Volume 27, Number 1, March 2026 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

ISSN 1234-7655

Head of Editorial Office/Secretary
Patryk Barszcz, *Statistics Poland, Warsaw, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 – 608 33 66*

Managing Editor
Adriana Nowakowska, *Statistics Poland, Warsaw, Poland, e-mail: a.nowakowska3@stat.gov.pl*

Technical Assistant
Rajmund Litkowiec, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: r.litkowiec@stat.gov.pl*

Address for correspondence
Statistics Poland, al. Niepodległości 208, 00-925 Warsaw, Poland, tel./fax: +48 22 – 825 03 95

CONTENTS

STATISTICS IN TRANSITION new series, *March 2026*
Vol. 27, No. 1, pp. III

Submission information for Authors

Statistics in Transition new series (SiTns) is an international journal published jointly by the Polish Statistical Association (PTS) and Statistics Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series*.

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl,
GUS/Statistics Poland,
Al. Niepodległości 208, R. 296, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: https://sit.stat.gov.pl/ForAuthors.

STATISTICS IN TRANSITION new series, *March 2026*
Vol. 27, No. 1, pp. V–VI

Policy statement

The broad objective of *Statistics in Transition new series* is to advance the statistical and associated methods used primarily by statistical agencies and other research institutions. To meet that objective, the journal encompasses a wide range of topics in statistical design and analysis, including survey methodology and survey sampling, census methodology, statistical uses of administrative data sources, estimation methods, economic and demographic studies, and novel methods of analysis of socio-economic and population data. With its focus on innovative methods that address practical problems, the journal favours papers that report new methods accompanied by real-life applications. Authoritative review papers on important problems faced by statisticians in agencies and academia also fall within the journal's scope.

***

Abstracting and indexing databases

Statistics in Transition new series is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

CEJSH (The Central European Journal of Social Sciences and Humanities)

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Emerging Sources Citation Index (ESCI) – Web of Science Core Collection

Electronic Journals Library

Elsevier – Scopus

Genamics JournalSeek

Google Scholar

Index Copernicus

J-Gate

JournalGuide

JournalTOCs

Keepers Registry

MIAR

Microsoft Academic

OpenAIRE

ProQuest – Summon

Publons

QOAM (Quality Open Access Market)

ReadCube

RePec

SCImago Journal & Country Rank

TDNet

Technische Informationsbibliothek (TIB) – German National Library of Science and Technology

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. VII–XI

From the Editor

The March issue of *Statistics in Transition new series*, which opens this year's series of our quarterly, contains a set of twelve articles by twenty-two authors from seven countries (in order of appearance): Lithuania, Poland, India, Morocco, Iran, Malaysia, United Kingdom.

Let me take this opportunity to share with our readers wonderful news about a long-awaited promotion of *Statistics in Transition new series* in the international publication space thanks to its selection for inclusion into the prestigious Emerging Sources Citation Index (ESCI) database. The Clarivate-Web of Science evaluation report states: "Congratulations! 'Statistics in Transition' has been selected for inclusion in Web of Science™. Welcome to Web of Science, the most trusted, publisher-independent global citation database." The indexing of 'SiTns' in the *Journal Citation Reports (JCR)* brings increased recognition of our periodical and increases its impact on the development of statistical science.

On behalf of the entire Editorial Office, I would like to personally thank all our colleagues and partners for their contribution to these successful efforts: authors, reviewers, readers, and promoters – members of the Editorial Board and Associate Editors.

Needless to say, this important achievement makes all of us not only more assertive about the way we proceed but also more motivated to do everything we can to ensure our journal is included in other prestigious indexing/abstracting databases as well.

This issue is structured as usual: *Original research papers* are followed by selected *Conference papers* and by *Research Communications and Letters*.

Original research papers

The first paper by **Dalius Pumputis** entitled ***Optimal sample allocation in multivariate stratified sampling: a comparison of deterministic and stochastic optimization algorithms*** addresses the problem of optimal sample allocation in multivariate stratified sampling, where survey accuracy and cost-efficiency are the key concerns. Two optimization formulations are examined: one aims to minimize the total survey cost subject to constraints on the precision of the estimators of the population totals, while the other seeks to minimize a weighted sum of the relative variances of these estimators, given a fixed total survey budget. Classical and modern optimization approaches are reviewed and evaluated, including Integer Programming Algorithms (IPA), Bethel's

Algorithm (BA), Constrained Optimization by Linear Approximations (COBYLA), and three stochastics, namely Generalized Simulated Annealing Algorithm (GSAA), Particle Swarm Optimization (PSOA) and Biased Random-Key Genetic Algorithm (BRKGA).

Jacek Białek's article ***Sampling techniques in the CPI measurement*** discusses general concepts and techniques of survey sampling that are crucial for the construction of price indices. Both probability and non-probability sampling techniques are discussed and illustrated with real data, also embracing sampling scanned products. One of the approaches used for such data is the dynamic approach, which involves monthly sampling by applying appropriate data filters. This technique can be seen as a special form of cut-off sampling. The empirical study investigates the effect of data filtering on the level of price indices. The main conclusion is that the low-sales filter has the most significant impact on reducing the size of the scanner dataset. The second important conclusion is that changing the order of data filtering has minimal impact on the value of the price index.

In the next paper**,** by **Mahamood Usman, *Optimality of classical difference estimators of finite population variance under random non-response with comparative study***, the issue of calculating the finite population variance when faced with random non-response is discussed with focus on dealing with data in such fields like medical sciences, environmental sciences and business studies. Using the range of an auxiliary variable across three different methodologies of random non-response, the author developed several novel difference-type estimators of population variance along with their optimal models. The properties of the proposed estimators under large sample approximations and determined their optimum situations in each strategy were studied. The introduced estimators can be viewed as an advancement of traditional difference estimators. A comparative analysis based on some real datasets as well as simulated datasets was conducted. The proposed estimators showed reduced variances when assessed in terms of the enhanced percentage relative efficiencies (PRE) compared to some standard ratio and difference-type estimators relevant to the respective methodologies.

The article ***The impact of Cyber Supply Chain Risk Management on Supply Chain 4.0***, by **Abdellah Sassi**, **Mohamed Ben Ali**, **Oumaima Oullada**, and **Said Rifai**, investigates the influence of Cyber Supply Chain Risk Management (CSCRM) on Supply Chain 4.0 (SC 4.0) using a causal model to evaluate the connection between Cyber Security (CS), CSCRM, and SC 4.0. The link between CS and CSCRM, and between CSCRM and the levers of SC 4.0 are under evaluation. The results highlight that CSCRM significantly influences various supply chain activities. The findings show that the integration of CSCRM, supported by CS is essential for improving the performance of SC 4.0.

The paper by **Abbas Eftekharian**, **Morad Alizadeh**, **Vahid Ranjbar**, **Omid Kharazmi**, and **Gholamhossein Hamedani,** ***An extended odd log-logistic-Lindley distribution with properties, applications and Bayesian estimation*** introduces a four-parameter extended odd log-logistic-Lindley distribution from which moments, hazard, and quantile functions are subsequently obtained. The statistical properties of this distribution show the high flexibility of the proposed distribution. The maximum likelihood and OLS estimators of the extended odd log-logistic-Lindley parameters are studied and a simulation study is carried out for evaluating the performance of the estimation methods; the usefulness of the new distribution is illustrated using two real data sets. Finally, Bayesian analysis and efficiency of Gibbs sampling are provided on the basis of two real data sets.

Julia Fidler's and **Łukasz Matysiak's** article ***Modeling the impact of demand, supply, and budget constraints on consumer preferences*** combines Paul Samuelson's classical theory of revealed preferences with dynamic demand and supply mechanisms, using Afriat's theorem and extensions by Varian and Mas-Colell to construct utility functions without survey data. Critical voices (e.g. Dryzek) prompt a reexamination of the assumptions of full information and fixed preferences, inspiring to propose the F(w, z) function that accounts for the strength of market fluctuations. Empirical simulations and an analysis of market equilibrium stability provide new insights into economic policy and marketing strategies.

The next paper entitled ***The comparison of the hidden Markov model with machine learning techniques in agricultural prediction*** by **Muraleedharan Vyshnavi** and **Madaswamy Muthukumar** demonstrates that Hidden Markov Models (HMMs) significantly outperform other predictive methods in forecasting agricultural data. HMMs exhibit the lowest Mean Absolute Error (MAE) and the highest R-squared values, signifying their superior accuracy and reliability. The residual analysis further confirms the model's robustness, as the residuals are randomly distributed with no identifiable patterns, indicating a strong model fit. HMMs have proven to be highly effective in capturing the temporal and stochastic characteristics of agricultural data, surpassing traditional machine learning techniques in performance. These results underline the potential of HMMs as a reliable tool for agricultural forecasting, enabling stakeholders to make data-driven decisions and develop strategic plans based on accurate predictions.

The article by **Jayanthi Arasan,** ***Jackknife-based diagnostics for non-monotonic hazard survival model with interval-censored data*** focuses on jackknife-based model diagnostics for a non-monotonic two-parameter hazard survival regression model (TBPR) when data is interval and right-censored. This distribution is very flexible, because it accommodates both monotonic and bathtub-shaped hazard rates. This research proposes a bias-corrected jackknife harmonic mean and a random imputation

technique to obtain the altered Cox-Snell (r∗Ci), adjusted Martingale (r∗Mi) and Schoenfeld (r∗ Si) residuals. Two simulation studies were conducted to assess the performances of the altered residuals and their ability to detect extreme observations and outliers at various censoring proportions (cp) and sample sizes (n) for this model. The results indicated that the altered residuals based on jackknife outperformed other residuals at cp and n levels. The proposed methods are employed to a real dataset on Hodgkin's Disease with the prior treatment group as the covariate. The results showed that the altered residuals work well to address model adequacy and identify potential outliers in the dataset.

Conference papers

XXXXII Multivariate Statistical Analysis 2024, Lodz, Poland

The next paper ***The application of BERTopic models to in the analysis of Polish research publications in the field of economics and management*** by **Paweł Lula** analyzes topics from the field of economics and management discussed in the Polish publications from 2000 to 2024. The research process allowed the identification of the main topics and the evaluation of their importance in subsequent years covered by the analysis. The BERTopic model was chosen as the main research method. The paper presents both the theoretical basis of the employed research method and the results of its application to the analysis of the Polish publication achievements registered in the Scopus database. The paper presents a description of topics identified, a specification of the relationship between them and changes in the importance of each topic between 2000 and 2024. All calculations were performed using computer programs prepared in Python language.

Wioletta Grzenda's and **Agnieszka Marszałek's** article ***The role of education and gender in shaping career paths of Polish millennials: a shared frailty survival model analysis*** examines the influence of gender and the level of education on job mobility among young employees, using the Polish labor market as an example. When analyzing job changes, the authors go beyond previous studies by considering the duration of individual job episodes and the time-varying nature of some characteristics in young people, such as the level of education or the marital status. The analysis was based on survival analysis methods, including frailty models. Using data from the Generation and Gender Survey, it was found that the impact of the examined factors on job mobility varied by gender. The influence of having a child on job mobility was significant only for women. Mothers have a lower risk of job change than childless women. The stabilization of men's careers takes place over time and is associated with leaving the family home and marriage. Moreover, having higher education has a greater impact on the risk of job changes for men than for women.

XV Scientific Conference MASEP 2024 – Measurement and Assessment of Social and Economic Phenomena, Warsaw, Poland

Agnieszka Choczyńska's paper ***Is the GPT model suitable for sentiment analysis? Testing for geographical, political and gender bias*** focusses on the GPT-4o-mini model by OpenAI which is tested for the presence of geographic, political and gender bias in the case of Polish economic news headlines. It has been found that the model consistently differs in sentiment scores for the same sentence, depending on the country mentioned. A remedy to this problem is proposed, which masks the references to countries and nationalities using the GPT model. Some differences in sentiment scores resulting from explicit references to gender or political parties are also identified, although these types of bias are considerably weaker than geographical bias.

Research Communicates and Letters

Olena Kaminska, in the paper ***Survey sampling in wartime: addressing challenges for cross-national and longitudinal studies,*** outlines the design of two samples: one for a cross-national European Social Survey in Ukraine (ESS), and the other for the longitudinal household panel study, UKRAINS. By carefully addressing these complex sampling challenges, it is possible to develop high-quality probability samples that account for population mobility and unpredictability in a wartime context. A cross-sectional study, such as the ESS, is easier to plan since its purpose is to capture a snapshot of the population at a single point in time. Although the population described in such a study may quickly become invalid, it remains relevant to a critical historical moment in the country's life. However, analysts must be aware of missing subgroups: IDPs living in non-private households, servicemen (who now constitute a larger proportion of the population than in non-conflict circumstances), and individuals who are abroad unless specifically covered.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 1–20, https://doi.org/10.59139/stattrans-2026-001
Received – 15.04.2025; accepted – 21.08.2025

Optimal sample allocation in multivariate stratified sampling: a comparison of deterministic and stochastic optimization algorithms

Dalius Pumputis[1]

Abstract

This study addresses the problem of optimal sample allocation in multivariate stratified sampling, where survey accuracy and cost-efficiency are the key concerns. Two optimization formulations are examined: one aims to minimize the total survey cost subject to constraints on the precision of the estimators of the population totals, while the other seeks to minimize a weighted sum of the relative variances of these estimators, given a fixed total survey budget. Classical and modern optimization approaches are reviewed and evaluated, including Integer Programming Algorithms (IPA), Bethel's Algorithm (BA), Constrained Optimization by Linear Approximations (COBYLA), and three stochastics, namely Generalized Simulated Annealing Algorithm (GSAA), Particle Swarm Optimization (PSOA) and Biased Random-Key Genetic Algorithm (BRKGA). Using synthetic and real-world populations, numerical experiments demonstrate that IPA consistently achieves the global minimum and serves as the benchmark. While BA underperforms, BRKGA emerges as a competitive alternative, closely matching IPA in most scenarios. Results also highlight the impact of variable skewness on allocation efficiency, with real-world datasets being more complex and thus having higher sampling demands. The findings underscore the importance of adaptive, integer-feasible optimization methods for accurate and cost-effective survey design.

Key words: constrained optimization by linear approximations, integer programming, multivariate stratified sampling, optimal sample allocation, stochastic optimization.

1. Introduction

To obtain accurate estimates, surveys often employ stratification of a finite population. This statistical technique involves dividing the survey population into several distinct, non-overlapping, and internally homogeneous groups known as strata. Independent samples are then drawn from each of these groups. When stratified sampling is selected as the sampling method, the initial task is to define the boundaries of the strata.

After the strata boundaries have been established and the total sample size n has been decided, the next step involves allocating the sample size across the strata. Various allocation strategies are available, such as equal, proportional, or Neyman allocation (Neyman 1934). Equal and proportional methods are generally efficient when within-stratum variances are similar. In contrast, the Neyman method is more appropriate when strata vary significantly, as it prioritizes drawing fewer samples from more homogeneous strata and more from those with greater internal variability.

[1] Vilnius Gediminas Technical University (VILNIUS TECH), Lithuania.
E-mail: dalius.pumputis@vilniustech.lt. ORCID: https://orcid.org/0000-0003-0954-0663.

Neyman allocation relies on a formula designed to minimize both the survey cost C and the variance of the estimator for a single study variable. However, modern surveys often focus on multiple variables. In such cases, an allocation optimized for one variable may not be optimal for others, resulting in what is known as the multivariate optimal sample allocation problem.

This issue has been addressed by several researchers, beginning with Yates (1960), who proposed minimizing a weighted sum of the variances of the estimates for all survey variables, under the constraint of a fixed total sample size. Later, Chatterjee (1967) extended this line of inquiry by deriving an expression for the increase in variance when a non-optimal allocation is used, offering a framework for quantifying the deviation from optimality in multivariate settings.

Ahsan and Khan (1982) formulated the multivariate allocation problem with stratum-level overhead costs as a nonlinear program, minimizing total cost subject to variance constraints. Bethel (1985) proposed a convex programming algorithm that is simple to implement and converges efficiently. He later extended this work (Bethel 1989) by incorporating linear variance constraints and deriving optimal allocations using Lagrangian multipliers, providing a practical algorithm with demonstrated convergence.

Subsequent work has focused on obtaining integer and compromise allocations. Khan et al. (1998), Khan and Ahsan (2003), and Khan et al. (2010) developed dynamic and goal programming methods to derive integer-valued, compromise solutions, incorporating auxiliary information where available. Swain (2013) and Varshney et al. (2014) also applied goal programming to balance efficiency and practicality in multivariate settings.

Kadane (2005) introduced a dynamic sampling plan that minimizes variance at every stage, extending Neyman's approach to sequential designs. Brito et al. (2015) proposed a binary integer programming model that offers improved performance over existing algorithms in complex survey scenarios.

Since study variable parameters are often unknown in advance, Dayal (1985) suggested using auxiliary variables correlated with the variable of interest to guide sample allocation, showing that proportional allocation based on such variables can outperform approximations of Neyman allocation. Reddy et al. (2018) used auxiliary data with dynamic programming to optimize stratum boundaries and sample sizes in health surveys, greatly improving estimation efficiency over traditional methods.

While many allocation methods rely on approximations or rounding to achieve practical sample sizes, these approaches may lead to suboptimal or even infeasible results. Wright (2017) addressed these limitations by proposing exact optimal allocation algorithms that avoid common issues with Neyman allocation, such as non-integer solutions, post-rounding inefficiencies, and allocations exceeding stratum sizes. Expanding on this, Wright (2020) developed an exact algorithm with cost and sample size bounds, using cost-weighted function decomposition to offer a flexible, efficient framework that includes several traditional methods as special cases.

Recent studies propose methodological advances for optimum and compromise allocation in multivariate stratified sampling, tackling issues like non-response, cost, uncertainty, fuzzy environments, and practical constraints.

Haq et al. (2020) tackled compromise allocation in multivariate stratified sampling under non-response and fixed costs by converting an integer non-linear problem into a binary goal programming model, solved with flexible fuzzy goals for population mean estimation. Mahfouz et al. (2023) proposed a stochastic compromise allocation model using multi-objective programming to minimize survey cost and stratum variances. Through chance-constrained programming and simulations, they showed it provides the most efficient allocations. Raghav et al. (2023) addressed compromise allocation under response and non-response using multi-objective intuitionistic fuzzy programming with optimistic and pessimistic strategies, demonstrating applicability through simulations in wildlife, agriculture, and marketing surveys. Jalil et al. (2023) proposed a hierarchical multi-level programming model for compromise allocation under non-response and budget constraints, using fuzzy methods to optimize allocations and improve survey efficiency, flexibility, and cost-effectiveness. Gupta et al. (2024) modeled compromise allocation as deterministic integer programming solved with intuitionistic fuzzy programming, showing via computations that it reduces variances and errors, improving precision in microeconomic surveys. Wesołowski et al. (2024) developed a recursive Neyman algorithm (RNABOX) for stratum sample sizes under box constraints, proving optimality with Karush-Kuhn-Tucker theory and implementing it in R as a generalization of classical Neyman allocation.

To obtain optimal or near-optimal solutions for multivariate sample allocation problems, general-purpose optimization techniques such as the Generalized Simulated Annealing Algorithm (Tsallis, 1996) and other metaheuristic methods can be applied, as demonstrated in this study.

