also a global minimum – which significantly simplifies computations and analyses when optimizating processes. Additionaly, the property of convexity of function is equivalent to the Jensen inequality. In operations research, the Jensen inequality is widely used in risk and expected return analysis, especially when the utility or risk function is convex or concave. The standard form:

$$\varphi\left(\mathrm{E}[X]\right) \leq \mathrm{E}\left[\varphi(X)\right]$$

enables estimating the expected value of a transformed random variable. For instance, if φ represents a risk-averse utility function (concave), the inequality underpins key results in decision theory under uncertainty, see Ivelić Bradanović and Lovričević, 2024. In portfolio optimization (e.g., the Markowitz model Markowitz, 1952), it helps show that the expected loss is lower for a diversified portfolio, thus improving the robustness of the model.

Acknowledgements

This article was created by the implementation of the grant project APVV-17-0258, APVV-19-0418, VEGA 1/0508/22, VEGA 1/0383/25, KEGA 020TUKE-4/2023, KEGA 003TUKE-4/2024 and VEGA 1/0243/23.

References

Alsina C, Frank MJ, Schweizer B (2006) Associative Functions: Triangular Norms and Copulas. World Scientific

Armstrong TE (1993/94) Chebyshev inequalities and comonotonicity. Real Analysis Exchange 19:266–268

Boczek M, Hovana A, Hutník O (2019a) General Chebyshev type inequality for seminormed fuzzy integral. In: et al. VT (ed) Modelling Decision for Artificial Intelligence, LNAI, vol 11676. Springer, Switzerland, Cham, p 3–16

Boczek M, Hovana A, Hutník O (2019b) General form of Chebyshev type inequality for generalized Sugeno integral. International Journal of Approximate Reasoning 115:1–12

Boczek M, Hovana A, Hutník O, et al (2020) Hölder-Minkowski type inequality for generalized Sugeno integral. Fuzzy Sets and Systems 396:51–71

Boczek M, Hovana A, Hutník O (2025) Generalizing comonotonicity: New insights and dual perspectives. Fuzzy Sets and Systems 510:109364

Borzová-Molnárová J, Halčinová L, Hutník O (2015) The smallest semicopula-based universal integrals I: Properties and characterizations. Fuzzy Sets and Systems 271:1–17

Bustince H, Barrenechea E, Sesma-Sara M, et al (2017) Ordered directionally monotone functions: Justification and application. IEEE Transactions on Fuzzy Systems 26:2237 –2250

Dellacherie C (1971) Quelques commentaires sur les prolongements de capacités. Séminaire de probabilités (Strasbourg) 5:77–81

Dhaene J, Denuit M, Goovaerts MJ, et al (2002a) The concept of comonotonicity in actuarial science and finance: applications. Insurance: Mathematics and Economics 31:133–161

Dhaene J, Denuit M, Goovaerts MJ, et al (2002b) The concept of comonotonicity in actuarial science and finance: theory. Insurance: Mathematics and Economics 31:3–33

Dybvig P (1988) Distributional analysis of portfolio choice. Journal of Business 61:360–393

Hardy GH, Littlewood JE, Pólya G (1934) Inequalities. Cambridge University Press

Ivelić Bradanović I, Lovričević A (2024) Generalized jensen and jensen–mercer inequalities for strongly convex functions. Journal of Inequalities and Applications 2024(1):79

Kaluszka M, Okolewski A, Boczek M (2014) On Chebyshev type inequalities for generalized Sugeno integrals. Fuzzy Sets and Systems 244:51–62

Klement EP, Mesiar R, Pap E (2000) Triangular Norms. Kluwer, Dordrecht

Lehmann EL (1966) Some concepts of dependence. Annals Mathematics Statistics 37:1137–1153

Markowitz HM (1952) Portfolio selection. The Journal of Finance 7(1):77–91

Niculescu CP, Persson LE (2018) Convex functions and their application. Springer

Ouyang Y, Li J, Mesiar R (2019) A sufficient condition of equivalence of the Choquet and the pan-integral. Fuzzy Sets and Systems 355:100–105

Schmeidler D (1986) Integral representation without additivity. Proceedings of the American Mathematical Society 97:255–261