Unlike prior work (e.g. Mahfouz et al., 2023) that developed specific stochastic models, our study introduces a broader comparison of stochastic, deterministic, and hybrid optimization algorithms under two canonical allocation formulations. We benchmark results against a globally optimal integer programming solution and apply the methods to both synthetic and real populations, including high-dimensional, skewed, and correlated data, providing new insights into practical performance under diverse conditions.

The structure of the paper is as follows. Section 2 outlines fundamental concepts and definitions related to stratified sampling and introduces two multivariate sample allocation problems along with relevant algorithms. Section 3 presents simulation results illustrating the performance of different allocation methods. Lastly, Section 4 offers concluding remarks.

2. Formulations and algorithms for sample allocation

Consider a finite population denoted by $\mathscr{U} = \{u_1, u_2, \ldots, u_N\}$, consisting of N distinct units. Assume there are m study variables $y^{(1)}, y^{(2)}, \ldots, y^{(m)}$, each defined over the population $\mathscr{U}$ and taking real values. For each variable $y^{(j)}$, the corresponding values for all population units are given by $y_1^{(j)}, y_2^{(j)}, \ldots, y_N^{(j)}$, where $j = 1, 2, \ldots, m$.

Now, suppose the population is partitioned into H non-overlapping and exhaustive strata, denoted by $\mathscr{U}_1, \mathscr{U}_2, \ldots, \mathscr{U}_H$, such that:

$$\mathscr{U} = \bigcup_{h=1}^{H} \mathscr{U}_h,$$

with each stratum $\mathscr{U}_h$ containing N_h units, for $h = 1,2,\ldots,H$. From each stratum, a simple random sample $\mathbf{s}_h \subset \mathscr{U}_h$ of size n_h is selected without replacement. The overall sample $\mathbf{s}$, the total population size N, and the total sample size n satisfy the following relationships:

$$\mathbf{s} = \bigcup_{h=1}^{H} \mathbf{s}_h, \quad N = \sum_{h=1}^{H} N_h, \quad n = \sum_{h=1}^{H} n_h.$$

The quantities of interest are the finite population totals for each variable:

$$t_j = \sum_{i=1}^{N} y_i^{(j)}, \quad j = 1,2,\ldots,m.$$

These totals, $t_1, t_2, \ldots, t_m$, can be estimated using the Horvitz-Thompson estimator (Horvitz and Thompson 1952):

$$\hat{t}_j = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}^{(j)}, \quad j = 1,2,\ldots,m,$$

where $y_{hi}^{(j)}$ denotes the i-th observed value of variable $y^{(j)}$ in the sample $\mathbf{s}_h$ from stratum $\mathscr{U}_h$.

The variance of the estimator $\hat{t}_j$ for each $j = 1,2,\ldots,m$ is given by:

$$V(\hat{t}_j) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{hj}^2}{n_h}, \tag{1}$$

where s_{hj}^2 denotes the variance of variable $y^{(j)}$ within stratum $\mathscr{U}_h$.

Because the variance (1) of the estimators is determined solely by the sample sizes chosen for each stratum – given that the number of strata (H), population sizes within strata (N_h), and within-stratum variances (s_{hj}^2 for $h = 1,2,\ldots,H$ and $j = 1,2,\ldots,m$) are fixed once the stratification is set – the level of variance can be managed through the appropriate selection of sample sizes $n_1, n_2, \ldots, n_H$. Consequently, decreasing these sample sizes leads to higher variance and a greater coefficient of variation in the total estimates, which in turn can reduce the accuracy of the results. Nonetheless, to limit the total survey cost, expressed as $\sum_{h=1}^{H} c_h n_h$, where c_h is the per-unit cost of sampling in stratum U_h, it is often necessary to reduce sample sizes. To address this trade-off, various sample allocation strategies – such as those developed by Kokan and Khan (1967), Bethel (1985, 1989), Ahsan and Khan (1982), Brito et al. (2015), among others – have been proposed to balance the need for precision in the survey variables of interest with cost efficiency.

Kish, L., (1976), Khan and Ahsan (2003), Garciá and Cortez (2006), Khan et al. (2011), and others have addressed the problem of optimal allocation in multivariate stratified sampling, focusing on optimizing allocation strategies with respect to the variances of estimators, under constraints such as total sample size or cost. Their work involves various mathematical programming approaches – including nonlinear, dynamic, and convex optimization – to balance precision and resource limitations in survey design.

The contributions discussed above allow us to distinguish two main directions in approaching the sample allocation problem. These perspectives form the basis for the two problem formulations presented below.

Problem 1. Find strata sample sizes $n_1, n_2, \ldots, n_H$, which minimize the total survey cost

$$C = \sum_{h=1}^{H} c_h n_h \tag{2}$$

and satisfy the following inequalities:

$$n_{\min} \leq n_h \leq N_h \quad (h = 1, \ldots, H), \tag{3}$$

$$\frac{\sqrt{V(\hat{t}_j)}}{t_j} \leq CV_j \quad (j = 1, \ldots, m), \tag{4}$$

where CV_j, for $j = 1, 2, \ldots, m$, are the pre-specified coefficients of variation of the estimators $\hat{t}_j$, $j = 1, 2, \ldots, m$.

In this formulation, constraint (3) ensures that each stratum receives a sample size between $n_{\min}$ and its population size. Constraint (4) keeps the coefficient of variation of each estimator within the target CV_j.

Problem 2. Find strata sample sizes $n_1, n_2, \ldots, n_H$ that minimize the weighted sum of the relative variances of the estimators of totals

$$\sum_{j=1}^{m} w_j \frac{1}{t_j^2} \sum_{h=1}^{H} s_{hj}^2 \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \tag{5}$$

and satisfy the following inequalities:

$$n_{\min} \leq n_h \leq N_h \quad (h = 1, \ldots, H), \tag{6}$$

$$\sum_{h=1}^{H} c_h n_h \leq C^*, \tag{7}$$

where C^* is the total cost, defined as a function of the available survey budget. The weights w_j, for $j = 1, 2, \ldots, m$, are predetermined values associated with the importance of each variable of interest, such that $0 < w_j < 1$ and $w_1 + w_2 + \ldots + w_m = 1$.

The constraint in (6) is identical to that in (3), while the constraint in (7) ensures that the total cost is less than or equal to C^*, which is defined based on the available survey budget.

Bethel's Algorithm (BA). Bethel (1985, 1989) solved Problem 1 without incorporating constraint (3), and developed an algorithm that is guaranteed to converge to a solution (when one exists). This was achieved by applying the Kuhn and Tucker (1951) Theorem and the method of Lagrange multipliers to tackle the optimization problem. Later, some implementations – such as the `bethel()` function in the `SamplingStrata` package for the R programming language – modified the algorithm to incorporate constraint (3).

In this work, we also employ the **Generalized Simulated Annealing Algorithm (GSAA)** introduced by Tsallis and Stariolo (1996), designed for global optimization of real-valued

functions $C(\mathbf{n}) : \mathbb{R}^H \to \mathbb{R}$, where $C(\mathbf{n})$, in the context of our paper, corresponds to the survey cost function or a weighted sum of the relative variances of the estimators of totals, and $\mathbf{n} = (n_1, n_2, \ldots, n_H)$. While originally formulated for unconstrained problems, GSAA can be extended to constrained settings by incorporating penalty terms into the objective or by applying constraint-preserving sampling strategies.

The algorithm is based on a generalized entropy functional from nonextensive statistical mechanics:

$$S_q = k \frac{1 - \sum_i p_i^q}{q - 1}, \quad q \in \mathbb{R},$$

which reduces to the Shannon (1948) entropy as $q \to 1$. Here, $\{p_i\}$ denotes the probabilities of the microscopic configurations, and k is a conventional positive constant. The two main parameters of GSAA are q_V and q_A, which control the sampling distribution and the acceptance probability, respectively.

Candidate moves are generated using a power-law visiting distribution $g_{q_V}(\Delta \mathbf{n}_t)$, controlled by the parameter q_V, and the annealing temperature $T_{q_V}^{(V)}(t)$ at iteration t.

The acceptance of uphill moves is governed by a generalized Metropolis rule:

$$P_{q_A}(\mathbf{n}_t \to \mathbf{n}_{t+1}) = \begin{cases} 1 & \text{if } C(\mathbf{n}_{t+1}) < C(\mathbf{n}_t), \\ \left(1 + (q_A - 1)\dfrac{C(\mathbf{n}_{t+1}) - C(\mathbf{n}_t)}{T_{q_A}^{(A)}(t)}\right)^{\frac{1}{1-q_A}} & \text{otherwise,} \end{cases}$$

where $T_{q_A}^{(A)}(t)$ is the acceptance temperature at iteration t.

The temperature follows a generalized cooling schedule:

$$T_{q_V}^{(V)}(t) = T_{q_V}(1) \frac{2^{q_V - 1} - 1}{(1+t)^{q_V - 1} - 1},$$

ensuring slow enough cooling to maintain ergodicity and eventual convergence to the global minimum.

Thus, at each iteration t, a candidate solution $\mathbf{n}_{t+1}$ is proposed by drawing a displacement $\Delta \mathbf{n}_t$ from the visiting distribution $g_{q_V}(\Delta \mathbf{n}_t)$ centered at the current solution $\mathbf{n}_t$. The candidate is accepted with probability $P_{q_A}(\mathbf{n}_t \to \mathbf{n}_{t+1})$, and the relevant temperatures are updated according to their respective cooling schedules. The process repeats until convergence criteria are met, typically when the energy stabilizes or the maximum number of iterations is reached.

For constrained problems, the objective may be modified as $C^*(\mathbf{n}) = C(\mathbf{n}) + a \cdot \text{Penalty}(\mathbf{n})$, where $a > 0$ controls the penalty strength.

Thus, GSAA offers a powerful and flexible global optimization method, especially suitable for complex landscapes and adaptable to both unconstrained and constrained scenarios.

In derivative-free optimization with nonlinear constraints, Powell (1994) introduced **Constrained Optimization by Linear Approximations (COBYLA)**. The method builds local linear models of the objective $C(\mathbf{n}) : \mathbb{R}^H \to \mathbb{R}$ and constraints $G_i(\mathbf{n}) : \mathbb{R}^H \to \mathbb{R}$, $i = 1, \ldots, r$, by interpolating their values at the vertices of a non-degenerate H-simplex. From

a simplex $\{x^{(0)}, x^{(1)}, \ldots, x^{(H)}\} \subset \mathbb{R}^H$ with full affine span, linear approximations $\tilde{C}(\mathbf{n})$ and $\tilde{G}_i(\mathbf{n})$ are constructed to match the true functions at each vertex.

At each iteration, COBYLA solves a linear subproblem

$$\begin{aligned} &\min_{\mathbf{n}\in\mathbb{R}^H} \tilde{C}(\mathbf{n}) \\ &\text{subject to} \quad \tilde{G}_i(\mathbf{n}) \geq 0, \quad i = 1, \ldots, r, \\ &\qquad\qquad \|\mathbf{n}_{t+1} - \mathbf{n}_t\|_2 \leq \Delta_t, \end{aligned}$$

where $\mathbf{n}_t$ is the current best vertex (minimizing a merit function) and $\Delta_t > 0$ is the trust-region radius. The radius is reduced if sufficient merit decrease is not achieved, regardless of simplex geometry.

The merit function used to compare candidate points is defined as

$$\Phi(\mathbf{n}) = C(\mathbf{n}) + \mu \cdot \max_{1\leq i\leq r} [-G_i(\mathbf{n})]_+ ,$$

with penalty parameter $\mu > 0$ dynamically updated to balance objective and constraint satisfaction. If the linearized subproblem is infeasible, COBYLA minimizes maximum constraint violation under the trust region. The simplex is then updated either by incorporating a new feasible point or by improving the interpolation geometry.

Although theoretical convergence guarantees are limited, COBYLA performs well in practice for low-dimensional problems without reliable derivatives, making it useful for black-box or noisy applications in engineering and science.

Consider the optimization of a real-valued objective function $C(\mathbf{n}) : \mathbb{R}^H \to \mathbb{R}$, defined over a bounded search domain $\Omega \subset \mathbb{R}^H$, where the goal is to find $\mathbf{n}^* \in \Omega$ such that $C(\mathbf{n}^*) = \min_{\mathbf{n}\in\Omega} C(\mathbf{n})$. To solve this, we use the **Particle Swarm Optimization Algorithm (PSOA)** (Kennedy and Eberhart 1995), which models a swarm of particles sharing positional information to locate the global minimum. Each particle i has a position $\mathbf{n}_{i,t} \in \Omega$, velocity $\mathbf{v}_{i,t}$, personal best position $\mathbf{p}_{i,t}$, and neighborhood best position $\mathbf{l}_{i,t}$. Velocities evolve as

$$\mathbf{v}_{i,t+1} = F(\mathbf{v}_{i,t}, \mathbf{p}_{i,t} - \mathbf{n}_{i,t}, \mathbf{I}_{i,t} - \mathbf{n}_{i,t}),$$

and positions update iteratively by

$$\mathbf{n}_{i,t+1} = \mathbf{n}_{i,t} + \mathbf{v}_{i,t+1},$$

with $C(\mathbf{n}_{i,t})$ guiding updates of $\mathbf{p}_{i,t}$ and $\mathbf{l}_{i,t}$.

To ensure convergence, Clerc and Kennedy (2002) introduced a constriction factor χ, yielding the standard PSOA update:

$$\mathbf{v}_{i,t+1} = \chi\left(\mathbf{v}_{i,t} + \alpha_1 \boldsymbol{\beta}_1 (\mathbf{p}_{i,t} - \mathbf{n}_{i,t}) + \alpha_2 \boldsymbol{\beta}_2 (\mathbf{l}_{i,t} - \mathbf{n}_{i,t})\right),$$

where α_1, α_2 are acceleration coefficients, and $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ are vectors of independent random samples drawn from the uniform distribution $U(0,1)$. Clerc (2012) later formalized Standard PSOA versions with reproducibility, rigorous topologies, and boundary handling.

Further developments include the phasor PSOA (PPSOA) (Ghasemi et al. 2018), which uses trigonometric phase-based updates for parameter-free adaptivity, and the multi-phase PSOA (Li et al. 2021), which segments optimization into phases with distinct strategies for improved performance.

For constrained optimization, modified PSOA methods incorporate strategies such as penalty functions, feasibility rules, and repair operators to enforce constraints while preserving swarm behavior (Rini et al. 2011).

PSOA has broad applicability. In statistical sampling, it optimizes stratum boundaries to minimize estimator variance under Neyman allocation (Al-Kassab and Ali 2015). In power systems, it is widely used for economic dispatch, optimal power flow, and reactive power control (del Valle et al. 2008). Numerous enhancements – such as inertia weight schedules, topology control, and hybridization with mutation operators – have been surveyed by Imran et al. (2013), underscoring the algorithm's adaptability and continued development.

In all cases, the central aim remains to iteratively improve candidate solutions $\mathbf{n}_{i,t}$ such that $C(\mathbf{n}_{i,t})$ approaches the global minimum, exploiting both individual and collective experience within the swarm framework.

In this study, we utilize the **Biased Random-Key Genetic Algorithm (BRKGA)** (Gonçalves and Resende 2011), an extension of the original Random-Key Genetic Algorithm (RKGA) proposed by Bean (1994), which can be applied to optimize a real-valued objective function $C(\mathbf{n}) : \mathbb{R}^H \to \mathbb{R}$. In the RKGA, candidate solutions are encoded as chromosomes – vectors of real numbers drawn from the interval $[0,1]$ – which are decoded into feasible solutions using a problem-specific mapping. This indirect encoding offers flexibility and is well-suited for combinatorial optimization problems.

The BRKGA, as applied by Brito et al. (2022), modifies the standard RKGA by introducing biased selection during crossover. In each generation, a population of N^* chromosomes is divided into an elite set (best-performing solutions), a non-elite set, and a set of mutant chromosomes randomly generated to preserve diversity. Crossover is performed between pairs where one parent is always selected from the elite set and the other from the non-elite set. A uniformly random auxiliary vector $\mathbf{v}_{aux} \in [0,1]^H$ and a predefined bias parameter $\delta_e > 0.5$ guide gene inheritance: if $v_{aux,i} \leq \delta_e$, the offspring gene at position i is inherited from the elite parent; otherwise, from the non-elite.

Each chromosome $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_H)$, where H is the number of strata, is decoded into a vector $\boldsymbol{\nu} = (n_1, \ldots, n_H)$ of sample sizes via a decoder. For the formulation that minimizes the total survey cost under precision constraints (Problem 1), assuming unit costs $c_h = 1$ for all strata, $h = 1, \ldots, H$, the decoding is given by:

$$n_h = n_{\min} + \operatorname{round}\left(\gamma_h \cdot (N_h - n_{\min})\right),$$

ensuring that $n_h \in [n_{\min}, N_h]$, $h = 1, \ldots, H$. For the formulation that minimizes a weighted sum of relative variances under a fixed total survey cost (Problem 2) – which reduces to the total sample size n when $c_h = 1$ for all $h = 1, \ldots, H$ – the decoding follows:

$$n_h = n_{\min} + \left((n - H n_{\min}) \cdot \frac{\gamma_h}{\sum_{k=1}^{H} \gamma_k} \right) \quad \text{for } h = 1, \ldots, H-1, \quad n_H = n - \sum_{h=1}^{H-1} n_h.$$

After decoding, each solution is evaluated using the objective function. To enforce feasibility, a penalty term is added if any constraint is violated. For instance, in Problem 1, the penalized objective becomes:

$$C_p = \sum_{h=1}^{H} n_h + P,$$

where $P = T^M$, $T \in \mathbb{R}$, if any $CV(\hat{t}_y^{(j)}) > CV_j$, and zero otherwise. Here, $M = \max_j \left\{ \frac{CV(\hat{t}_y^{(j)})}{CV_j} \right\}$.

This procedure ensures that only feasible or near-feasible solutions persist across generations. The BRKGA evolves the population by keeping elites, adding mutants, and producing biased offspring, balancing exploration and exploitation. As shown by Brito et al. (2022), it yields high-quality integer-feasible solutions under nonlinear constraints, rivaling exact integer programming methods.

Most approaches used to derive optimal sample sizes face challenges related to rounding, which can be particularly problematic in certain scenarios. These include: (1) surveys involving small areas, where adding or removing even a single unit from the sample can notably affect the variance estimates, and (2) surveys with a very high number of strata, where the total sample size n may differ considerably from the sum of the individually rounded sample sizes allocated to each stratum.

To address these issues, Brito et al. (2015) proposed **Integer Programming Algorithms (IPA)** to solve Problems 1 and 2, with the following additional constraint imposed: $n_h \in \mathbb{Z}_+$, for $h = 1, \ldots, H$. They used simple algebraic techniques to achieve linearity either in the objective function or in the constraints. Specifically, Brito et al. (2015) introduced a new binary variable z defined as:

$$z_{hk} = \begin{cases} 1, & \text{if the sample size } k \in \{n_{min}, \ldots, N_h\}, h = 1, \ldots, H, \text{ is allocated to stratum } U_h; \\ 0, & \text{otherwise.} \end{cases}$$

Through this variable, the second constraint in Problem 1 – which is originally nonlinear – can be reformulated as a linear expression in terms of the values of the binary variable z:

$$\sum_{h=1}^{H} N_h p_{hj} \sum_{k=n_{min}}^{N_h} \frac{z_{hk}}{k} - \sum_{h=1}^{H} p_{hj} \leq 1, \quad p_{hj} = \frac{N_h s_{hj}^2}{t_j^2 CV_j^2}, \quad j = 1, 2, \ldots, m.$$

Similarly, the nonlinear objective function in Problem 2 can be reformulated as a linear expression in terms of the binary variable z:

$$\sum_{j=1}^{m} w_j \frac{1}{t_j^2} \sum_{h=1}^{H} \left(\sum_{k=n_{\min}}^{N_h} \frac{z_{hk}}{k} \right) N_h^2 s_{hj}^2.$$

Although not shown here, the remaining constraints in Problems 1 and 2, as well as the linear objective function in Problem 1, are also reformulated in terms of the binary variable z, resulting in fully linear expressions.

After achieving linearity either in the objective function or in the constraints, Brito et al. (2015) solved the resulting integer programming problems using the Branch and Bound method (Wolsey 1998). This optimization approach guarantees attainment of the global minimum.

3. Numerical comparisons

This section presents the findings from a comparison of various multivariate optimal allocation techniques applied to a specific subset of population datasets. All computations were performed using the R programming language. The evaluation focuses on several algorithms employed to solve Problem 1, including Integer Programming (IPA) (Brito et al. 2015a), Bethel's Algorithm (BA) (Bethel 1985, 1989), the Generalized Simulated Annealing Algorithm (GSAA) (Tsallis and Stariolo 1996), Constrained Optimization by Linear Approximations (COBYLA) (Powell 1994), Particle Swarm Optimization (PSO) (Kennedy and Eberhart 1995), and the Biased Random Key Genetic Algorithm (BRKGA) (Gonçalves and Resende 2011; Brito et al. 2022). The IPA, GSAA, COBYLA, PSOA, and BRKGA algorithms are also applied to solve Problem 2, along with the textbook method given in Cochran (1977), which is denoted as TBA. Specifically, according to this method, the optimal sample size n_h from stratum $\mathscr{U}_h$ is calculated using the following formula:

$$n_h = n\frac{\sqrt{\sum_{j=1}^{m}\left(n_{hj}^{(N)}\right)^2}}{\sum_{h=1}^{H}\sqrt{\sum_{j=1}^{m}\left(n_{hj}^{(N)}\right)^2}},$$

where $n_{hj}^{(N)}$ denotes the optimum sample size in stratum $\mathscr{U}_h$ for variable j, calculated according to the Neyman (1934) allocation.

Initially, the comparisons are performed on two synthetic populations, each containing $N = 10000$ units. In each population, four study variables are specified. To establish a predetermined dependence structure among them, a Gaussian copula is first constructed. This copula serves as a probability distribution where each of the four random variables has a uniform marginal distribution. Next, these uniformly distributed variables are converted into the target distributions by applying the inverse transform method.

Thus, for Population 1, the variables are simulated from asymmetric distributions: $y^{(1)} \sim \mathscr{E}(0.005)$, $y^{(2)} = |y|$, where $y \sim t(3)$, $y^{(3)} \sim \Gamma(1,2)$, $y^{(4)} \sim \chi^2(2)$, with $\rho(y^{(1)},y^{(2)}) = 0.13$, $\rho(y^{(1)},y^{(3)}) = 0.39$, $\rho(y^{(1)},y^{(4)}) = -0.31$, $\rho(y^{(2)},y^{(3)}) = 0.12$, $\rho(y^{(2)},y^{(4)}) = 0.13$, $\rho(y^{(3)},y^{(4)}) = 0.30$.

Population 2 consists of study variables following a combination of normal, exponential, and Fisher distributions: $y^{(1)} \sim \mathscr{E}(0.005)$, $y^{(2)} \sim \mathscr{N}(3000,300)$, $y^{(3)} \sim \mathscr{F}(5,4)$, and $y^{(4)} \sim \mathscr{N}(100,20)$. Here, the second parameter in the normal distributions represents the standard deviation. The correlations between these variables are given as follows: $\rho(y^{(1)},y^{(2)}) = 0.19$, $\rho(y^{(1)},y^{(3)}) = 0.13$, $\rho(y^{(1)},y^{(4)}) = 0.27$, $\rho(y^{(2)},y^{(3)}) = 0.12$, $\rho(y^{(2)},y^{(4)}) = 0.20$, $\rho(y^{(3)},y^{(4)}) = 0.17$.

For extended analysis, Population 3 is introduced, originating from a statistical survey on the area and yield of agricultural plants in Lithuanian agricultural companies and en-

terprises, with a total population size of $N = 6\,204$. To assess different sample allocation methods, the following four skewed variables are selected: $y^{(1)}$ - total yield of cereals and oilseed rape, $y^{(2)}$ - total yield of cereals and oilseed rape after cleaning and drying, $y^{(3)}$ - total area of cereals and oilseed rape, and $y^{(4)}$ - total harvested area of cereals and oilseed rape. The relationships between these variables are characterized by the following correlation coefficients: $\rho(y^{(1)}, y^{(2)}) = 0.64$, $\rho(y^{(1)}, y^{(3)}) = 0.66$, $\rho(y^{(1)}, y^{(4)}) = 0.71$, $\rho(y^{(2)}, y^{(3)}) = 0.86$, $\rho(y^{(2)}, y^{(4)}) = 0.87$, $\rho(y^{(3)}, y^{(4)}) = 0.93$.

All populations are stratified using the traditional k-means approach implemented in the `stats` package in R (R Core Team 2023). The number of strata, H, along with their respective sizes, $N_1, N_2, \ldots, N_H$, are detailed in Table 1.

Table 1. Population strata sizes and number of strata

Population	Number of strata (H)	Strata sizes ($N_1, N_2, \ldots, N_H$)
1	10	1024, 531, 501, 1253, 1761, 649, 1616, 1211, 731, 723
2	7	1083, 831, 245, 1397, 2719, 1123, 2602
3	6	680, 438, 557, 695, 2596, 1238

For the numerical analysis conducted in each population, unit survey costs are assumed to be the same across all strata, and the weights w_j, for $j = 1, 2, \ldots, m$, are considered equal for all survey variables. A minimum sample size per stratum of $n_{\min} = 2$ is maintained for all methods and populations examined. The predefined coefficients of variation for the total estimators in Problem 1 are set at 5%, 10%, and 15%. In Problem 2, the total sample sizes for allocation are determined based on sampling fractions of 5%, 10%, and 20% of the respective population sizes (N).

The algorithms are implemented using their respective R packages, with specific parameter configurations as detailed below. The Integer Programming Algorithms (IPA) employ the functions `BSSM_FC()` and `BSSM_FD()` from the `MultAlloc` package (Brito et al. 2015b) to solve Problem 1 and Problem 2, respectively. Bethel's Algorithm (BA) is executed via the `bethel()` function from the `SamplingStrata` package (Barcaroli 2014), where precision constraints are defined in terms of coefficients of variation for each studied variable. The textbook method (TBA) is developed by us using the R programming language. For the generalized simulated annealing algorithm (GSAA), the `GenSA()` function from the `GenSA` package (Xiang et al. 2013) is used, with the parameters set as follows: `temperature = 1000`, `parameter for visiting distribution = 2.63`, and `parameter for acceptance distribution = -12`. Constrained Optimization by Linear Approximations (COBYLA) employs the `nloptr()` function from the `nloptr` package, with the `algorithm` parameter assigned the value `NLOPT_LN_COBYLA`. Particle Swarm Optimization (PSO) is carried out using the `psoptim()` function from the `pso` package (Bendtsen 2022), where the `swarm size` is configured as 300. By default, this algorithm adheres to the Standard PSOA 2007 framework established by Clerc (2012). Lastly, the Biased Random Key Genetic Algorithm (BRKGA) is executed using the `brkga()` function from the `BRKGA` package (Brito et al. 2023), with the following parameter settings: `size of the algorithm population = 2000`, `percentage of elite chromosomes = 0.2`, `percentage of mutant chromosomes = 0.2`, `crossover probability = 0.6`,

`number of generations = 2 000`, and `penalty factor = 1 000`. Any other hyperparameters that are not explicitly stated remain at their default values. The R code defining the objective function and constraints for GSAA, COBYLA, PSOA, and BRKGA is created externally from their respective function environments.

Table 2. Comparison of Algorithms for Population 1 and Problem 1

Alg.:	IPA	BA	GSAA	CBLA	PSOA	BRKGA
$CV_j = 5\%$						
$\sum n_i$	**457**	462	458	460	460	**457**
$f(\text{in }\%)$	4.570	4.620	4.580	4.600	4.600	4.570
$CV(\hat{t}_y^{(1)})$	3.408	3.386	3.389	3.389	3.418	3.401
$CV(\hat{t}_y^{(2)})$	4.996	4.967	4.999	4.978	4.997	4.999
$CV(\hat{t}_y^{(3)})$	4.550	4.522	4.519	4.537	4.582	4.581
$CV(\hat{t}_y^{(4)})$	4.485	4.459	4.479	4.472	4.504	4.492
$CV_j = 10\%$						
$\sum n_i$	**119**	124	**119**	**119**	120	120
$f(\text{in }\%)$	1.190	1.240	1.190	1.190	1.200	1.200
$CV(\hat{t}_y^{(1)})$	6.809	6.695	6.778	6.809	6.985	6.702

Alg.:	IPA	BA	GSAA	CBLA	PSOA	BRKGA
$CV(\hat{t}_y^{(2)})$	9.983	9.777	9.999	9.983	9.998	9.980
$CV(\hat{t}_y^{(3)})$	9.047	8.896	9.049	9.047	9.185	8.953
$CV(\hat{t}_y^{(4)})$	8.948	8.750	8.948	8.948	8.921	8.959
$CV_j = 15\%$						
$\sum n_i$	**54**	59	**54**	**54**	**54**	**54**
$f(\text{in }\%)$	0.540	0.590	0.540	0.540	0.540	0.540
$CV(\hat{t}_y^{(1)})$	10.313	9.786	10.214	10.306	10.214	10.070
$CV(\hat{t}_y^{(2)})$	14.937	14.269	14.974	14.889	14.968	14.987
$CV(\hat{t}_y^{(3)})$	13.636	12.831	13.520	13.576	13.432	13.495
$CV(\hat{t}_y^{(4)})$	13.418	12.750	13.388	13.276	13.353	13.447

Table 3. Comparison of Algorithms for Population 2 and Problem 1

Alg.:	IPA	BA	GSAA	CBLA	PSOA	BRKGA
$CV_j = 5\%$						
$\sum n_i$	**773**	776	**773**	774	774	**773**
$f(\text{in }\%)$	7.730	7.760	7.730	7.740	7.740	7.730
$CV(\hat{t}_y^{(1)})$	2.497	2.473	2.459	2.495	2.455	2.478
$CV(\hat{t}_y^{(2)})$	0.192	0.190	0.189	0.192	0.189	0.190
$CV(\hat{t}_y^{(3)})$	4.998	4.982	4.999	4.992	4.995	5.000
$CV(\hat{t}_y^{(4)})$	1.002	0.994	0.995	1.001	0.983	0.995
$CV_j = 10\%$						
$\sum n_i$	**305**	308	**305**	306	306	**305**
$f(\text{in }\%)$	3.050	3.080	3.050	3.060	3.060	3.050
$CV(\hat{t}_y^{(1)})$	4.266	4.140	4.166	4.264	4.144	4.127

Alg.:	IPA	BA	GSAA	CBLA	PSOA	BRKGA
$CV(\hat{t}_y^{(2)})$	0.328	0.318	0.319	0.328	0.318	0.317
$CV(\hat{t}_y^{(3)})$	9.988	9.928	9.999	9.967	9.997	9.999
$CV(\hat{t}_y^{(4)})$	1.710	1.665	1.681	1.709	1.648	1.655
$CV_j = 15\%$						
$\sum n_i$	**153**	156	**153**	155	**153**	154
$f(\text{in }\%)$	1.530	1.560	1.530	1.550	1.530	1.540
$CV(\hat{t}_y^{(1)})$	6.042	5.714	5.773	6.038	5.903	5.702
$CV(\hat{t}_y^{(2)})$	0.465	0.438	0.442	0.464	0.451	0.436
$CV(\hat{t}_y^{(3)})$	14.976	14.823	14.993	14.868	14.998	14.993
$CV(\hat{t}_y^{(4)})$	2.419	2.301	2.347	2.418	2.402	2.303

Table 4. Comparison of Algorithms for Population 3 and Problem 1

Alg.:	IPA	BA	GSAA	CBLA	PSOA	BRKGA
$CV_j = 5\%$						
$\sum n_i$	**745**	747	**745**	747	**745**	**745**
$f(\text{in }\%)$	12.008	12.041	12.008	12.041	12.008	12.008
$CV(\hat{t}_y^{(1)})$	4.996	4.987	4.999	4.987	4.998	5.000
$CV(\hat{t}_y^{(2)})$	3.758	3.748	3.771	3.754	3.817	3.735
$CV(\hat{t}_y^{(3)})$	4.242	4.238	4.293	4.240	4.400	4.077
$CV(\hat{t}_y^{(4)})$	3.521	3.516	3.579	3.519	3.649	3.427
$CV_j = 10\%$						
$\sum n_i$	**229**	231	**229**	**229**	**229**	**229**
$f(\text{in }\%)$	3.691	3.723	3.691	3.691	3.691	3.691
$CV(\hat{t}_y^{(1)})$	9.984	9.936	10.000	9.976	9.995	9.982

Alg.:	IPA	BA	GSAA	CBLA	PSOA	BRKGA
$CV(\hat{t}_y^{(2)})$	7.390	7.271	7.337	7.407	7.587	7.309
$CV(\hat{t}_y^{(3)})$	7.455	7.592	7.605	7.821	8.443	7.790
$CV(\hat{t}_y^{(4)})$	6.421	6.351	6.376	6.618	6.997	6.575
$CV_j = 15\%$						
$\sum n_i$	**107**	110	**107**	**107**	**107**	**107**
$f(\text{in }\%)$	1.725	1.773	1.725	1.725	1.725	1.725
$CV(\hat{t}_y^{(1)})$	14.953	14.735	14.989	14.953	14.982	14.992
$CV(\hat{t}_y^{(2)})$	10.795	10.649	10.825	10.795	10.848	10.824
$CV(\hat{t}_y^{(3)})$	10.858	10.797	10.908	10.858	10.874	10.001
$CV(\hat{t}_y^{(4)})$	9.100	9.028	9.130	9.100	9.121	8.656

Table 5. Comparison of Algorithms for Population 1 and Problem 2

Alg.:	IPA	TBA	GSAA	CBLA	PSOA	BRKGA
$n = 500$, $f = 5\%$						
$CV(\hat{t}_y^{(1)})$	3.228	3.603	3.247	3.238	3.205	3.238
$CV(\hat{t}_y^{(2)})$	4.787	5.281	4.788	4.789	4.806	4.789
$CV(\hat{t}_y^{(3)})$	4.255	4.581	4.272	4.254	4.292	4.254
$CV(\hat{t}_y^{(4)})$	4.275	4.532	4.250	4.266	4.318	4.266
$\sum CV(\hat{t}_y^{(i)})$	**16.545**	17.997	16.557	16.547	16.621	16.547
$n = 1000$, $f = 10\%$						
$CV(\hat{t}_y^{(1)})$	2.206	2.479	2.212	2.214	2.211	2.209
$CV(\hat{t}_y^{(2)})$	3.287	3.635	3.290	3.290	3.287	3.287
$CV(\hat{t}_y^{(3)})$	2.933	3.164	2.944	2.933	2.929	2.929
$CV(\hat{t}_y^{(4)})$	2.940	3.122	2.931	2.942	2.940	2.941
$\sum CV(\hat{t}_y^{(i)})$	**11.366**	12.400	11.377	11.379	11.367	**11.366**
$n = 2000$, $f = 20\%$						
$CV(\hat{t}_y^{(1)})$	1.455	1.654	1.465	1.453	1.453	1.453
$CV(\hat{t}_y^{(2)})$	2.179	2.437	2.184	2.177	2.177	2.178
$CV(\hat{t}_y^{(3)})$	1.950	2.124	1.970	1.954	1.954	1.953
$CV(\hat{t}_y^{(4)})$	1.961	2.096	1.957	1.961	1.961	1.961
$\sum CV(\hat{t}_y^{(i)})$	**7.545**	8.311	7.576	**7.545**	**7.545**	**7.545**

Table 6. Comparison of Algorithms for Population 2 and Problem 2

Alg.:	IPA	TBA	GSAA	CBLA	PSOA	BRKGA
$n = 500$, $f = 5\%$						
$CV(\hat{t}_y^{(1)})$	2.919	2.362	2.926	2.726	2.920	2.914
$CV(\hat{t}_y^{(2)})$	0.222	0.187	0.223	0.208	0.223	0.222
$CV(\hat{t}_y^{(3)})$	7.199	21.043	7.198	7.499	7.198	7.201
$CV(\hat{t}_y^{(4)})$	1.189	0.859	1.190	1.113	1.189	1.194
$\sum CV(\hat{t}_y^{(i)})$	**11.529**	24.451	11.537	11.546	11.530	11.531
$n = 1000$, $f = 10\%$						
$CV(\hat{t}_y^{(1)})$	1.847	1.627	1.859	1.854	1.854	1.847
$CV(\hat{t}_y^{(2)})$	0.140	0.129	0.142	0.141	0.141	0.141
$CV(\hat{t}_y^{(3)})$	4.142	14.438	4.136	4.137	4.137	4.141
$CV(\hat{t}_y^{(4)})$	0.755	0.591	0.759	0.758	0.758	0.755
$\sum CV(\hat{t}_y^{(i)})$	**6.884**	16.785	6.896	6.890	6.890	**6.884**
$n = 2000$, $f = 20\%$						
$CV(\hat{t}_y^{(1)})$	1.155	1.093	1.184	1.162	1.162	1.161
$CV(\hat{t}_y^{(2)})$	0.087	0.087	0.090	0.088	0.088	0.088
$CV(\hat{t}_y^{(3)})$	2.463	9.475	2.453	2.458	2.458	2.458
$CV(\hat{t}_y^{(4)})$	0.472	0.396	0.480	0.476	0.476	0.475
$\sum CV(\hat{t}_y^{(i)})$	**4.177**	11.051	4.207	4.184	4.184	4.182

Table 7. Comparison of Algorithms for Population 3 and Problem 2

Alg.:	IPA	TBA	GSAA	CBLA	PSOA	BRKGA
$n = 310$, $f = 5\%$						
$CV(\hat{t}_y^{(1)})$	8.825	9.895	8.881	8.844	8.825	8.825
$CV(\hat{t}_y^{(2)})$	5.843	7.218	5.852	5.827	5.843	5.843
$CV(\hat{t}_y^{(3)})$	4.963	5.225	4.937	4.984	4.963	4.963
$CV(\hat{t}_y^{(4)})$	4.392	4.966	4.359	4.409	4.392	4.392
$\sum CV(\hat{t}_y^{(i)})$	**24.023**	27.304	24.029	24.064	**24.023**	**24.023**
$n = 620$, $f = 10\%$						
$CV(\hat{t}_y^{(1)})$	5.906	6.697	5.905	5.940	5.930	5.904
$CV(\hat{t}_y^{(2)})$	3.888	4.896	3.884	3.948	3.901	3.889
$CV(\hat{t}_y^{(3)})$	3.376	3.556	3.382	3.363	3.359	3.377
$CV(\hat{t}_y^{(4)})$	2.970	3.379	2.976	2.976	2.950	2.971
$\sum CV(\hat{t}_y^{(i)})$	**16.140**	18.528	16.147	16.227	**16.140**	16.141
$n = 1241$, $f = 20\%$						
$CV(\hat{t}_y^{(1)})$	3.657	4.292	3.638	3.657	3.657	3.657
$CV(\hat{t}_y^{(2)})$	2.377	3.182	2.392	2.378	2.378	2.378
$CV(\hat{t}_y^{(3)})$	2.177	2.323	2.189	2.177	2.177	2.177
$CV(\hat{t}_y^{(4)})$	1.886	2.209	1.895	1.886	1.886	1.886
$\sum CV(\hat{t}_y^{(i)})$	**10.097**	12.006	10.114	10.098	10.098	10.098

Tables 2-4 present the sample allocation resulting from the solution to Problem 1, along with the total sample size, sampling fraction f, and the achieved coefficients of variation for the estimators of all variables across the populations, as well as the pre-specified coefficients of variation. Note that COBYLA is abbreviated as CBLA in the tables to ensure a better fit within the table format.