Yaari ME (1987) Dual theory of choice under risk. Econometrica 55:95–115

Received: 2025-06-08 Last revision: 2025-08-12 Accepted: 2025-08-12

The Industry-Level Heterogeneous Impacts of Investor Sentiment on Return, Volume, and Volatility: Evidence from the S&P500

Bahate Maidiya [1]

Abstract. This study examines the differentiated impacts of investor sentiment across various industries represented in the S&P 500 index constituents. Using extended Fama-French 3-factor and 5-factor models augmented by sentiment spread variables, combined with trading volume and return volatility as additional response variables, we estimate firm-level sensitivities via OLS regressions. The sample is classified by GICS sectors. A multi-stage statistical procedure, including Welch ANOVA and Games-Howell post-hoc tests, is applied to identify significant cross-industry differences in sentiment sensitivity. The analysis shows that industries such as Health Care, Financials, and Information Technology exhibit more pronounced return or volatility responses to shifts in sentiment, whereas other sectors display comparatively lower sensitivity. This research contributes to understanding the behavioral aspect of asset pricing and provides new industry-specific evidence on the role of investor sentiment.

Keywords: investor sentiment; S&P 500; behavioral finance

JEL Classification: G12, G14, G41
AMS Classification: 62P05, 91G80

1 Introduction

Numerous studies (Baker et al., 2006; Baker et al., 2007; Kostopoulos et al., 2018; Lei et al., 2012) have identified investor sentiment as a key behavioural factor shaping asset prices and trading patterns. Prior research has shown that sentiment affects different industries to varying degrees (Huang et al., 2014; Niu et al., 2023; Sayim et al., 2013; Tiwari et al., 2022; Xie, 2024). However, some studies report that investor sentiment from Internet message postings does not predict stock returns, volatility, or trading volume.(Kim et al., 2014). Conversely, prior research (Anghel et al., 2024) reports that investor sentiment exerts a significant positive influence on volatility connectedness within the energy sector, but a negative effect on defensive industries such as consumer services and utilities.

Building on this context, the present study investigates whether firms across different industries display varying degrees of sensitivity to investor sentiment, with a particular focus on trading volume and return volatility. By focusing on S&P 500 index constituents, this study contributes to the literature in two ways. First, it extends the Fama-French factor model by introducing sentiment diffusion variables to capture behavioural effects at the firm level. Second, it introduces the analysis of trading volume and volatility. Finally, it uses a robust multistage statistical procedure to systematically compare sentiment sensitivity across different industries. Specifically, we hypothesize: **H1**: Investor sentiment has a stronger effect on trading volume and volatility than on stock returns. **H2**: Industries vary in their sensitivity of both trading volume and volatility to changes in investor sentiment.

The findings indicate that while sentiment has only a modest effect on corporate returns, its influence on trading volume and volatility is both statistically significant and economically substantial across industries. There are significant cross-industry differences in the impact of investor sentiment on market activity.

The remainder of this paper is organized as follows. Section 2 introduces the data and methodology. Section 3 presents the empirical results and discusses the findings. Section 4 concludes the paper.

2 Data and methodology

The stock dataset comprises all S&P 500 constituents as of December 31, 2024, obtained from Yahoo Finance (2025). The S&P 500 index is chosen because it represents large-cap U.S. equities with high liquidity and diverse industry coverage, which reduces microstructure noise and ensures the results are representative of the broad