Among all the methods examined, only the Integer Programming Algorithm (IPA) attains the global minimum. Thus, other methods can be compared to IPA's results to evaluate deviations from the minimum objective value. Bolded values in the tables indicate the

global optimum, making it easier to visually compare the results of IPA with those of the other algorithms.

As shown in the tables, the total sample size n produced by Bethel's Algorithm (BA) is never smaller than that obtained by the other methods. The only case where BA matches another method is in Population 3 for $CV_j = 5\%$, where it yields the same total sample size as COBYLA. The greatest deviations from the global minimum for BA are observed in the skewed Population 1. Notably, BA does not achieve the global minimum in any of the analyzed populations.

Tables 2-4 also show that the algorithms GSAA, COBYLA, PSOA, and BRKGA achieve the global minimum in 88.89%, 44.44%, 55.56%, and 77.78% of the cases, respectively. In the remaining cases, the differences from the global minimum for these methods are not substantial. The lowest percentage of global minimum attainment, when GSAA, COBYLA, PSOA, and BRKGA are considered together, is observed in Population 2. In real Population 3, the methods discussed in this paragraph reach the global minimum in nearly all cases. The proportion of cases in which the global minimum is reached in Population 1 is similar to that in Population 2. We further observe that even if an algorithm finds the global minimum, the specific solution it yields may differ from that produced by IPA. As an example, in Population 2 where $CV_j = 10\%$, IPA produces the values $n_1, n_2, \ldots, n_H$ as 24, 42, 118, 8, 43, 48, and 22. Although GSAA also reaches the global minimum, it results in a different allocation: 24, 46, 113, 9, 42, 49, and 22. It is also worth noting that a method capable of achieving the global minimum for a fixed population and a particular value of CV_j may not consistently reach the global minimum when CV_j varies. As an illustration, in Population 2, PSOA successfully attains the global minimum at $CV_j = 15\%$, yet it does not achieve this outcome when CV_j is set to 5% or 10%.

The comparison of artificial Populations 1 and 2 indicates that both stratum and total sample sizes are sensitive to the distributional characteristics of the study variables. In Population 2, despite two variables being normally distributed, the presence of two skewed variables – particularly the one following a heavily skewed Fisher distribution – leads to a notable increase in the total sample size. Compared to other populations, real Population 3 demonstrates the most pronounced increase in sampling fraction for every pre-specified coefficient of variation.

Tables 5–7 present the comparative results of six algorithms – IPA, TBA, GSAA, COBYLA, PSOA, and BRKGA – for the previously considered populations under Problem 2. The objective in this setting is to minimize the weighted relative variance of the Horvitz-Thompson estimators of totals across multiple survey variables, given fixed overall sample sizes corresponding to sampling fractions of 5%, 10%, and 20%. Each table presents the sample allocations across strata, the corresponding coefficients of variation for each survey variable, and the total sum of coefficients of variation, $\sum CV(\hat{t}_y^{(i)})$, which serves as a summary measure of overall efficiency across all survey variables. The global minimum of this total is indicated in bold.

In these experiments, IPA is again used as a benchmark, as it achieves the global minimum in every case. Other methods are evaluated against IPA's performance in terms of both efficiency and allocation stability.

Thus, IPA serves as the reference algorithm, consistently achieving the lowest possible value of $\sum CV(\hat{t}_y^{(i)})$ across all populations and for each fixed total sample size. Its allocations are well-balanced and establish the best-case baseline against which all other methods are compared. BRKGA closely matches IPA in all settings, often reaching the same total *CV* values and producing well-balanced and robust allocations. It proves to be a competitive and stable alternative to IPA. COBYLA performs slightly worse than BRKGA, sometimes matching IPA and BRKGA in total *CV*. It provides consistent and efficient allocations, especially in Populations 1 and 3, making it quite a reliable method in practice. GSAA delivers results similar to COBYLA and BRKGA in some instances but shows occasional variability in allocation that slightly affects performance. PSOA performs moderately well across all populations, with results typically falling slightly above the optimal $\sum CV(\hat{t}_y^{(i)})$ values. Its allocations are generally balanced, although not as consistently efficient as IPA or BRKGA. TBA, while effective in select settings, often exhibits unstable behavior, particularly in Population 2. It tends to heavily over- or under-sample certain strata, which leads to significantly inflated coefficients of variation for some estimators (e.g. estimator $\hat{t}_y^{(3)}$). These outliers frequently result in high total *CV* values, thereby undermining the method's overall reliability.

4. Conclusions

The study finds that the Integer Programming Algorithm (IPA) is the most robust and accurate method for multivariate optimal allocation in stratified sampling. As an exact approach, it guarantees the global minimum, directly handles integer constraints, and avoids rounding errors – crucial for many strata or small-area surveys.

Bethel's Algorithm (BA), while widely used, consistently underperforms. It never reaches the global minimum and often yields the largest sample sizes, particularly struggling in populations with skewed distributions. Despite typically converging, Bethel's Algorithm often produces suboptimal results compared to more advanced methods.

Among stochastic approaches, the Biased Random-Key Genetic Algorithm (BRKGA) proves to be the most competitive. It frequently finds solutions that match or closely approximate IPA's and consistently delivers stable, well-balanced allocations. Its efficiency and adaptability make it a strong practical alternative, especially in scenarios where flexibility or faster approximate solutions are preferred over exact methods. The Generalized Simulated Annealing Algorithm (GSAA) also performs well, often reaching the global minimum, although with slightly more variability. COBYLA is reliable, especially with real-world data in Problem 1, although less consistent in achieving optimality in Problem 2. Particle Swarm Optimization (PSOA) offers reasonable results but tends to be more variable and less efficient under tighter precision constraints in certain cases.

The textbook method (TBA) is unstable in variance minimization, often yielding poor allocation balance and high coefficients of variation in some strata and variables.

The study shows that variable distribution affects allocation: skewed variables require larger samples in Problem 1 and yield higher variation in Problem 2, stressing the need for adaptive, precise methods.

In the real population case (Population 3), under Problem 1, for each predefined precision level ($CV_j = 5\%, 10\%$, and 15%), the sampling fraction is consistently higher than in the artificial populations, highlighting the greater complexity and variability inherent in real-world data. In Problem 2, which aims to minimize overall variance given a fixed total sample size, the highest coefficients of variation across all survey variables are also observed in the real data case.

This increase – whether in the sampling fraction under Problem 1 or in the coefficients of variation under Problem 2 – is primarily driven by the skewed distribution of the study variables, which necessitates larger samples to meet precision requirements in Problem 1 and leads to higher variances in Problem 2.

This study offers a comparative analysis of exact and approximate optimization methods. By benchmarking against exact integer programming and testing on diverse real and synthetic populations, it stands as one of the most comprehensive empirical studies of multivariate stratified sampling allocation. These results not only affirm the strengths and limitations of different algorithm classes but also provide actionable guidance for practitioners and survey designers facing real-world complexity in cost and variance trade-offs.

Acknowledgment

I sincerely thank Vilma Nekrašaitė-Liegė (VILNIUS TECH) for sharing the real-world dataset used in this study.

References

Ahsan, M. J., Khan, S. U., (1982). Optimum allocation in multivariate stratified random sampling with overhead cost. *Metrika*, 29, pp. 71–78. Available from: https://doi.org/10.1007/BF01893366.

AL-Kassab, M. M., Ali, A. A., (2015). Using particle swarm optimization to determine the optimal strata boundaries. *J. Adv. Math.*, 11(1). Available from: https://rajpub.com/index.php/jam/article/view/1290.

Barcaroli, G., (2014). SamplingStrata: An R package for the optimization of stratified sampling. *J. Stat. Softw.*, 61(4), pp. 1–24. Available from: https://doi.org/10.18637/jss.v061.i04.

Bean, J. C., (1994). Genetic algorithms and random keys for sequencing and optimization. *ORSA J. Comput.*, 6(2), pp. 154–160. Available from: https://doi.org/10.1287/ijoc.6.2.154.

Bendtsen, C., (2022). *pso: Particle Swarm Optimization*. Available from: https://cran.r-project.org/web/packages/pso/index.html. R package version 1.0.4.

Bethel, J., (1985). An optimum allocation algorithm for multivariate surveys. *Proc. Surv. Res. Methods Sect.*, pp. 209–212. Available from: http://www.asasrms.org/Proceedings/papers/$1985_0$35.pdf.

Bethel, J., (1989). Sample allocation in multivariate surveys. *Surv. Methodol.*, 15(1), pp. 47–57. Available from: https://www.istat.it/en/files/2016/10/Sample-Allocation-in-Multivariate-Surveys.pdf.

Brito, J. A., do Nascimento Silva, P. L., Semaan, G. S. and Maculan, N., (2015a). Integer programming formulations applied to optimal allocation in stratified sampling. *Surv. Methodol.*, 41(2), pp. 427–442. Available from: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2015002/article/14249-eng.pdf?st=P7ZqwcD1.

Brito, J. A., do Nascimento Silva, P. L., Maculan, N. and Semaan, G. S., (2015b). *MultAlloc: Optimal Allocation in Stratified Sampling*. R package version 1.2.

Brito, J. A., Fadel, A. and Semaan, G. S., (2022). A genetic algorithm applied to optimal allocation in stratified sampling. *Commun. Stat. Simul. Comput.*, 51(7), pp. 3714–3732. Available from: https://doi.org/10.1080/03610918.2020.1722832.

Brito, J. A., Semaan, G. S. and Fadel, A., (2023). *BRKGA: Biased Random Key Genetic Algorithm for Optimization Problems*. R package version 0.1.0.

Chatterjee, S., (1967). A note on optimum allocation. *Scand. Actuar. J.*, 50, pp. 40–44. Available from: https://doi.org/10.1080/03461238.1967.10406206.

Clerc, M., (2012). *Standard particle swarm optimisation*, Preprint on HAL Open Archive. Available from: https://hal.science/hal-00764996v1.

Clerc, M., Kennedy, J., (2002). The particle swarm – explosion, stability, and convergence in a multidimensional complex space. *IEEE Trans. Evol. Comput.*, 6(1), pp. 58–73. Available from: https://doi.org/10.1109/4235.985692.

Cochran, W. G., (1977). *Sampling techniques*, 3rd ed. New York: Wiley. Available from: https://books.google.lt/books?id=xbNn41DUrNwC.

Dayal, S., (1985). Allocation of sample using values of auxiliary characteristic. *J. Stat. Plan. Inference*, 11(3), pp. 321–328.

del Valle, Y., Venayagamoorthy, G. K., Mohagheghi, S., Hernandez, J. C. and Harley, R. G., (2008). Particle swarm optimization: Basic concepts, variants and applications in power systems. *IEEE Trans. Evol. Comput.*, 12(2). Available from:https://doi.org/10.1109/TEVC.2007.896686.

Garciá, J. A. D., Cortez, L. U., (2006). Optimum allocation in multivariate stratified sampling: multi-objective programming. *Comunic. Del Cimat*, no I-06-07/28-03-2006. Available from: https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/656/1/I-06-07.pdf.

Ghasemi, M., Akbari, E., Rahimnejad, A., Razavi, S. E., Ghavidel, S. and Li, L., (2018). Phasor particle swarm optimization: a simple and efficient variant of pso. *Soft Comput.*, 23, pp. 9701–9718. Available from: https://doi.org/10.1007/s00500-018-3536-8.

Gonçalves, J. F., Resende, M. G. C., (2011). Biased random-key genetic algorithms for combinatorial optimization. *J. Heuristics*, 17(5), pp. 487–525. Available from: https://doi.org/10.1007/s10732-010-9143-1.

Gupta, S., Haq, A. and Varshney, R., (2024). Problem of compromise allocation in multivariate stratified sampling using intuitionistic fuzzy programming. *Ann. Data Sci.*, 11, pp. 425–444. Available from: https://doi.org/10.1007/s40745-022-00410-y.

Haq, A., Ali, I. and Varshney, R., (2020). Compromise allocation problem in multivariate stratified sampling with flexible fuzzy goals. *J. Stat. Comput. Simul.*, 90(9), pp. 1557–1569. Available from: https://doi.org/10.1080/00949655.2020.1734808.

Horvitz, D. G., Thompson, D. J., (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, 47, pp. 663–685. Available from: https://doi.org/10.1080/01621459.1952.10483446.

Imran, M., Hashim, R. and Abd Khalid, N. E., (2013). An overview of particle swarm optimization variants. *Procedia Eng.* , 53, pp. 491–496. Available from: https://doi.org/10.1016/j.proeng.2013.02.063.

Jalil, S. A., Haq, A., Owad, A. A., Hashmi, N. and Adichwal, N. K., (2023). A hierarchical multi-level model for compromise allocation in multivariate stratified sample surveys with non-response problem, *Knowl.-Based Syst.*, 278. Available from: https://doi.org/10.1016/j.knosys.2023.110839.

Kadane, J. B., (2005). Optimal dynamic sample allocation among strata. *J. Off. Stat.*, 21(4), pp. 531–541. Available from: https://doi.org/10.1184/R1/6586808.v1.

Kennedy, J., Eberhart, R., (1995). Particle swarm optimization. *Proc. ICNN'95 – Int. Conf. Neural Netw.*, pp. 1942–1948. Available from: https://doi.org/10.1109/ICNN.1995.488968.

Khan, M. F., Ali, I. and Ahmad, Q. S., (2011). Chebyshev approximate solution to allocation problem in multiple objective surveys with random costs. *Am. J. Comput. Math.*, 01(04), pp. 247–251. Available from: https://doi.org/10.4236/ajcm.2011.14029.

Khan, M. G. M. and Ahsan, M. J., (2003). A note on optimum allocation in multivariate stratified sampling. *South Pac. J. Nat. Appl. Sci.*, 21(1), pp. 91–95. Available from: https://doi.org/10.1071/SP03017.

Khan, M. G. M., Ahsan, M. J. and Jahan, N., (1998). Compromise allocation in multivariate stratified sampling: An integer solution. *Nav. Res. Logist.*, 44(1). Available from: https://doi.org/10.1002/(SICI)1520-6750(199702)44:1<69::AID-NAV4>3.0.CO;2-K.

Khan, M. G. M., Maiti, T. and Ahsan, M. J., (2010). An optimal multivariate stratified sampling design using auxiliary information: an integer solution using goal programming approach. *J. Off. Stat.*, 26, pp. 695–708. Available from: https://www.semanticscholar.org/paper/An-optimal-multivariate-stratified-sampling-design-Khan-Maiti/a8cfea23255468fd838e09ef09b2f9976f984985.

Kish, L., (1976). Optima and proxima in linear sample designs. *J. R. Stat. Soc. Ser. A*, 139(1), pp. 80–95. Available from: https://doi.org/10.2307/2344384.

Kokan, A. R., Khan, S., (1967). Optimum allocation in multivariate surveys : An analytical solution. *J. R. Stat. Soc. Ser. B (Methodol.)*, 29(1), pp. 115–125. Available from: https://doi.org/10.1111/j.2517-6161.1967.tb00679.x.

Kuhn, H. W., Tucker, A. W., (1951). Nonlinear programming. *Proc. 2nd Berkeley Symp. Math. Stat. Prob.*, pp. 481 – 492.

Li, J., Sun, Y. and Hou, S., (2021). Particle swarm optimization algorithm with multiple phases for solving continuous optimization problems. *Discret. Dyn. Nat. Soc.*. Available from: https://doi.org/10.1155/2021/8378579.

Mahfouz, M. I., Rashwan, M. M. and Khadr, Z. A., (2023). Optimal Stochastic Allocation in Multivariate Stratified Sampling. *Math. Stat.*, 11(4), pp. 676–684. Available from: https://doi.org/10.13189/ms.2023.110409.

Neyman, J., (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection (with discussion). *J. R. Stat. Soc.*, 97, pp. 558–625. Available from: https://doi.org/10.2307/2342192.

Powell, M. J. D., (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation. In Gomez, S. and Hennart J. P. (Eds.), *Advances in Optimization and Numerical Analysis*, pp. 51–67, Kluwer Academic, Dordrecht. Available from: https://doi.org/10.1007/978-94-015-8330-5.

R Core Team, (2023). *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, Vienna, Austria. Available from: https://www.R-project.org/.

Raghav, Y. S., Haq, A. and Ali I., (2023). Multiobjective intuitionistic fuzzy programming under pessimistic and optimistic applications in multivariate stratified sample allocation problems. *PLoS ONE*, 18(4). Available from: https://doi.org/10.1371/journal.pone.0284784.

Reddy, K. G., Khan, M. G. M. and Khan, S., (2018). Optimum strata boundaries and sample sizes in health surveys using auxiliary variables. *PLoS ONE*, 13(4).

Rini, D. P., Shamsuddin, S. M. and Yuhaniz, S. S., (2011). Particle swarm optimization: Technique, system and challenges. *Int. J. Comput. Appl.*, 14(1). Available from: https://doi.org/10.5120/1810-2331.

Shannon, C. E., (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, 27(3), pp. 379–423. Available from: https://people.math.harvard.edu/ ctm/home/text/ others/shannon/entropy/entropy.pdf.

Swain, A. K., (2013). A note on optimum allocation in stratified random sampling. *Invest. Oper.*, 34(2).

Tsallis, C., Stariolo, D. A., (1996). Generalized simulated annealing. *Physica A*, 233(1), pp 395–406. Available from: https://doi.org/10.1016/S0378-4371(96)00271-3.

Varshney, R., Khan, M. G. M., Fatima, U. and Ahsan, M. J., (2014). Integer compromise allocation in multivariate stratified surveys. *Ann. Oper. Res.*, 226(1), pp. 659–668. Available from: https://doi.org/10.1007/s10479-014-1734-z.

Wesołowski, J., Wieczorkowski, R. and Wójciak, W., (2024). Recursive Neyman algorithm for optimum sample allocation under box constraints on sample sizes in strata. *Surv. Methodol.*, 50(2). Available from: https://www150.statcan.gc.ca/n1/pub/12-001-x/2024002/article/00003-eng.pdf.

Wolsey, L. A., (1998). *Integer Programming*, John Wiley & Sons, New York. Available from: https://books.google.lt/books/about/Integer$_P$*rogramming.html*?*id* = *x*7*RvQgAA CAAJredir*$_e$*sc* = *y*.

Wright, T., (2017). Exact optimal sample allocation: More efficient than Neyman. *Stat. Probab. Lett.*, 129, pp. 50–57. Available from: https://doi.org/10.1016/j.spl.2017.04. 026.

Wright, T., (2020). A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. *Stat. Probab. Lett.*, 165. Available from: https://doi.org/ 10.1016/j.spl.2020.108829.

Xiang, Y., Gubian, S., Suomela, B. and Hoeng, J., (2013). Generalized simulated annealing for global optimization: The GenSA package. *R J.*, 5(1), pp. 13–28.Available from: https://doi.org/10.32614/RJ-2013-002.

Yates, F., (1960). *Sampling Methods for Censuses and Surveys*, Charles Griffin and Co., London. Available from: https://archive.org/details/samplingmethodsf0000fran/page/ n5/mode/2up.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 21–42, https://doi.org/10.59139/stattrans-2026-002
Received – 31.10.2024; accepted – 26.08.2025

Sampling techniques in the CPI measurement

Jacek Białek[1]

Abstract

The procedure used by a National Statistical Office (NSO) for collecting prices to produce the Consumer Price Index (CPI) is based on sample surveys. The universe (or population) of items has three dimensions: product, geographical, and time, all of which are described in the paper. This paper presents and discusses general concepts and techniques of survey sampling that are crucial for the construction of price indices. In particular, both probability and non-probability sampling techniques are discussed and illustrated with the real-world examples. A separate section discusses sampling scanned products. One of the approaches used for such data is the *dynamic approach*, which involves monthly sampling by applying appropriate data filters. This technique can be seen as a special form of *cut-off sampling*. The empirical study investigates the effect of data filtering on the level of price indices. The main pragmatic conclusion is that the low-sales filter has the most significant impact on reducing the size of the scanner dataset. The second important conclusion is that changing the order of data filtering has minimal impact on the value of the price index.

Key words: probability sampling, non-probability sampling, Consumer Price Index, scanner data, dynamic approach, multilateral indices

1. Introduction

The procedure used for price collection by a National Statistical Office (NSO) when producing a Consumer Price Index (CPI) is a *sample survey*. Here, the CPI (or the Harmonised Index of Consumer Prices, HICP) can play a role of a *target quantity* which is defined with respect to (CPI Manual, 2004): (1) *a universe* that comprises finite population of units (e.g., products or outlets); (2) *variables*, which are defined for the units in the universe (e.g., prices and quantities of products or expenditure shares of outlets); (3) *a parameter*, which is a single value obtained on the basis of values of those variables (e.g., the Jevons (1865) or the Laspeyres (1871) price index).

In general, there are three sampling dimensions (HICP Methodological Manual, 2018; CPI Manual: Concepts and methods, 2020): (I) *a product dimensions*, which consists of all purchased products and varieties of products; (II) *a geographical and outlet dimension*, which consists of all places (e.g., small shops, supermarkets, petrol stations, web-pages, etc.) where the product is sold; (III) *a time dimension*, which comprises those days of the month for which the applicable price index is determined.

For each of these dimensions, there is a *general population* from which a sample will be drawn. The population (universe) of products from the CPI basket is divided into COICOP

[1]Department of Statistical Methods, University of Łódź, Łódź, Poland. E-mail: jacek.bialek@uni.lodz.pl & Department of Prices and Services, Statistics Poland, Poland. E-mail: J.Bialek@stat.gov.pl.
ORCID: https://orcid.org/0000-0002-0952-5327.

5-digit sub-classes, although this division can go down to a lower level of data aggregation (e.g., COICOP 6-digit level) for web-scraped data or scanner data (see Section 5). A sample of products is drawn from each product sub-class, with a common practice being to decide on *representative products* in each sub-class. The population (universe) of outlets includes all places that sell consumer products in a given COICOP product group. Since outlets have specific locations on the country map, the outlet universe has a geographical character. For the time dimension, the universe consists of all sub-periods of the month since the consumer may buy products on any day of the month. The CPI Manual (2004) pays less attention to the time dimension because "price variation is usually smaller over a short time span." However, at least for web-scraped datat, this aspect may be more relevant.

Depending on the product group, the above-mentioned dimensions have differing degrees of importance when collecting data to measure inflation (HICP Methodological Manual, 2004). For instance, *fresh fruits* (COICOP 5: 01161) have highly volatile prices within a month, so the price sampling strategy should not focus only on the product and outlet dimension but also on the time dimension. In contrast, prices for *actual rentals paid by tenants* (COICOP 0411) are generally fixed for at least a month; thus, the time dimension is irrelevant in the sampling procedure.

Sampling is an alternative to conducting a full survey on all observations from a population, which is obviously impossible in practice and would be too costly. Additionally, excessive workload for interviewers in the field could substantially reduce their efficiency and the quality of the data collected. However, while this remark is valid for *traditional data collection*, it has limited applicability when dealing with *alternative data sources*. For example, in the case of scanner data, the use of multilateral indices generally does not require any random sampling of products (Eurostat, 2022). Instead, all data from all outlets from a given retail chain are included. However, to ensure the representativeness of the data while simultaneously reducing analysis time, various data filters are then applied. This thread will be discussed in more detail in Section 5.

In the simplest terms, *probability sampling* involves selecting units in such a way that each one (e.g., product, outlet, or day) has a known non-zero probability of being included in the sample. For example, in an outlet draw, we can determine that each outlet has an equal probability of being included in the sample or that this probability is proportional to the number of people employed at that outlet or its sales revenue. In contrast, in *non-probability sampling*, the probability of selecting any particular unit is unknown (and often impossible to determine). Although *probability sampling* is the recommended approach for sampling in statistical surveys, *non-probability sampling* techniques still dominate the CPI measurement in most countries.

The main aim of the paper is to discuss and verify the effectiveness of selected sampling techniques used in the CPI measurement. This article addresses two gaps in the existing literature. First, it illustrates the sampling techniques discussed for constructing an inflation basket with respect to scanner data. To the best of the author's knowledge, there is a lack of studies in the literature that apply these techniques using scanner data; this article not only fills that gap but also provides practical scripts written in the R environment. Second, no existing research examines the impact of the sequence in which data filters are applied to scanner data on the resulting price index. Data filtering is a recommended method for

selecting scanned product samples within the so-called dynamic approach (see Section 5.2). Therefore, addressing this second research gap may be of substantial importance to statistical offices that employ the chain Jevons index to estimate inflation based on scanner data.

The structure of the paper is as follows: Section 2 discusses the main probability sampling techniques; Section 3 describes non-probability sampling methods; Section 4 presents the main results obtained when trying to estimate bilateral population price indices; Section 5 discusses the problem of sample selection when using scanner data to compile the CPI, and Section 6 is an empirical study which compares the effectiveness of selected sampling techniques based on real scanner data sets. The main conclusions from the empirical study are discussed in Section 7.

2. Probability sampling techniques

This section presents, discusses and illustrates methods of survey sampling that are implemented when measuring a CPI. In particular, the sub-sections focus on three main probability sampling techniques, i.e., *simple random sampling*, *systematic sampling* (in two variants) as well as *probability proportional to size sampling* (for a broader overview on that topic see Särndal et. al. (2003)). Please note that both Section 2 and Section 3 concern the traditional CPI data collection, and Section 5 discusses sample selection for scanner CPI data.

The survey sampling approach assumes that the universe (population) consists of a finite number N of observational units. The sampling procedure selects a sample S that comprises n units out of N available, where the inclusion probability $\pi_i = P(i \in S)$ is known for each unit $i \in \{1,2,...,N\}$.

The universe can be divided into strata, denoted here by $h \in \{1,2,...,H\}$. Each stratum, which can be treated as mini-universe with sampling taking place independently in each one, consists of N_h units, where $\sum_{h=1}^{H} N_h = N$. For example, for outlet dimension the universe of outlets can be divided into four disconnected sub-populations: online stores, small neighborhood stores, supermarkets and hypermarkets.

A *sampling frame* is a list of all (or most) of the N units in a given universe. Sampling frames for the outlet dimension could be business registers or any records of local administrations (CPI Manual, 2004). A products list obtained from sellers or a product list obtained from price collectors can be used as sampling frames for the product dimension.

2.1. Simple random sampling and systematic simple sampling

In *simple random sampling* and *systematic simple sampling* each unit is drawn with equal inclusion probability, which means that $\pi_i = \frac{n}{N}$. In simple random sampling, all units are sampled with replacement. Simple random sampling without replacement is not addressed in the CPI Manual (2004), likely because it entails a changing selection probability for each unit as the population size diminishes with successive draws. Therefore, this article only considers simple random sampling with replacement. In systematic sampling, only the first element is drawn randomly in that way, and the remaining units are selected at equal distances from each other in the sampling frame (CPI Manual, 2004).

2.2. Probability proportional to size sampling

In *probability proportional to size (pps) sampling*, the inclusion probability is proportional to an auxiliary variable x_i (CPI Manual, 2004; HICP Methodological Manual, 2018). This can be expressed as $\pi_i = nx_i/\sum_{j=1}^{N} x_j$. CPI Manual (2004) on page 69 states: "Units for which initially this quantity is larger than one are selected with certainty, whereafter the inclusion probabilities are calculated for the remainder of the universe". For example, when drawing outlets, an auxiliary variable could be the number of people employed at the outlet or the sales volume from the last year of operation (if this information is available).

While it is theoretically possible to consider a fixed or random sample size, when compiling a CPI in practice a fixed sample size is typically considered in each stratum (CPI Manual, 2004, p. 70). Specifically, a statistical office can consider various sampling techniques that provide fixed-size *pps* samples. One such technique is *systematic pps sampling*, which follows a similar concept to *simple sampling*, but the first sample element is drawn in the *pps* scheme. Another techique is *order pps sampling*, which is described below.

Order pps sampling is a commonly accepted technique for selecting *pps* samples, and is widely discussed in Rosén (1997a, 1977b). Once the auxiliary variable x_i is determined, the procedure begins by assigning each $i-$th unit in the population a uniform random number $U_i \in (0,1)$. The units are then assigned a number Q_i as the value of a differential function with arguments x_i and U_i, i.e., $Q_i = f(x_i, U_i)$. The units in the population are then sorted in ascending order relative to the value of Q_i. The n units with the smallest Q_i values are sampled. The CPI Manual (2004) discusses two important cases of the above-mention approach, i.e., *sequential pps sampling* with $Q_i = U_i/z_i$, and *Pareto pps sampling* with $Q_i = (U_i(1-z_i))/(z_i(1-U_i))$, where $z_i = nx_i/\sum_{j=1}^{N} x_j$. Rosen (1997b) showed that for estimating mean and variance, these order sampling techniques are only approximately *pps*. *Pareto pps* is marginally better than *sequential pps* and should therefore be preferred in the price index contex (CPI Manual, 2004, p.71). For more detailed information about *Pareto pps* see, for instance, Lindblom and Teterukovsky (2007), where a case with strata is considered. As it was mentioned above, probability sampling is less commonly used by statistical agencies than non-probability sampling. However, Lindblom (2003) provides details concerning the probabilistic approach used in Sweden, while probability sampling methods used by the US Bureau of Labor Statistics are described in Sections 5.24 - 5.26 in the CPI Manual (2004).

2.3. Empirical illustration

The empirical illustration of probability sampling concerns the selection of outlets of retail chain operating in Poland, with the objective of determining price indices for an elementary group of coffee products. For demonstration purposes, we will use scanner data on coffee sales as available in the *PriceIndices* R package (Białek, 2021). A more extensive discussion of scanner data is given in Section 5, and thus, a detailed description of the structure of this type of data is omitted here.

The 'coffee' dataset contains transaction data of coffee sales in $N = 20$ outlets representing the population for this study. We assume that we need to draw a sample of $n = 4$ outlets. The sales data includes three types of coffee: instant coffee, coffee beans and ground coffee, and we will focus on the period from January 2019 to December 2019. It means that we ob-

served 79 coffee products with a total of 14,392 records. The script that implements the outlet sampling is available at https://github.com/JacekBialek/important_documents/blob/main/SIT_illustration_1.Rmd.

As a result of running the R script, the user receives a table of results on the basis of which the selection of outlets is made. Let us first discuss the columns of this table. The first column (*Outlet ID*) indicates the outlet identification number assigned by the retail chain. The x_i column contains the values of the size variable, which in our illustration is the annual coffee sales revenue of each outlet (in PLN). The next column, x_i^{cum}, contains the cumulative values of the size variable (PLN). Column z_i contains values of the intermediate variable described in Section 2.2, which will be used in the *pps* method. Uniform random values between 0 and 1 are in the column labelled U_i. The values of $Q_i = f(x_i, U_i)$, depending on whether the *sequential pps sampling* or *Pareto pps sampling* technique is implemented, are in columns Q_i^{seq} and Q_i^{Par} respectively. Finally, the last four columns indicate the four outlets drawn, depending on the probabilistic sampling method. Specifically, we have the results for: simple random sampling (*simple*), systematic pps sampling (*systematic*), sequential (order) pps sampling (*seq*) and Pareto (order) pps sampling (*Pareto*).

Table 1. Selection of outlets using the sampling techniques

Outlet ID	x_i	x_i^{cum}	z_i	U_i	Q_i^{seq}	Q_i^{Par}	simple	systematic	seq	Pareto
2183	747848.76	747848.76	0.18	0.81	4.53	19.51	-	-	-	-
2381	859283.40	1607132.16	0.21	0.07	0.34	0.29	-	✓	✓	✓
2681	844018.93	2451151.09	0.20	0.61	3.03	6.23	-	-	-	-
3782	702174.50	3153325.59	0.17	0.13	0.79	0.76	✓	-	-	-
4080	928925.11	4082250.70	0.22	0.42	1.88	2.51	-	-	-	-
4281	938415.01	5020665.71	0.22	0.54	2.42	4.10	-	-	-	-
4380	796774.77	5817440.48	0.19	0.15	0.81	0.78	✓	✓	-	-
4580	1159091.58	6976532.06	0.28	0.08	0.28	0.22	✓	-	✓	✓
4681	807040.59	7783572.65	0.19	0.64	3.33	7.47	-	-	-	-
4780	894942.71	8678515.36	0.21	0.25	1.15	1.20	-	-	-	-
4883	770725.98	9449241.34	0.18	0.51	2.79	4.68	-	-	-	-
5480	826464.99	10275706.33	0.20	0.10	0.52	0.47	-	✓	-	-
6681	809634.89	11085341.22	0.19	0.97	5.01	123.33	✓	-	-	-
7081	728462.60	11813803.82	0.17	0.43	2.48	3.61	-	-	-	-
7481	854912.08	12668715.90	0.20	0.36	1.76	2.18	-	-	-	-
7482	626678.63	13295394.53	0.15	0.05	0.34	0.30	-	-	✓	✓
8480	1153981.94	14449376.47	0.28	0.59	2.13	3.73	-	✓	-	-
8580	846678.61	15296055.08	0.20	0.42	2.07	2.84	-	-	-	-
9082	755509.54	16051564.62	0.18	0.03	0.14	0.12	-	-	✓	✓
9182	712747.31	16764311.93	0.17	0.89	5.21	37.85	-	-	-	-

While *simple random sampling*, *sequential pps sampling* and *Pareto pps sampling* methods have been sufficiently described in Section 2.2, the results in Table 1 on *systematic pps sampling* still require additional clarification. After determining the cumulative values of size variable x_i^{cum}, one integer I_x from the interval $(0, max)$ is drawn (*one_sample_number* in R script), where *max* is the floor value of the last cumulative size variable value (i.e., the total sum of size variable x_i) divided by $n = 4$. In the next step, the next three numbers are determined non-randomly: $I_x + max$, $I_x + 2max$ and $I_x + 3max$. At this stage we have four threshold values. The final stage involves selecting those outlets for which the cumulative value of the size variable has exceeded the given threshold for the first time. In our empirical illustration we obtain: $max = 4191078$, $I_x = 1469607$, $I_x + max = 5660685$,

$I_x + 2max = 9851763$ and $I_x + 3max = 14042841$, which leads to the following sample of outlets: $\{2381, 4380, 5480, 8480\}$. The sample structure is exactly the same in the case of *sequential (order) pps sampling* and *Pareto (order) pps sampling*, although this in not always guaranteed. In our case these two techniques lead to the following sample of outlets: $\{2381, 4580, 7482, 9082\}$, while *simple random sampling* provides the following sample of outlets: $\{3782, 6681, 4580, 4380\}$ (the outlet with $ID = 4580$ appears in both samples).

3. Non-probability sampling techniques

Probability sampling is more advanced than non-probability sampling and is thus more demanding on the researcher (statistician). Therefore, non-probability sampling is easier to implement, and perhaps this is one of the reasons why this approach is more common in the practices of statistical offices. Another technical reason could be the lack of availability of a sampling frame, especially for the product dimension. An argument for using non-probability sampling may also be the low measurement bias it generates as a result. Moreover, de Haan, Opperdoes and Schut (1999) verified the bias that results from non-probability sampling based on scanner data and found that the mean square error (MSE) was often smaller than that for *pps* sampling. Furthermore, when there is a shortages of interviewers, it may be cheaper to collect prices close to where the interviewers live. Sending interviewers to new locations and training them each time a new sample is drawn is certainly both time-consuming and costly. Finally, in probabilistic sampling, statisticians must often contend with oversampling, such as when the population of individuals is already small at the outlets. Thus, below we will discuss the main approaches commonly used in non-probability sampling.

3.1. Cut-off sampling

Cut-off sampling refers to the situation when the n 'largest' sampling units are selected with certainty, and the remaining units have zero chance of being included in the sample (CPI Manual, 2004). The term 'largest' units refers to units with the highest values of size variable that are highly correlated with the target variable. In general, the *cut-off sampling* method provides biased estimators; however, if we are primarily concerned with reducing MSE, this method may be a good way of sampling. This is because any estimator from *cut-off sampling* has zero variance (de Haan, Opperdoes and Schut, 1999).

A specific case of *cut-off sampling* is the filtering of scanner data using a *dynamic approach*. On the one hand, the automation of the collection of electronic transaction data and its full availability (provided that the retail chain signs an agreement with the statistical office) means that there is no need to sample products, varieties or points in time when using scanner data (CPI Manual, 2004, p. 74). On the other hand, to reduce chain drift bias and account for the impact of clearance sales and dump prices, some statistical offices choose to use the chain Jevons index while first filtering the scanner data (e.g., by eliminating relatively low sales or product with extreme price changes from the dataset). This process is known as a *dynamic approach*, where samples are selected in this way from month to month. It will be discussed in more detail in Section 5.