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

market. All data were collected from 1 January 2000 to 31 December 2024. The 25-year window provides a sufficiently large time series for reliable estimation of GARCH models and cross-sectional regressions, especially when working with firm-level and industry-level breakdowns. Since some newly added stocks in the index have only a short history of price records, we excluded stocks with time series lengths of less than 2,000 observations (approximately 8 years of daily data), reducing the database from 503 stocks to 478 stocks. This threshold helps eliminate newly listed or frequently replaced stocks with irregular data, ensuring better comparability across industries. While this may exclude a small number of firms, it significantly improves the robustness of the analysis. Industries are classified using the Global Industry Classification Standard (GICS) sectors, resulting in 11 industry groups. The market sentiment index used in this study is derived from the Investor Sentiment Survey conducted by the American Association of Individual Investors (AAII) (AAII, 2025). We use the bull-bear spread (percentage of bullish minus bearish responses) as the sentiment proxy. We employ the Fama-French three-factor model (FF3F) and five-factor model (FF5F) for analysis of returns. All data for the Fama-French factor model are sourced from Kenneth French's data library. In our analysis, each factor is treated as a time-series index representing market-wide characteristics. These factors include: Market risk premium (Rm-Rf, hereafter denoted as MKT): The value-weighted market return minus the risk-free rate. Size premium (SMB): The average return of small-capitalization stocks minus the average return of large-capitalization stocks. Value premium (HML): The average return of high book-to-market stocks minus the average return of low book-to-market stocks. Profitability premium (RMW): The average return of robust profitability stocks minus the average return of weak profitability stocks. Investment style premium (CMA): The average return of conservative investment stocks minus the average return of aggressive investment stocks. The risk-free rate (RF) is the one-month Treasury bill rate. The goal of the model is to more accurately predict and explain the impact of sentiment on stock returns by taking these factors into account.

We apply OLS for each firm individually. The estimation for both Fama-French 3-factor (FF3F) and 5-factor (FF5F) models, extended by a sentiment spread variable:

$$E\ (R_{i,t}) - R_{f,t} = \alpha_i + \beta_{1,i} \cdot MKT_t + \beta_{2,i} \cdot SMB_t + \beta_{3,i} \cdot HML_t + \gamma_i \cdot SENT_t\ , \quad (1)$$

$$E\ (R_{i,t}) - R_{f,t} = \alpha_i + \beta_{1,i} \cdot MKT_t + \beta_{2,i} \cdot SMB_t + \beta_{3,i} \cdot HML_t + \beta_{4,i} \cdot RMW_t + \beta_{5,i} \cdot CMA_t + \gamma_i \cdot SENT_t \quad (2)$$

where $E\ (R_{i,t})$ is the expected return of stock i at time t, $R_{f,t}$ is the risk-free return at time t, α_i is the intercept term, and all $\beta_{k,i}$ (k=1,...,5) parameters quantify the extent to which each factor affects stock returns. MKT_t refers to the market excess return at time t, SMB_t is the size premium (Small Minus Big) at time t, defined as the difference between the returns of small-capitalization stocks and the returns of large-capitalization stocks at time t. HML_t is the value premium (High Minus Low) at time t, defined as the difference between the returns of high book-to-market ratio stocks and low book-to-market ratio stocks at time t. RMW_t is the profitability premium (Robust Minus Weak) at time t, defined as the difference between the returns of strongly profitable firms and those of less profitable firms at time t. CMA_t is the investment style premium (Conservative Minus Aggressive) at time t, defined as the difference between the returns of low investment firms and those of highly invested firms at time t. $\gamma_i$is the parameters for sentiment, and $SENT_t$ is the sentiment index at time t.

Volatility is calculated using the GARCH(1,1) model, which is a widely accepted method for modeling time-varying volatility in financial time series. The GARCH(1,1) model is chosen due to its parsimony and its ability to capture key characteristics of financial volatility, such as volatility clustering and persistence (Bollerslev, 1986). The model assumes that fluctuations in an asset's price are not random, but depend on previous fluctuations. Given our objective of estimating firm-level volatility for over 470 stocks across multiple sectors over a long time period, the GARCH(1,1) model strikes a balance between computational feasibility and statistical adequacy. The formula of the GARCH model we used in this study is as follows:

$$R_{i,t} = \mu_i + \sigma_{i,t} \cdot \epsilon_{i,t}\ , \quad (3)$$

$$\sigma_{i,t}^2 = \omega + a \cdot \sigma_{i,t-1}^2 + b \cdot \epsilon_{t-j}^2\ , \quad (4)$$

where $R_{i,t}$ is the return on asset i at time t, μ_i the average return on asset i, which is usually assumed to be a constant, $\sigma_{i,t}$ is the volatility of asset i at time t, and $\epsilon_{i,t}$ is the random disturbance term of the standard normal distribution (i.e., $\epsilon_{i,t} \sim$ N(0,1)). $\sigma_{i,t}^2$ is the square of the volatility of asset i at time t. ω, a, and b are model parameters that need to be estimated from historical data. ω is the mean of the long-run volatility, a is the effect of the previous period's volatility on the current volatility, and b is the effect of innovations (i.e., stochastic perturbations) in the previous j periods on the current volatility.