3.2. Quota sampling

Quota sampling is a non-random selection method for survey samples (Cochran, 1977). The share (number or percentage) of units in the sample is determined in such a way that it is proportional to their actual share in the entire survey population. Although a sample obtained through quota sampling is not selected using a random technique, it can still be representative of the entire population, albeit to a limited extent. This representativeness largely depends on the level of detail available about the population under study.

It is important to note that *quota sampling* requires central management of the whole sampling process, which may limit its usefulness in practice. Additionally, the standard error of any estimate cannot be determined in the case of *quota sampling*, further limiting this method (CPI Manual, 2004).

3.3. The representative item method

The representative item method is a traditional CPI method where the statistical office compiles a list of product types along with their specifications (CPI Manual, 2004). If the product type specification is very precise and, therefore, narrow, interviewers receive exact guidelines on which product should be added to the sample. However, this precision may make searching for a product that is compliant with the specification much more difficult or even impossible in a given area or a given period. Conversely, if the type-specification is relatively broad, interviewers have greater freedom in selecting a sample of the most popular products locally. As a rule, this approach leads to better representativity of the sample compared to the narrow type-specification variant.

3.4. Sampling in time

The CPI (like the HICP) refers to a month. Prices of goods and services usually fluctuate throughout the month; however, in the practice of statistical offices, interviewers collect prices on a specific day of the month. The CPI Manual (2004) gives the 15th day of the month as an example of a reference day for price measurement. In Poland, price quotations are carried out by interviewers from the 5th to the 22nd of each month. As a rule, prices of goods are collected once a month, but for some products, price quotations are more frequent (e.g., prices of fresh fruit in Poland are collected twice a month due to their high price volatility). However, even quoting prices twice a month may be insufficient when the price of a product or service varies substantially and depends, for example, on the day of the week (e.g., cinema tickets are more expensive on weekends). Nevertheless, this practice results more from statistical offices' limited funds and human resources rather than from methodological guidelines.

A separate issue when using scanner data to compile a CPI is *sampling in time*. In this case, the expectations of the statistical office must be confronted with the cooperation offered by the retail chains. However, assuming that the agreement between the retail chain and the statistical office leaves a lot of freedom in selecting the period from which the data should come within a month, the individual product usually covers the first three (or sometimes even four) weeks of the month (Eurostat, 2022). For more details, see Section 5.

3.5. Empirical illustration

This section illustrates the *cut-off sampling* method using scanner data on coffee sales, as described in Section 2. This method will be used to sample $n = 9$ products from $N = 79$ coffees products available for sale in 2019. In this empirical illustration, we will consider two size variables: the value of coffee sales and the number of coffee product sold. The total value of coffee sales in the population is 16,764,311.93 PLN, which corresponds to the total number of coffee products sold: 358575. Each product (coffee) is identified based on an internal code (ID) assigned by the retail chain, which has been verified to have a 1:1 relationship with the EAN (European Article Number) barcode. Additionally, two levels of data aggregation are considered: the higher COICOP 5 level (where, as mentioned above, the population contains 79 units) and the lower level of aggregation, i.e. COICOP 6 level, where sales are divided into three product subgroups: coffee beans (23 IDs), ground coffee (37 IDs) and instant coffee (19 IDs).

Since *cut-off sampling* involves considering the $n = 9$ largest products in terms of sales value and sales volume, let us first look at the histograms for these two size variables (Fig. 1 and Fig. 2). At first glance, noticeable differences can be observed between the distributions of size variables, supporting the initial hypothesis that the choice of size variable may substantially impact the final sample selection using the discussed technique.

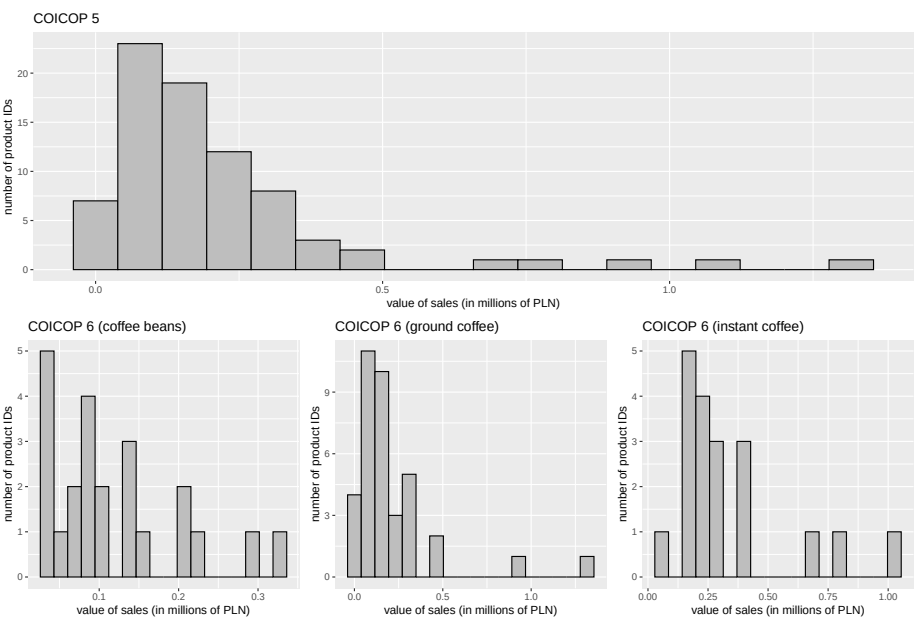

Figure 1. Histogram of sales values determined for the coffee products population, presented at both COICOP 5 and COICOP 6 aggregation levels

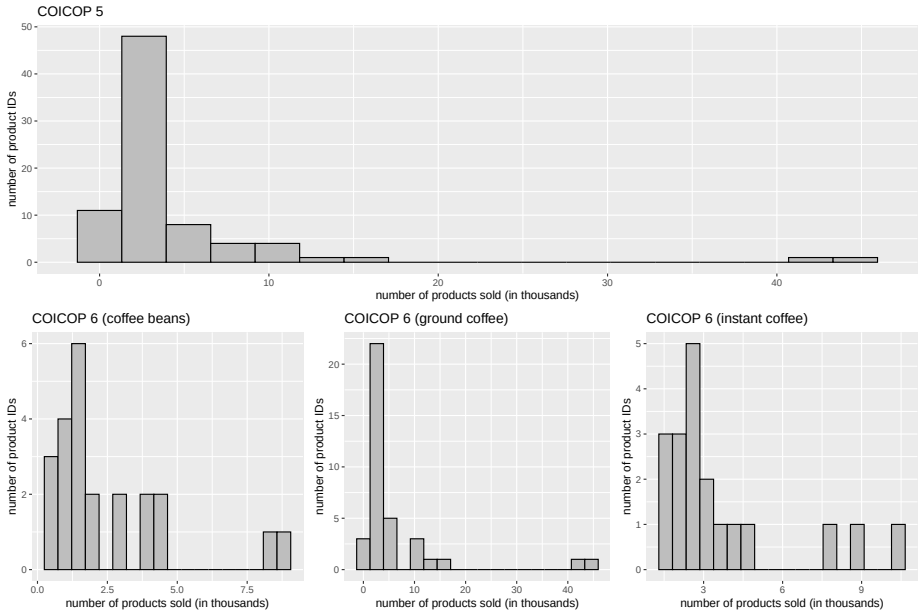

Figure 2. Histogram of the number of products sold determined for the coffee products population, presented at both COICOP 5 and COICOP 6 aggregation levels

Our second hypothesis to be verified is that the samples obtained by selecting three units from each of the three coffee subgroups (based on the given size variable) do not necessarily give the same result as the initial sample of 9 units taken at the COICOP 5 level. Both working hypotheses were confirmed in our empirical illustration, but the first hypothesis holds only at the higher level of data aggregation (COICOP 5). Table 2 and Table 3 highlight the ID numbers of coffee products that appeared in all considered cut-off sampling variants in bold. As one can see, both the selection of the size variable and the choice of the level of data aggregation are important with regard to the structure of the sample obtained using the *cut-off sampling* method.

Table 2. Characteristics of samples of coffee products, presented at both COICOP 5 and COICOP 6 aggregation levels (a size variable is the value of coffee sales)

	COICOP 5	**COICOP 6**		
Characteristics	call coffee products	ccoffee beans	cground coffee	cinstant coffee
sample product IDs	**2401950**, **2401947**, **2402723**, 2401948, **2400379**, **2400915**, **2402453**, 2400368, 32308	33955, 75096, 22687	**2401950**, **2402723**, **2400915**	**2401947**, **2401948**, **2400379**
total sales (PLN)	6436417.44	845251.21	2669982.48	2555225.97
population share (%)	38.39	31.06	34.98	39.84

Table 3. Characteristics of samples of coffee products, presented at both COICOP 5 and COICOP 6 aggregation levels (a size variable is the number of coffee product sold)

	COICOP 5	**COICOP 6**		
Characteristics	all coffee products	coffee beans	ground coffee	instant coffee
sample product IDs	**2402723**, **2401950**, **2400915**, **2402453**, 2400655, 2401380, **2401947**, 2403353, **2400379**	33955, 22687, 89025	**2401950**, **2402723**, **2400915**	**2401947**, **2401948**, **2400379**
no. of sold products	168776	21825	102351	27213
population share (%)	47.06	37.23	44.44	39.05

Coffee bean products were underrepresented at the COICOP 5 level, which is due to the low sales value within this product group. However, at the COICOP 6 level, this group has representatives in the sample (see Table 2 and Table 3). Notably, at the COICOP 6 level there are almost no differences in the sample structure due to the size variable (in fact, the samples differ in only one coffee bean). At the higher level of data aggregation (COICOP 5), samples designated for different size variables overlap in only 2/3 of cases. We encourage the reader to conduct similar experiments for a larger sample size since the script that implements the presented coffee product cut-off sampling is available at https://github.com/JacekBialek/important_documents/blob/main/SIT_illustration_2.Rmd

4. Price indices in the sampling approach

Measurement of the CPI begins at the elementary level, where interviewers note the prices of representatives of each elementary group of products sold in various outlets in the survey regions (e.g., Poland has 207 such regions). At this level, elementary (unweighted) indices are used to determine price dynamics, which are discussed in detail in Section 4.1. At higher levels of aggregation, where information is available on both prices and consumption levels, weighted price indices are used, selectively discussed in Section 4.2. An exception arises with scanner data, where knowledge of consumption levels is already available at the lowest level of data aggregation (the bar-code level), and therefore there are no restrictions on the choice of the price index formula at the elementary level of data aggregation (see Section 5).

Let us suppose we have a population (universe) of N goods and we are interested in estimating a target (population) price index $P^{0,t}$, which compares a current period t with a base one 0. To achieve this aim we collect a sample $S \subset \{1,2,..,N\}$ of goods for which, depending on the information available, we can obtain full observations $\{p_i^0, p_i^t, q_i^0, q_i^t : i \in S\}$ or limited observations $\{p_i^0, p_i^t : i \in S\}$, where p_i^τ and q_i^τ denote the price and quantity of the $i-th$ unit in a period $\tau \in \{0,t\}$, respectively. Based on the drawn sample S of n units we estimate the population price index $P^{0,t}$ using the sample price index $\hat{P}^{0,t}$.

4.1. Population and sample unweighted indices

Well-established elementary price indices include the Dutot, Carli and Jevons indices (von der Lippe, 2007; CPI Manual, 2004; CPI Manual: Concepts and methods, 2020). Chronologically, the first formal proposal of an elementary price index comes from the French economist Nicolas Dutot (1738). The population Dutot price index can be presented as a ratio of unweighted arithmetic means of prices from compared periods, i.e.,

$$P_D^{0,t} = \frac{\frac{1}{N}\sum_{i=1}^{N} p_i^t}{\frac{1}{N}\sum_{i=1}^{N} p_i^0}. \tag{1}$$

In 1764, the Italian economist Gian Rinaldo Carli proposed an elementary index as an unweighted arithmetic mean of price relatives, known as the Carli (1804) index. It can be expressed as follows:

$$P_C^{0,t} = \frac{1}{N}\sum_{i=1}^{N} \frac{p_i^t}{p_i^0}. \tag{2}$$

However, due to its superior axiomatic properties, the most recommended elementary price index formula is the Jevons (1865) index (Levell (2015)). This index uses an unweighted geometric mean of price relatives and can be written in terms of the natural logarithm of prices as follows:

$$P_J^{0,t} = \left(\prod_{i=1}^{N} \frac{p_i^t}{p_i^0}\right)^{\frac{1}{N}}. \tag{3}$$

The last elementary index presented is the Balk-Mehrhoff-Walsh (BMW) index, independently obtained by Mehrhoff and Balk as a linear approximation of the Walsh (1901) index (Eurostat (2018), p. 176; Balk (2005), p. 689). It is formulated as:

$$P_{BMW}^{0,t} = \frac{\sum_{i=1}^{N} \sqrt{\left(\frac{p_i^t}{p_i^0}\right)}}{\sum_{i=1}^{N} \sqrt{\left(\frac{p_i^0}{p_i^t}\right)}}. \tag{4}$$

The sample counterparts of these formulas are denoted in the paper by $\hat{P}_D^{0,t}$, $\hat{P}_C^{0,t}$, $\hat{P}_J^{0,t}$ and $\hat{P}_{BMW}^{0,t}$ respectively. For instance, the sample Jevons price index can be written as follows:

$$\hat{P}_J^{0,t} = \left(\prod_{i\in S} \frac{p_i^t}{p_i^0}\right)^{\frac{1}{n}}. \tag{5}$$

Silver and Heravi (2007) compared the population elementary indices. The statistical approach, which treats calculated elementary indices as estimators of population indices, has been discussed in several studies, including Balk (2005), McClelland and Reinsdorf (1999), and Dorfman, Leaver, and Lent (1999). For instance, McClelland and Reinsdorf (1999) highlight the small sample bias associated with the sample Jevons index when used as an estimator of its population counterpart. Białek (2020) extends Silver and Heravi's (2007) findings by considering the case with correlated prices. Specifically, he demon-

strates that the Carli population price index is very sensitive to changes in the level of price correlations when prices are log-normally distributed. Furthermore, Białek (2022) uses a very general continuous-time stochastic approach to compare elementary indices. In particular, he compares expected values and variances of sample Dutot, Carli and Jevons indices under the assumption that prices are described by a geometric Brownian motion (GBM).

As the purpose of this paper is not to discuss the properties of sample price indices in detail, the formulas for their variances or Mean Square Errors (MSEs) are omitted. For readers interested in more detail in this area, we recommend Balk (2005). In the following sections, we present only the most important findings regarding sample elementary price indices, which can serve as estimators for the weighted price indices discussed in Section 4.2. These main results are presented in Table 4, where s_i^0 and s_i^t denote the expenditure share of the $i-$th population unit in the base and current periods, respectively.

The term "approximately unbiased" estimator is used when presenting the results in Table 4 and Table 5. Following the CPI Manual (2004), we understand this term to refer to an estimator whose bias is small and decreases as the sample size increases, indicating that the estimator is therefore asymptotically unbiased.

Table 4. Selected estimation finding concerning unweighted sample indices $(*)$

Probability sampling method	Proportionality of weights	Estimation finding
simple random sampling	no weighting scheme	$\hat{P}_D^{0,t}$ is the approximately unbiased estimator of $P_D^{0,t}$
pps sampling method	$p_i^0 / \sum_{j=1}^N p_j^0$	$\hat{P}_C^{0,t}$ is the unbiased estimator of $P_D^{0,t}$
pps sampling method	s_i^0	$\hat{P}_J^{0,t}$ is the approximately unbiased estimator of $P_T^{0,t}$
pps sampling method	s_i^0	$\hat{P}_C^{0,t}$ is the unbiased estimator of $P_L^{0,t}$
pps sampling method	q_i^0	$\hat{P}_D^{0,t}$ is the approximately unbiased estimator of $P_L^{0,t}$
pps sampling method	$\sqrt{s_i^0 s_i^t}$	$\hat{P}_{BMW}^{0,t}$ is the approximately unbiased estimator of $P_W^{0,t}$

$*$ *The weighted population indices are described in Section 4.2*

4.2. Weighted population and sample indices

At higher levels of data aggregation, the Laspeyres (1871) index is used to calculate the price dynamics of the CPI basket (see formula (9)). This is due to the fact that consumption data comes from the Household Budget Survey, which is conducted at a certain frequency (e.g., once a year). Consequently, the weighting system based on consumption levels from the base period is, in practice, already outdated in the current period. From both axiomatic and economic perspectives, it would be ideal to use superlative indices (von der Lippe, 2007), which are discussed below. For scanner data (see Section 5), superlative indices can be used even at the lowest data aggregation level.

Superlative price indices, as discussed by Diewert (1976), are the most frequently recommended index formulas for the Cost of Living Index (COLI) approximation. The list

of population superlative indices begins with the Walsh (1901) and Törnqvist (1936) price indices, which are given by:

$$P_W^{0,t} = \frac{\sum_{i=1}^{N} \sqrt{q_i^0 q_i^t} \cdot p_i^t}{\sum_{i=1}^{N} \sqrt{q_i^0 q_i^t} \cdot p_i^0}, \tag{6}$$

and

$$P_T^{0,t} = \prod_{i=1}^{N} \left(\frac{p_i^t}{p_i^0} \right)^{\frac{s_i^0 + s_i^t}{2}}. \tag{7}$$

where s_i^0 and s_i^t denote the expenditure shares of matched products in months 0 and t.

Another commonly known superlative price index is the Fisher (1922) formula, which can be written as:

$$P_F^{0,t} = \sqrt{P_{La}^{0,t} \cdot P_{Pa}^{0,t}}, \tag{8}$$

where $P_{La}^{0,t}$ and $P_{Pa}^{0,t}$ denote the Laspeyres (1871) price index and the Paasche (1874) price index respectively, given by

$$P_{La}^{0,t} = \frac{\sum\limits_{i \in G_{0,t}} q_i^0 p_i^t}{\sum\limits_{i \in G_{0,t}} q_i^0 p_i^0}, \tag{9}$$

and

$$P_{Pa}^{0,t} = \frac{\sum\limits_{i \in G_{0,t}} q_i^t p_i^t}{\sum\limits_{i \in G_{0,t}} q_i^t p_i^0}. \tag{10}$$

The sample counterparts of the weighted formulas are denoted in the paper by $\hat{P}_{La}^{0,t}$, $\hat{P}_W^{0,t}$, $\hat{P}_F^{0,t}$ and $\hat{P}_T^{0,t}$ respectively. For instance, the sample Walsh price index can be written as follows:

$$\hat{P}_W^{0,t} = \frac{\sum_{i \in S} \sqrt{q_i^0 q_i^t} \cdot p_i^t}{\sum_{i \in S} \sqrt{q_i^0 q_i^t} \cdot p_i^0}. \tag{11}$$

When the target index is a weighted index, the *pps* draw scheme seems to be unnecessary. Table 5 presents the most important findings regarding sample superlative price indices obtained using the *simple random sampling* method (Balk, 2005).