To assess cross-industry heterogeneity, Welch ANOVA and Games-Howell post-hoc tests are employed, accommodating unequal variances and sample sizes.

3 Result

First, we ran four sets of regressions for all 479 stocks, grouped into 11 industry classifications: FF3F, FF5F, trading volume, and volatility, using the specifications in Eqs. (1) and (2) for the FF3F and FF5F return models, respectively, and Eqs. (3) and (4) for the volatility. We then examined, at the industry level, the proportion of firms with significant coefficients. Table 1 reveals a marked asymmetry across the three channels under study. Across all sectors, only a small fraction of firms show significant sentiment loadings in the FF3F or FF5F return regressions (below 13%). In contrast, the proportions rise sharply when the dependent variable is trading volume or volatility—precisely the pattern anticipated by H1. A distinct "high–high" cluster emerges here, comprising Real Estate, Energy, Materials, Financials, and Utilities. These industries are often cyclical or sensitive to interest rates and commodity prices, and many operate with relatively high fixed or financial leverage. In such settings, even modest swings in investor mood can amplify order flow and widen uncertainty premia. The high sensitivity in cyclical sectors (Materials, Energy, Real Estate) can be attributed to their exposure to macroeconomic uncertainty and commodity price fluctuations, which amplify investor sentiment effects (Niu et al., 2023; Patel et al., 2025). Real estate investment trusts (REITs) within the real estate industry are particularly vulnerable; both institutional and individual investors in this space have been shown to react strongly to shifts in sentiment (Das et al., 2015). The picture is more nuanced for other sectors. Consumer Staples shows a strong link to volume but much weaker sensitivity in volatility, a profile consistent with its reputation as a defensive sector—steady cash flows may draw rebalancing trades without triggering outsized price moves. Communication Services and Information Technology, by contrast, exhibit relatively low changes in trading volume but heightened volatility sensitivity. This pattern aligns with the role of intangible assets, rapid news cycles, and derivative activity, all of which can reprice risk without proportionately shifting share turnover. Industrials and Consumer Discretionary fall into a middle-to-high zone on both measures, reflecting their demand elasticity and broader macroeconomic exposure. Health Care, finally, appears muted on both dimensions, perhaps because its key risks—such as drug pipeline outcomes or regulatory approvals—are highly idiosyncratic and less tied to broad-based retail sentiment.

Industry	FF3_Ratio	FF5_Ratio	Volume_Ratio	Volatility_Ratio
Industrials	0.0423	0.0423	0.8732	0.7887
Health Care	0.1228	0.1228	0.7018	0.6667
Information Technology	0.0448	0.0448	0.4925	0.7910
Utilities	0.0333	0.0333	0.9333	0.8667
Financials	0.1111	0.0972	0.9028	0.9444
Materials	0.0435	0.0435	0.9565	0.9565
Consumer Discretionary	0.0612	0.0612	0.7755	0.8571
Real Estate	0.0345	0.0345	1.0000	0.9655
Communication Services	0.0476	0.0476	0.6667	0.9524
Consumer Staples	0.1351	0.1081	0.8378	0.6757
Energy	0.0455	0.0909	0.9545	1.0000

Table 1 Industry-level Proportion of Firms with Significant Sentiment Sensitivity ($p < 0.05$)

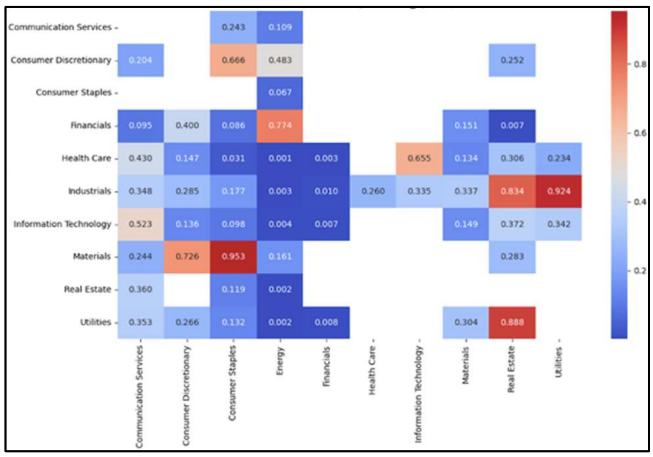


Figure 1 Heatmap of Industry Differences in Trading Volume Responses to Investor Sentiment.

Secondly, the cross-industry differences in volume sensitivity. Figure 1 shows the p-value heatmap of pairwise comparisons using the Games–Howell test, used to determine whether there are differences in the sensitivity of trading volume to sentiment (slope) across industries. Among the 55 pairs formed by the 11 industries, approximately 11 pairs had $p < 0.05$, indicating that there are cross-industry differences in trading volume channels, but the significance is concentrated in a few industry pairs and not across all industries : Energy vs. Health Care ($p = 0.001$), Real Estate ($p = 0.002$), Utilities ($p = 0.002$), Industrials ($p = 0.003$), and Information Technology ($p = 0.004$); Financials vs. Health Care ($p = 0.003$), Real Estate ($p = 0.007$), Information Technology ($p = 0.007$), Utilities ($p = 0.008$), and Industrials ($p = 0.010$); Health Care vs. Consumer Staples ($p = 0.031$).

Within the Cross-industry differences in volatility sensitivity. From Figure 2, we can see a much broader and stronger pattern of between-industry dispersion in the volatility slope. Extremely small p-values (rounded as 0.000 in the plot) are widespread—for example across Health Care, Information Technology, Energy, Financials, Utilities, and Consumer Staples—while only a few pairs remain non-significant (e.g., Financials–Real Estate $p = 0.921$). This result provides strong support for H2 in terms of volume and volatility channels.

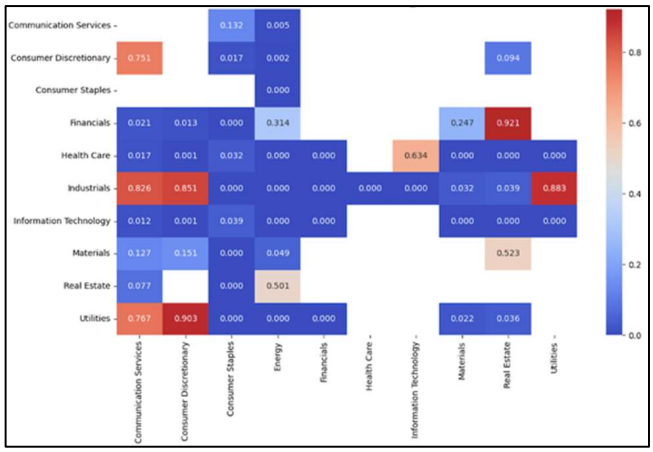


Figure 2 Heatmap of Industry Differences in Trading Volatility Responses to Investor Sentiment.

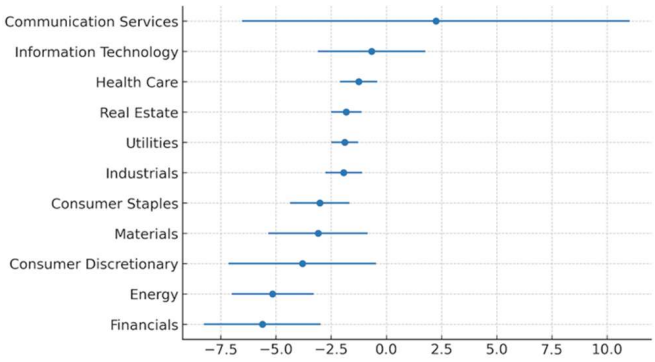


Figure 3 Industry-Level Mean (±95% CI) of Trading Volume Sensitivity to Investor Sentiment (mil.).