Table 5. Selected estimation finding concerning weighted sample indices

Probability sampling method	Proportionality of weights	Estimation finding
simple random sampling	no weighting scheme	$\hat{P}_T^{0,t}$ is the approximately unbiased estimator of $P_T^{0,t}$
simple random sampling	no weighting scheme	$ln(\hat{P}_F^{0,t})$ is the approximately unbiased estimator of $ln(P_F^{0,t})$
simple random sampling	no weighting scheme	$\hat{P}_W^{0,t}$ is the approximately unbiased estimator of $P_W^{0,t}$

4.3. Empirical illustration

This section illustrates the selected sampling methods for drawing scanner products to calculate sample unweighted and weighted indices, as described in Sections 4.1 and 4.2. The demonstration is based on the scanner data on milk sales, which is implemented in the *PriceIndices* R package (Białek, 2021). The *milk* data set contains $N = 61$ milk products observed over the time interval between Dec, 2018 - Dec, 2019. The following methods are used to sample $n \in \{10, 20, 30\}$ products out of the $N = 61$ milk products available for sale: *cut-off sampling* using total sales value as the size variable, *simple random sampling*, and *pps sampling* with weights proportional to base period expenditure shares.

Table 6 presents the above-discussed population and sample indices for the three selected sampling methods. The columns labelled *cut_off_10, cut_off_20* and *cut_off_30* present sample index results obtained after using the *cut-off sampling* procedure and for sample sizes: $n = 10$, $n = 20$ and $n = 30$, respectively. The columns labelled *simple_10, simple_20* and *simple_30* present sample index results obtained after using the *simple sampling* procedure for the same sample sizes. The columns labelled *pps_10, pps_20* and *pps_30* present sample index results obtained after using the *pps sampling* procedure for the same sample sizes. For the last two probabilistic sampling techniques, the presented index numbers are the results of the simulation experiment in which the sampling procedure was repeated $K = 200$ times, and the mean of the obtained index values was taken.

Table 6. Population indices and mean values of sample indices for the three selected sampling methods

Index name	population_index	cut_off_10	cut_off_20	cut_off_30	simple_10	simple_20	simple_30	pps_10	pps_20	pps_30
Dutot	0.951437	0.988638	1.001703	1.022526	0.992398	0.966445	0.965991	1.002938	0.999994	1.005274
Carli	1.041709	0.986144	0.995187	1.006313	1.041499	1.038971	1.045865	1.060286	1.062067	1.062032
Jevons	1.024937	0.981871	0.992935	1.004036	1.026928	1.023023	1.028455	1.038357	1.038215	1.039196
BMW	1.025366	0.981873	0.992930	1.004035	1.027230	1.023403	1.028890	1.039075	1.039077	1.040043
Laspeyres	1.001400	0.998832	0.999979	1.001354	1.000120	0.999190	1.002136	1.007905	1.006967	1.009558
Paasche	0.972483	0.959394	0.969410	0.972084	0.980302	0.971805	0.975483	0.994187	0.992919	0.992916
Fisher	0.986835	0.978915	0.984576	0.986611	0.990075	0.985339	0.988682	1.000974	0.999887	1.001177
Tornqvist	0.986757	0.978668	0.984461	0.986539	0.989994	0.985244	0.988525	1.000652	0.999512	1.000721
Walsh	0.985306	0.976565	0.982900	0.985069	0.989185	0.983949	0.986995	0.999171	0.997832	0.998695

Table 7. Biases of the sample indices for the three selected sampling methods

Index name	cut_off_10	cut_off_20	cut_off_30	simple_10	simple_20	simple_30	pps_10	pps_20	pps_30
Dutot	0.037200	0.050266	0.071088	0.040960	0.015007	0.014553	0.051501	0.048557	0.053837
Carli	-0.055565	-0.046522	-0.035396	-0.000210	-0.002738	0.004156	0.018577	0.020358	0.020323
Jevons	-0.043066	-0.032002	-0.020902	0.001991	-0.001914	0.003518	0.013420	0.013278	0.014259
BMW	-0.043493	-0.032436	-0.021331	0.001864	-0.001963	0.003524	0.013709	0.013711	0.014677
Laspeyres	-0.002568	-0.001421	-0.000046	-0.001280	-0.002210	0.000736	0.006505	0.005567	0.008158
Paasche	-0.013088	-0.003073	-0.000399	0.007819	-0.000678	0.003001	0.021704	0.020437	0.020433
Fisher	-0.007921	-0.002259	-0.000225	0.003239	-0.001496	0.001846	0.014139	0.013052	0.014342
Tornqvist	-0.008089	-0.002297	-0.000219	0.003237	-0.001513	0.001768	0.013895	0.012754	0.013964
Walsh	-0.008741	-0.002405	-0.000236	0.003880	-0.001356	0.001690	0.013865	0.012527	0.013390

Table 7 presents biases of the sample indices, i.e., differences between their mean values (expected values) and the corresponding population indices. Using the *cut-off method*, there was no simulation procedure, and the sample was taken once (the sales value is fixed for a given period). Table 8 presents standard deviations of the sample indices obtained

in a simulation study, i.e., it concerns only the *simple sampling* and *pps sampling* procedures. The R script that implements the discussed sampling methods in the context of price index estimates is available at:
https://github.com/JacekBialek/important_documents/blob/main/SIT_illustration_3.Rmd

Table 8. Standard deviations of the sample indices obtained using probabilistic sampling methods for two selected sampling methods

index name	simple_10	simple_20	simple_30	pps_10	pps_20	pps_30
Dutot	0.086655	0.077570	0.064017	0.033060	0.031947	0.024423
Carli	0.065156	0.041950	0.033380	0.074991	0.042128	0.023385
Jevons	0.052950	0.034015	0.026580	0.055243	0.031691	0.018728
BMW	0.053305	0.034293	0.026819	0.055927	0.032093	0.018902
Laspeyres	0.039094	0.028847	0.021635	0.028870	0.020125	0.016812
Paasche	0.038158	0.028326	0.022333	0.014513	0.010779	0.010118
Fisher	0.036257	0.026305	0.020186	0.020425	0.013983	0.011862
Tornqvist	0.036046	0.025999	0.019899	0.020035	0.013574	0.011484
Walsh	0.035144	0.024953	0.018960	0.018347	0.012007	0.010101

As shown in Table 7, *cut-off sampling* works much better for weighted sample price indices (when the target indices are their population counterparts) than for unweighted sample indices. Surprisingly, when using this method, the measurement bias of the weighted sample price index is considerably less than the bias generated by the weighted sample index obtained when using probabilistic techniques to draw products. *Simple sampling* works well for both categories of sample indices, although for weighted sample indices, it works worse than *cut-off sampling* but better than *pps sampling*. Nevertheless, similar to the *pps sampling*, increasing the sample size does not lead to a clear reduction in the sample index bias (Table 7). However, the standard deviation - and thus the variance - of the estimators for both unweighted and weighted sample indices noticeably decreases as the sample size increases (see Table 8).

Please note that the population Dutot price index is the most difficult to estimate (Table 7). Perhaps this is due to the fact that this index - as recommended by the CPI Manual (2004) and Eurostat (2018) - should only be used for highly homogeneous product groups. However, in the this study, the *milk* collection contains clearly disjointed subgroups of milk group, e.g., goat's and cow's milk, UHT and pasteurized milk, and low-fat and high-fat milk products. Therefore, the homogeneity condition may be weakened here, which consequently generates an additional bias in measuring the Dutot price index.

5. Sample selection for scanner data

Scanner data refer to electronic transaction data that specify product prices and expenditures obtained from supermarket IT systems by scanning product bar codes, such as the Global Trade Item Number (GTIN), European Article Number (EAN) or Stock Keeping Unit (SKU). Scanner data are a relatively new and cheap data source for calculating the Consumer Price Index (CPI) and the main advantage of using these data is that they provide full information about products, even at the lowest data aggregation level (see Figure 3).

	date	outlet	segment	category	product number	EAN	label	price	quantity
1	2024-03-21	00-199	RYBY SAMOOBSLUGA	PROD. RYBNE PRZETWORZONE	32994	7311170032443	PASTA Z TUŃCZYKA 145G ABBA	9.98	23.00
2	2024-03-21	00-199	WARZYWA	WARZYWA	34021	220003100000	POMIDOR UKŁADANY LUZ	12.97	179.98
3	2024-03-21	00-199	OWOCE	OWOCE PODSTAWOWE	34041	220205500000	BANAN LUZ	3.72	2212.75
4	2024-03-21	00-199	ZDROWA ZYWNOSC	ZYWNOSC EKOLOGICZNA	81189	5905699160163	BIO SYROP MALINOWY 500ML Z DOMU REMBOWSKICH	25.89	4.00
5	2024-03-21	00-199	KONSERWY ZUPY DANIA GOTOW	PASZTETY	103793	3596710010783	) PASZTET Z ŻOŁĄDK.DROBIOWY. 180G. MP	5.49	4.00
6	2024-03-21	00-199	WARZYWA	WARZYWA GOTOWE	143537	5900449007163	SURÓWKA Z MARCHEWKI 300G MAGA	3.59	46.00
7	2024-03-21	00-199	DESERY I DODATKI	BUDYN	170498	5900983025098	BUDYŃ MLECZNA CZEKOLADA KLEKS 42G DELECTA	1.34	5.00
8	2024-03-21	00-199	HIGIENA TOALETOWA	MYDLA W KOSTCE	188510	5900536348735	MYDŁO W KOSTCE COTTON 90G LUKSJA	2.18	17.00
9	2024-03-21	00-199	TLUSZCZE, MLEKO, JAJA	TLUSZCZE	188827	4001954160266	KERRYGOLD MASŁO 200G	6.59	96.00
10	2024-03-21	00-199	PRODUKTY MACZNO ZBOZOWE	KASZA	284681	5906827003109	KASZA PĘCZAK KUJAWSKI 0,9KG MELVIT	5.09	13.00

Figure 3. Sample scanner data frame from a Polish supermarket

Processing scanner data poses a number of challenges, including the automatic classification of products into COICOP groups, matching products over time, data filtering, as well as the selection of a price index formula and the aggregation of partial results (e.g., over outlets). These processes are described in detail by Białek and Beręsewicz (2021). However, the issue of selecting a sample of products for determining a price index on the basis of scanner data is often overlooked. In practice, we can consider the *time dimension*, the *outlet dimension* and the *product dimension* when using CPI scanner data, and each of these aspects can play a measurable role in shaping the final price index (see our empirical study presented in Section 6). These above-mentioned dimensions are described in Section 5.1, while Section 5.2 discusses two main approaches in scanner sample selection. Section 5.3 describes the most popular multilateral indices that are considered in the empirical study.

Multilateral price indices designed for scanner data are much more complex than the bilateral indices discussed in Sections 4.1 and 4.2. Perhaps this is why the literature lacks theoretical results on population and sample-based multilateral indices that are analogous to the results presented in Table 4 and Table 5.

5.1. The time, outlet and product dimensions

The time dimension. According to Eurostat (2022, p. 10) we can read: "If all points in time during a certain period are equivalent to the consumer and there are no price level differences between weekdays and hours of the day, then the whole time period (month or week) can be considered as homogeneous for the purpose of price aggregation". It recommends aggregating data across a period that covers as much of the reference month as possible. In practice, however, statistical offices are limited by the terms of data transmission established with specific retail chains: for example, contracts may stipulate that the data are aggregated from the 5th to the 20th day of the month. A commonly used approach involves collecting scanner data that cover the first three weeks of sales from subsequent months of the retail chain's operations.

The outlet dimension. When working with CPI scanner data, its is generally recommended to specify individual products at the level of a single outlet. Retail chain often have different pricing policies in different outlets, depending on local conditions (e.g., demand for products or competitors' prices). However, determining the price index for each outlet separately is a time-consuming task. In some cases, there are reasons to aggregate scanner data across outlets, e.g., when the chain has an identical pricing policy within a specific region. This strategy can effectively reduce the computation time needed for multilateral price index calculations.

The product dimension. Typically, barcodes are used to identify products at the lowest level of aggregation, e.g., GTIN (Global Trade Item Number), EAN (European Article Number) or SKU (Stock Keeping Unit). However, the problem with disaggregated data is that over a longer period, we can observe *product churn*), i.e., a large number of products emerging and disappearing from the market. This means that the life cycle of a given product code may last a few months. The second problem observed at the bar-code level is identifying *relaunches*. Relaunches may occur when there are changes in the size or colour of the packaging. A change in size requires quality correction and price standardization, while the latter case does not affect product quality but may mean a change in its bar-code. Both scenarios should be detected automatically, which can be achieved by the procedure of matching products in time based not only on the bar-code, but also using the code assigned by the retail chain or the product description. The detection procedure (*data_matching*) is implemented in the *PriceIndices* R package (Białek, 2021).

If homogeneous product are defined too broadly, there is a risk of unit value bias. Conversly, operating at the bar-code level or defining homogeneous products too tightly may lead to problems with detecting relaunches (Eurostat, 2022). The MARS methods can be seen as a solution of this problem since it is a compromise between the above-mentioned two objectives (Chessa, 2021).

5.2. Static vs dynamic approach

There are two approaches that have emerged for using scanner data in the CPI measurement, i.e., *static* and *dynamic*. The *static* approach aligns with traditional data collection methods based on field surveys, while the *dynamic* approach uses the concept of monthly matched samples with the chain Jevons index as a target index.

In the *static* approach, a sample of items is selected at the beginning of each year and these items are monitored and maintained over time. Every month, prices for the selected products are taken from the scanner data files. Similar to the practices of price collectors from the field, if a particular item becomes unavailable, a replacement item is selected and used for further price index calculations.

Due to the high dynamics of scanner data related to product rotation and product seasonality, implementing a *dynamic* approach seems to be a better choice. This approach involves selecting the best-selling items available in two consecutive months each month to measure a monthly price changes. In practice, sample selection is carried out using the *cut-off method*, which is implemented by applying data filters.

The dynamic basket is determined using turnover figures of individual products in two adjacent months, i.e., the product is included in the sample if its turnover is above a fixed threshold determined by the number of products in a given product group. Van Loon and Roels (2018) provided the following condition for the above mentioned rule, which indicates whether the i-th product is taken into consideration when comparing months $t-1$ and t:

$$\frac{s_i^{t-1}+s_i^t}{2} > \frac{1}{n\lambda}, \tag{12}$$

where n is the number of considered products and λ is a fixed parameter (usually set to

1.25). This kind of data filter can be called a *low sale filter*. Proponents of using filters also believe that products displaying extreme price changes from one month to another should also be excluded from the sample (*extreme price filter*). For example, Statistics Poland uses the *extreme price filter* to remove products from the sample whose price has increased more than threefold or decreased more than fourfold. The list of possible data filters is extensive, e.g. Statistics Belgium implements a filter for dump prices (Van Loon and Roels, 2018). With this *dump price filter*, products are eliminated from the sample if a simultaneous, clear decrease in price and sales value is observed. These products will most likely be withdrawn from sale in the near future and, therefore, they are no longer representative.

Data filtering can also be considered when using multilateral indices, which are, in fact, specifically designed for scanner data cases (see Section 5.3). For instance, the *low sales filter* and *dump price filter* are mentioned as a part of *data pre-processing* serving as an initial step before computing multilateral price indices (see Eurostat (2022), p. 4). In particular, the aforementioned document recommends using *dumping filters* together with the CCDI multilateral index (p. 25). It seems that the same remark concerns the GEKS and GEKS-W price indices, since they give more weight to the price decrease of the dumped products (see Sections 5.3 and our *Empirical study*).

5.3. Multilateral indices

As it was mentioned above, multilateral indices are recommended for statistical offices to determine the dynamics of scanner prices (Eurostat, 2020). Commonly known and accepted methods include the GEKS method (Gini, 1931; Eltetö and Köves, 1964), the Geary-Khamis method (Geary, 1958; Khamis, 1972), the CCDI method (Caves et al., 1982), or the Time Product Dummy Methods (de Haan and Krsinich, 2018). Multilateral indices operate on a time window $[0,T]$ and therefore take into account phenomena such as product rotation or product seasonality. Moreover, due to the *transitivity* property, multilateral indices eliminate *chain drift bias* (Eurostat, 2022). The chain drift effect occurs when prices and quantities of products sold return to their original values (e.g., after the season) but the index deviates from the expected value of one. The most commonly used multilateral indices can be also found in Eurostat (2022).

6. Empirical study

This section examines the impact of scanner data sampling methods (under the *dynamic approach*) on the value of the multilateral price index. For this purpose, we will use the data filters discussed in Section 5.2 and the full-window multilateral price indices discussed in Section 5.3. The empirical study is based on the basis of scanner data collection on sales of *cleaning and preservatives* (COICOP: 056111) and *cosmetics and hygiene products* (COICOP: 121321) obtained from a Polish retail chain. The data covers the period: Dec, 2022 - Dec, 2023. The author of the study has not received receive permission to share these datasets, so the R script without the data is available for download from: https://github.com/JacekBialek/important_documents/blob/main/SIT_Empirical%20study.Rmd.

In particular, the study will consider the following data filtering variants: (**v1**) data sets without filtering, (**v2**) the *low sales filter* (**f1**) used with $\lambda = 1.25$; (**v3**) the *extreme price filter* (**f2**) with thresholds: $lower = 0.25$ for price decrease and $upper = 3$ for price increase; (**v4**) the *dump price filter* (**f3**) with thresholds: $lower1 = 0.25$ for price decrease and $lower2 = 0.3$ for sales decrease; (**v5**) all data filters **{f1, f2, f3}** working independently; (**v6**) data filters implemented in order (**f1, f2, f3**); (**v7**) data filters implemented in order (**f1, f3, f2**); (**v8**) data filters implemented in order (**f2, f1, f3**); (**v9**) data filters implemented in order (**f2, f3, f1**); (**v10**) data filters implemented in order (**f3, f1, f2**) and (**v11**) data filters implemented in order (**f3, f2, f1**). The results concerning these variants - specifically regarding dataset reduction and its impact on multilateral price index levels - are presented in Tables 9 and 10. In particular, columns labelled *sample size* and *normalized sample size* describe the number of different products after applying the given type of filter, with the first row in these tables indicating the situation with no filtering.