When explaining the direction and magnitude of differences, we can see in Figure 3 that the average slope for Communication Services is in the positive range (around 0 to +2.5 million, leaning towards +2.5), even though it is distributed across both positive and negative ends. This indicates that when sentiment rises, trading volume in this industry is more likely to increase. This industry composition includes media/entertainment platforms, interactive social media, communication services, etc., and the chain of 'sentiment—retail investor participation—speculation/options activity' is more pronounced. When sentiment is on the rise, attention and speculative intent increase in tandem, leading to higher trading volumes (with a positive average). The average slopes for Financials, Energy, and Consumer Discretionary are in a significant negative range (around −7.5 to −2.5 million, with Financials and Energy having larger absolute values), indicating that when sentiment rises, trading volume in these industries is more likely to decrease. When market sentiment is high, it often means that the economic environment is expected to improve, and there will be fewer 'marginal surprises' in terms of interest rates or credit expansion, leading to a decline in short-term trading interest. Large institutional investors have relatively stable positions in financial stocks, and high sentiment may actually reduce the frequency of reallocation. High sentiment may

correspond to market expectations of stable or moderate growth in energy demand, and such expectations are unlikely to trigger large-scale position adjustments. The remaining industries mostly fall within a range of slightly negative to near-zero values (approximately −3.5 to 0 million), with a negative direction but moderate intensity. Consistency with Figure 1 (p-value heatmap). The means between other sectors are closer (mostly weak negative/near zero), so the coverage of significant pairs in Figure 1 is more limited.

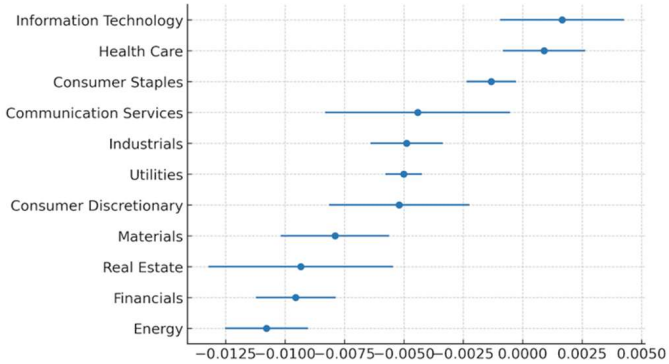


Figure 4 Industry-Level Mean (±95% CI) of Volatility Sensitivity to Investor Sentiment.

As shown in Figure 4, the average slopes for Information Technology and Health Care are in the positive range (IT is further to the right, followed by HC), indicating that when sentiment improves, the volatility of these two industries is more likely to increase (they are more sensitive to risk repricing). IT and healthcare are sectors with longer durations and more sensitive expectations. When optimism prevails, positive expectations are quickly amplified, leading to more active options/leveraged trading and prices that tend to 'move more aggressively,' thereby increasing volatility. Uncertainty comes often from R&D/regulatory/product milestones, and when sentiment is positive, expectations differences are exaggerated (some are more optimistic, while others wait and see), raising short-term volatility. The average slopes for Financials, Energy, and Real Estate are in a more significant negative range (around −0.01), indicating that when sentiment improves, the volatility of these industries is more likely to decrease. Financials, energy, and real estate are more sensitive to interest rates, oil prices, and rental discounts. Optimism is often accompanied by a convergence of risk premiums and a more predictable discount rate path, resulting in more stable prices and lower volatility (negative slope). Other sectors are mostly in negative territory, but the intensity is not as strong as the aforementioned three (with smaller absolute values). Consistency with Figure 2 (p-value heatmap). Figure 4 clearly shows a 'positive group (IT/HC) – negative group (Fin/Ene/RE)' differentiation. Therefore, in Figure 2, you see a significant number of paired values (large areas of dark blue): when the positive and negative groups are paired with most industries, the mean differences are generally larger, resulting in smaller p-values and broader coverage.

4 Conclusion

This study provides empirical support for both H1 and H2. Consistent with H1, investor sentiment shows a much stronger. In the FF3F and FF5F return regressions, the proportion of significant sentiment loadings is low across all sectors, while in the trading-volume and volatility models, significance rates are substantially higher, indicating that sentiment primarily affects turnover and risk premia rather than returns. In line with H2, significant cross-industry differences in sentiment sensitivity are identified. For trading volume, such differences are concentrated in a limited set of industry contrasts—most notably those involving extreme-mean sectors such as Energy and Financials—whereas for volatility, significant differences are widespread across industries. Mean ± 95% CI plots further show a clear polarity in both direction and magnitude: Communication Services displays strong positive volume sensitivity, while Financials and Energy are strongly negative; for volatility, Information Technology and Health Care show strong positive sensitivity, whereas Financials, Energy, and Real Estate are strongly negative.

These results have practical implications for execution, portfolio allocation, and risk management. In volatility-positive sectors such as IT and Health Care, volatility and tail-risk limits should be tightened during sentiment upswings, while in volatility-negative sectors such as Financials, Energy, and Real Estate, risk monitoring should focus on potential volatility rebounds when sentiment reverses. Execution strategies should anticipate sentiment-driven volume expansion in Communication Services and possible contraction in Financials and Energy. Moreover, the observed high-/low-sensitivity structure is not a static cyclical–defensive divide but is shaped by sectoral exposures to interest rates, commodity prices, news flow, and investor composition—factors that merit explicit inclusion, or interaction with sentiment measures, in future robustness checks.

Acknowledgements

The research was financially supported by the SGS project no. SP2025/043 of VSB – Technical University of Ostrava.

References

Anghel D, Caraiani P (2024) The volatility connectedness of US industries: The role of investor sentiment. Economics Letters, 235, 111578. https://doi.org/10.1016/j.econlet.2024.111578

Baker M, Wurgler J (2006). Investor sentiment and the cross‐section of stock returns. The journal of Finance, 61(4), 1645-1680. https://doi.org/10.1111/j.1540-6261.2006.00885.x

Baker M, Wurgler J (2007). Investor Sentiment in the Stock Market. Journal of Economic Perspectives, 21(2), 129–151. https://doi.org/10.1257/jep.21.2.129

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

Das P, Freybote J, Marcato G (2015). An Investigation into Sentiment-Induced Institutional Trading Behavior and Asset Pricing in the REIT Market. The Journal of Real Estate Finance and Economics, 51(2), 160–189. https://doi.org/10.1007/s11146-014-9490-z

Huang C, Yang X, Yang X, Sheng H (2014). An empirical study of the effect of investor sentiment on returns of different industries. Mathematical Problems in Engineering, 2014(1), 545723. https://doi.org/10.1155/2014/545723

Kim S, Kim D (2014). Investor Sentiment from Internet Message Postings and Predictability of Stock Returns. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2423752

Kostopoulos D, Meyer S (2018). Disentangling investor sentiment: Mood and household attitudes towards the economy. Journal of Economic Behavior & Organization, 155, 28–78. https://doi.org/10.1016/j.jebo.2018.08.003

Lei V, So S, Zou M (2012). Investor Sentiment—Relationship between VIX and Trading Volume. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2136802

Niu H, Lu Y, Wang W (2023) Does investor sentiment differently affect stocks in different sectors? Evidence from China. International Journal of Emerging Markets, 18(9), 3224–3244. https://doi.org/10.1108/IJOEM-11-2020-1298

Patel V, Amilan S, Vairasigamani P (2025) The relationship of market sentiment and sector return across time and frequency – a wavelet coherence analysis. The Economics and Finance Letters, 12(2), 388–402. https://doi.org/10.18488/29.v12i2.4255

Sayim M, Morris P, Rahman H (2013) The effect of US individual investor sentiment on industry‐specific stock returns and volatility. Review of Behavioural Finance, 5(1), 58‐76. https://doi.org/10.1108/RBF-01-2013-0006

Tiwari A, Abakah E, Bonsu C, Karikari N, Hammoudeh S (2022) The effects of public sentiments and feelings on stock market behavior: Evidence from Australia. Journal of Economic Behavior & Organization, 193, 443–472. https://doi.org/10.1016/j.jebo.2021.11.026

Xie J (2024) The Impact of Investor Sentiment on Stock Returns of Different Industries in China. Advances in Economics, Management and Political Sciences. 122: 152-159. https://doi.org/10.54254/2754-1169/202-4.17746

Data

American Association of Individual Investors. AAII Investor Sentiment Survey. Retrieved June 1, 2025, from https://www.aaii.com/sentimentsurvey

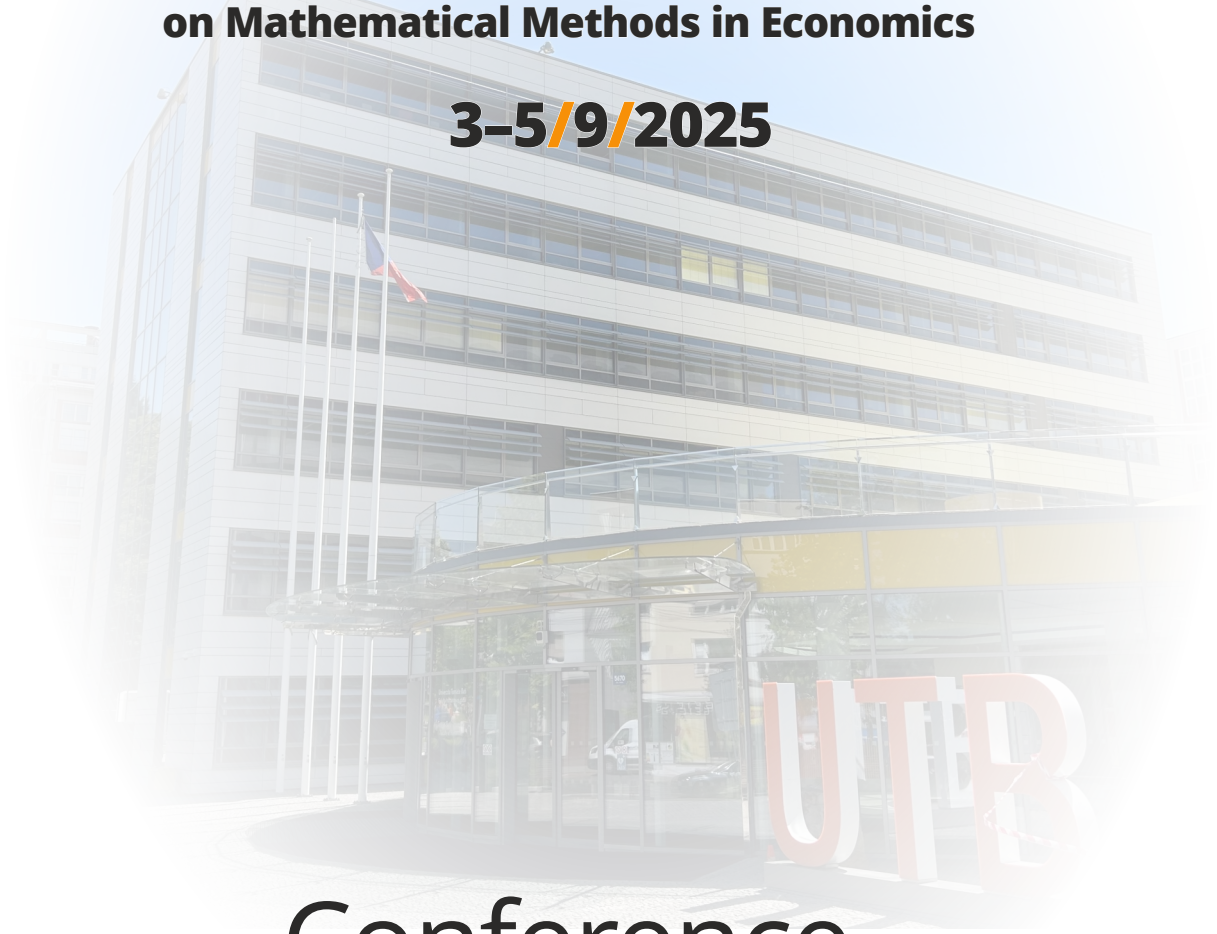
French, K. R. Data Library. Retrieved June 1, 2025, from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Yahoo Finance. Historical Stock Prices. Retrieved June 1, 2025, from https://finance.yahoo.com

MME 2025

The 43rd International Conference on Mathematical Methods in Economics

3–5/9/2025



Conference Proceedings

Tomas Bata University in Zlín

Tomas Bata University in Zlín
Faculty of Applied Informatics

Tomas Bata University in Zlín
Faculty of Management and Economics

ČSOV

SOLVERTECH