Table 9. Different variants of data filtering and their impact on sample size and multilateral index values (*cleaning and preservatives*)

Filter variant	Sample size	Normalized sample size	Chain Jevons	Geary-Khamis	GEKS	CCDI	TPD
v1	2078	100	1.05733	1.14268	1.13548	1.13480	1.14140
v2	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15406
v3	1914	92.11	1.05452	1.14287	1.13465	1.13404	1.14233
v4	1915	92.16	1.05733	1.14287	1.13463	1.13399	1.14233
v5	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15406
v6	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15406
v7	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15406
v8	903	43.45	1.11527	1.15401	1.14328	1.14414	1.15401
v9	903	43.45	1.11527	1.15401	1.14328	1.14414	1.15401
v10	905	43.55	1.11460	1.15408	1.14330	1.14416	1.15405
v11	903	43.45	1.11527	1.15401	1.14328	1.14414	1.15401

Table 10. Different variants of data filtering and their impact on sample size and multilateral index values (*cosmetics and hygiene products*)

Filter variant	Sample size	Normalized sample size	Chain Jevons	Geary-Khamis	GEKS	CCDI	TPD
v1	5966	100	0.97274	1.09909	1.09594	1.09465	1.09829
v2	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v3	5393	90.39	0.96662	1.09945	1.09649	1.09540	1.09830
v4	5395	90.43	0.97458	1.09949	1.09654	1.09544	1.09743
v5	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v6	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v7	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v8	1994	33.42	1.08567	1.11877	1.11542	1.11598	1.11984
v9	1994	33.42	1.08567	1.11877	1.11542	1.11598	1.11984
v10	1995	33.44	1.08593	1.11889	1.11553	1.11609	1.12386
v11	1994	33.42	1.08567	1.11877	1.11542	1.11598	1.11984

7. Conclusions

The ultimate aim of CPI sampling techniques is to obtain the most accurate estimate of inflation. A general conclusion from the empirical illustrations presented is that the sample structure depends not only strongly on the sampling technique adopted (particularly on the choice between random and non-random sampling), but also on the level of data aggregation (see the empirical illustration in Section 3.5). The considered sampling technique may turn out to be better than other techniques at COICOP level 5, but worse at COICOP level 6. Further, if we take the bias of the final estimated price index as an evaluation criterion, it may turn out that the considered method performs better or worse depending on whether we estimate a weighted or unweighted index. For instance, in Section 4.3 we found that *cut-off sampling* works much better than *simple random sampling* and *pps sampling* for estimating weighted population price indices.

An important practical conclusion of the empirical study (see Section 6) is that the *low sales filter* has the greatest impact on reducing the size of the scanner dataset. In both analyzed scanner datasets, the product sample size was reduced by more than 55% after applying this filter. In contrast, the other two types of data filters (i.e., the *extreme price filter* and the *dump price filter*) reduced the sample size in a similar yet smaller way (by less than 10%), although they substantially affected the price index value. We can also conclude that the order in which the scanner data filters are applied has no effect on either the sample structure or the value of the price index (see Tables 9 and 10). In other words, changing the order of data filtering has little impact on the value of the price index. As a consequence, each of the filters can be applied independently.

Finally, as expected, the chain Jevons index proved to be much more sensitive to the choice of the data filter than multilateral indices. It is important to note that data filtering is essential if a statistical office intends to use the chain Jevons index as part of a *dynamic approach*. With weighted multilateral indices, while data filtering may not seem necessary, it can effectively reduce the sample size and, thus, the time needed to estimate the index.

References

Balk, B. M., (2005). Price indexes for elementary aggregates: the sampling approach. *Journal of Official Statistics*, 4(21), pp. 675–679.

Bialek, J., (2020). Comparison of elementary price indices. *Communications in Statistics - Theory and Methods*, 49(19), pp. 4787–4803.

Białek, J., (2021). PriceIndices – a new R package for bilateral and multilateral price index calculations. *Statistika – Statistics and Economy Journal*, 36(2), pp. 122–141.

Białek, J., Beręsewicz, M., (2021). Scanner data in inflation measurement: from raw data to price indices. *The Statistical Journal of the IAOS*, 37, pp. 1315–1336.

Bialek, J., (2022). Elementary price indices under the GBM price model. *Communications in Statistics - Theory and Methods*, 51(5), pp. 1232–1251.

Caves, D. W., Christensen, L. R. and Diewert, W. E., (1982). Multilateral comparisons of output, input, and productivity using superlative index numbers. *Economic Journal*, 92(365), pp. 73–86.

Chessa, A. G., (2021). A Product Match Adjusted R Squared Method for Defining Products with Transaction Data. *Journal of Official Statistics*, 37(2), pp. 411–432.

Cochran, W. G., (1977). *Sampling techniques*. New York: John Wiley.

de Haan, J., Krsinich, F., (2018). Time dummy hedonic and quality-adjusted unit value indexes: Do they really differ? *Review of Income and Wealth*, 64(4), pp. 757–776.

Diewert, W. E., (1976). Exact and superlative index numbers. *Journal of Econometrics*, 4(2), pp. 115–145.

Diewert, W. E., (2004). *On the Stochastic Approach to Linking the Regions in the CPI*. Department of Economics, University of British Columbia.

Diewert, W. E., Fox, K. J., (2018). *Substitution bias in multilateral methods for CPI construction using scanner data*. UNSW Business School Research Paper (2018–13).

Dorfman, A. H., Leaver, S. and Lent, J., (1999). *Some observations on price index estimators*. Bureau of Labor Statistics working paper no. 324, Washington DC.

de Haan, J., Opperdoes, E. and Schut, C. ,(1999). *Item Sampling in the Consumer Price Index: A Case Study using Scanner Data*. Research Report, Statistics Netherlands, Voorburg.

Eltetö, O., Köves, P., (1964). On a problem of index number computation relating to international comparison. *Statisztikai Szemle*, 42(10), pp. 507–518.

Eurostat, (2018). *Harmonised Index of Consumer Prices (HICP) Methodological Manual*. Luxembourg: Publications Office of the European Union.

Eurostat, (2022). *Guide on Multilateral Methods in the Harmonised Index of Consumer Prices*. Luxembourg: Publications Office of the European Union.

Fisher, I., (1922). The making of index numbers: a study of their varieties, tests, and reliability. *Number 1. Houghton Mifflin*.

Geary, R. C., (1958). A note on the comparison of exchange rates and purchasing power between countries. *Journal of the Royal Statistical Society. Series A (General)*, 121(1), pp. 97–99.

Gini, C., (1931). On the circular test of index numbers. *Metron*, 9(9), pp. 3–24.

International Labour Office, (2004). *Consumer price index manual: Theory and practice*, Geneva.

International Monetary Fund, (2020). *Consumer Price Index manual: Concepts and methods*. Washington, D.C.

Khamis, S. H., (1972). A new system of index numbers for national and international purposes. *Journal of the Royal Statistical Society: Series A (General)*, 135(1), pp. 96–121.

Levell, P., (2015). Is the Carli index flawed?: assessing the case for the new retail price index RPIJ. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(2), pp. 303–336.

Lindblom, A., Teterukovsky, A., (2007). *Coordination of Stratified Pareto πps Samples and Stratified Simple Random Samples at Statistics Sweden*. Papers presented at the ICES-III, Montreal, Quebec, Canada.

Laspeyres, K., (1871). Ix. Die berechnung einer mittleren waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik*, 16(1), pp. 296–318.

Lindblom, A., (2003). AMU - The system for coordination of frame populations and samples from the Business Register at Statistics Sweden, *Background Facts on Economic Statistics*, 2003:3, Statistics Sweden.

McClelland, R., Reinsdorf. M., (1999). *Small sample bias in geometric mean and seasoned CPI component indexes*. Bureau of Labor Statistics working paper no. 324, Washington DC.

Paasche, H., (1874). Über die preisentwicklung der letzten jahre nach den hamburger börsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, pp. 168–178.

Rosén, B., (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62(2), . 135–158.

Rosén, B., (1997b). On sampling with Proportionality Proportional to Size. *Journal of Statistical Planning and Inference*, 62(2), pp. 159–191.

Särndal, C.E., Swensson, B. and Wretman, J., (2003). *Model Assisted Survey Sampling*. Springer Science and Business Media, Berlin, Heidelberg.

Silver, M., Heravi, S., (2007). Why elementary price index number formulas differ: Evidence on price dispersion. *Journal of Econometrics*, 140(2), pp. 874–883.

Törnqvist, L., (1936). *The bank of Finland's consumption price index*. Bank of Finland Monthly Bulletin, pp. 1–8.

van Loon, K. V. Roels, D., (2018). *Integrating big data in the Belgian CPI*. Paper presented at the meeting of the group of experts on consumer price indices, 8-9 May 2018, Geneva, Switzerland.

von der Lippe, P., (2007). *Index Theory and Price Statistics*. Peter Lang, Berlin, Germany.

Walsh, C. M., (1901). *The Measurement of General Exchange Value*. Macmillan and Co.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 43–64, https://doi.org/10.59139/stattrans-2026-003
Received – 12.12.2023; accepted – 23.06.2025

Optimality of classical difference estimators of finite population variance under random non-response with comparative study

Mahamood Usman[1]

Abstract

In this study, we address the challenge of calculating the finite population variance when faced with random non-response. Such issues are commonly encountered in various fields like medical sciences, environmental sciences and business studies when dealing with data. Using the ranking of an auxiliary variable across three different methodologies of random non-response, we developed several novel difference-type estimators of population variance along with their optimal models. The strategies are shaped by using the varying levels of information available regarding the auxiliary variable. We have studied the properties of the proposed estimators under large sample approximations and determined their optimum situations in each strategy. The introduced estimators can be viewed as an advancement of traditional difference estimators. Within the associated methodologies, we conducted a comparative analysis based on some real datasets as well as simulated datasets, whereby the proposed estimators showed reduced variances when assessed in terms of the enhanced percentage relative efficiencies (PRE) compared to some standard ratio and difference-type estimators relevant to the respective methodologies.

Key words: study variable, population variance, dual use of auxiliary variable, percentage relative efficiency, random non-response.

AMS Subject Classification: 62D05.

1. Introduction

The measurement of variation provides a dynamic idea about the data. For example, a company sales representative may analyze the variations in sales records or the population of customers monthly to help them decide how to improve sales or customer satisfaction. Similarly, a marketing analyst of a company may be interested in analyzing the variability of company sales in a particular area over time to see which products the customers like most. To measure the variation of such kind of data, the survey practitioners often use the term variance. Variance measures the variability of data and is extensively used by analysts in various fields such as agriculture, forestry, medical science, politics, finance, population traits, etc. It plays an important role in the testing of hypotheses and the construction of confidence intervals for population parameters. The attention to the variance estimation techniques has been paid by researchers since long ago. Singh et al. (1973) have proposed

[1]Department of Mathematics (School of Advanced Sciences), Vellore Institute of Technology, Vellore-632014, India. E-mail: mahmoodu33@gmail.com. ORCID: https://orcid.org/0000-0003-1234-4401.

the estimator of population variance using the priori information about the population coefficient of kurtosis and compared it with the usual unbiased estimator. Das and Tripathi (1978) have introduced the estimator using the information on an auxiliary variable. Later on, ratio and regression-type estimators using the information on auxiliary variables have been discussed by various authors such as Isaki (1983), Singh et al. (1988), Upadhyay and Singh (2001), Yasmeen et al. (2019), Zaman and Bulut (2022), among many others. Belili et al. (2023) and Khodija et al. (2023) have presented some improved probability distributions and studied their mathematical properties. They have examined the efficiencies of the estimators through comparative studies. Ahmad et al. (2023), Zaman and Bulut (2024) and Daraz et al. (2025) have proposed some ratio and difference-type estimators of population mean and variance and investigated their optimal behaviors using real and simulated datasets. The above authors have studied the estimation procedures of population parameters in the presence of complete response.

When the non-response is observed in the sample, the problem of estimation of population variance has also been discussed by various authors in the context of random non-response, introduced by Rubin (1976). The authors, notably Singh and Joarder (1998), Kumar (2014), Sharma and Singh (2020), and Bhusan and Pandey (2021) have suggested improved estimators of finite population variance in the presence of random non-response using the information on single and multi-auxiliary variables. It is a well-known fact that the efficiency of the estimators may be increased by using the multiple auxiliary variables. When the information on multi-auxiliary variables is not available, the researchers like Yaqub et al. (2017), Hussain and Haq (2019), Irfan et al. (2020) and many others have published the higher efficient estimators just by recalling the dual or rank of an auxiliary variable. Singh and Usman (2022) have established improved estimators of population variance using the rank of an auxiliary variable in a customary way in the case of random non-response. Recently, the authors like Javed et al. (2023), Almulhim et al. (2024), Bhusan and Pandey (2025), etc., have suggested improved and optimal estimation procedures for estimation of population parameters in the related areas.

Inspired by the aforementioned researchers, the motivation of the present work can be stated as follows:

- Enhancing the efficiency of the classical difference estimator of population variance to the next level.
- Efficient utilization of the rank of an auxiliary variable in the construction of a new model.
- Investigation of the behavior of the new model in three distinct strategies of random non-response.
- Comparison of the new model with existing ones based on numerical and simulation studies.

In this study, we have developed some new models along with their optimal versions for the estimation of finite population variance under the missing at random (MAR) non-response mechanism. We have efficiently employed the rank of an auxiliary variable in the

construction of new estimators in three distinct strategies of random non-response. The novelty of the present work may be stated as the extension of classical difference estimator by utilizing the rank of an auxiliary variable in order to achieve an enhanced level of efficiency. The role of the rank (dual) of an auxiliary variable in the formulation of newly suggested estimators may be easily recognized in terms of higher percentage relative efficiencies compared to existing estimators considered in this study.

The rest part of the paper is constituted as follows. In *Section 2*, the methodology and notations are presented and some customary estimators have been discussed in *Section 3*. The proposed estimators have been formulated in *Section 4*, and their optimal situations have been stated in *Section 5*. In *Section 6*, the properties of proposed estimators have been compared with some relevant existing estimators under a comparative study based on real and simulated datasets. Finally, the conclusions have been made in *Section 7*.

2. Methodology and notations

Consider a finite population $\Omega = \{\Omega_1, \Omega_2, ..., \Omega_N\}$ of size N in which the study variable y and auxiliary variable x are properly correlated with an amount of correlation ρ_{yx}. Suppose that $Z_x = \{z_{x_1}, z_{x_2}, ..., z_{x_N}\}$ denote the ranks of corresponding values of variable $X = \{x_1, x_2, ..., x_N\}$ on which the information is already available in Ω. Draw a sample of size n from Ω using the simple random sampling without replacement (SRSWOR) technique where the information cannot be received on $m\{m = 0, 1, 2, ..., (n-2)\}$ units due to random non-response (MAR) for target variable y only. As a result, the (n-m) responding units that remain are treated as the sample based on the technique of simple random sampling. We assume that the information is missing for auxiliary variable x on corresponding units of y, as per the situations discussed in the present study. If the probability of non-response among the $(n-2)$ possible values of m non-responses is denoted by p, then m follows the distribution given by

$$P(m) = \frac{n-m}{nq+2p}\binom{n-2}{m}p^m q^{n-2-m}; \quad m = 0, 1, 2, ..., (n-2) \tag{1}$$

where $p+q=1$ (*instantly* see Singh et al., 2000). Here, p can be estimated using the maximum likelihood estimation method based on the distribution given in (2.1).

Singh and Joarder (1998) have obtained the maximum likelihood estimator of p as

$$\hat{p} = \frac{(n-1+m) - \sqrt{(n-1+m)^2 - \frac{4nm(n-3)}{(n-2)}}}{2(n-3)}$$

and therefore $\hat{q} = 1 - \hat{p}$.

Now, we define the following notations:
$\bar{Y} = \sum_{i=1}^{N} y_i/N$: Mean of y for entire population
$S_y^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/(N-1)$: Variance of y for entire population.
$\bar{x}^* = \sum_{i=1}^{n-m} x_i/(n-m)$: Respondent mean of x

$\bar{x} = \sum_{i=1}^{n} x_i/n$: Mean of x for selected sample
$\bar{X} = \sum_{i=1}^{N} x_i/N$: Mean of x for entire population
$s_x^{*2} = \sum_{i=1}^{n-m}(x_i - \bar{x}^*)^2/(n-m-1)$: Respondent variance of x
$s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$: Variance of x for selected sample
$S_x^2 = \sum_{i=1}^{N}(x_i - \bar{X})^2/(N-1)$: Variance of x for entire population
$S_{yx} = \sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})/(N-1)$: Population Covariance between y and x
$\bar{z}_x^* = \sum_{i=1}^{n-m} z_{x_i}/(n-m)$: Respondent mean of Z_x
$\bar{z}_x = \sum_{i=1}^{n} z_{x_i}/n$: Mean of Z_x for selected sample
$\bar{Z}_x = \sum_{i=1}^{N} z_{x_i}/N$: Mean of Z_x for entire population
$s_{z_x}^{*2} = \sum_{i=1}^{n-m}(z_{x_i} - \bar{z}_x^*)^2/(n-m-1)$: Respondent variance of Z_x
$s_{z_x}^2 = \sum_{i=1}^{n}(z_{x_i} - \bar{z}_x)^2/(n-1)$: Variance of Z_x for selected sample
$S_{z_x}^2 = \sum_{i=1}^{N}(z_{x_i} - \bar{Z}_x)^2/(N-1)$: Variance of Z_x for entire population
$S_{yz_x} = \sum_{i=1}^{N}(y_i - \bar{Y})(z_{x_i} - \bar{Z}_x)/(N-1)$: Population Covariance between y and Z_x
$S_{xz_x} = \sum_{i=1}^{N}(x_i - \bar{X})(z_{x_i} - \bar{Z}_x)/(N-1)$: Population covariance between x and Z_x.
$\rho_{yx} = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X})}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{Y})^2}\sqrt{\sum_{i=1}^{N}(x_i-\bar{X})^2}}$: Population Correlation between y and x
$\rho_{yz_x} = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})(z_{x_i}-\bar{Z}_x)}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{Y})^2}\sqrt{\sum_{i=1}^{N}(z_{x_i}-\bar{Z}_x)}}$: Population Correlation between y and z_x
$\rho_{xz_x} = \frac{\sum_{i=1}^{N}(x_i-\bar{X})(z_{x_i}-\bar{Z}_x)}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{X})^2}\sqrt{\sum_{i=1}^{N}(z_{x_i}-\bar{Z}_x)}}$: Population Correlation between x and z_x

To obtain the biases and MSEs of the proposed estimators, we assume the following transformations in terms of errors:

$$s_y^{*2} = S_y^2(1+\varepsilon_0),$$
$$\bar{x}^* = \bar{X}(1+\varepsilon_1^*), \qquad \bar{x} = \bar{X}(1+\varepsilon_1),$$
$$\bar{z}_x^* = \bar{Z}_x(1+\varepsilon_2^*) \quad \text{and} \quad \bar{z}_x = \bar{Z}_x(1+\varepsilon_2)$$

such that
$E(\varepsilon_0) = E(\varepsilon_1^*) = E(\varepsilon_1) = E(\varepsilon_2^*) = E(\varepsilon_2) = 0$ and $E(\varepsilon_0^2) = f_2(\lambda_{400}-1) = f_2\lambda_{400}^*$, $E(\varepsilon_1^{*2}) = f_2C_x^2$, $E(\varepsilon_1^2) = f_1C_x^2$, $E(\varepsilon_2^{*2}) = f_2C_{z_x}^2$, $E(\varepsilon_2^2) = f_1C_{z_x}^2$, $E(\varepsilon_0\varepsilon_1^*) = f_2\lambda_{210}C_x$, $E(\varepsilon_0\varepsilon_1) = f_1\lambda_{210}C_x$, $E(\varepsilon_0\varepsilon_2^*) = f_2\lambda_{201}C_{z_x}$, $E(\varepsilon_0\varepsilon_2) = f_1\lambda_{201}C_{z_x}$, $E(\varepsilon_1^*\varepsilon_2^*) = f_2C_{xz_x}$, $E(\varepsilon_1^*\varepsilon_2) = f_1C_{xz_x}$, $E(\varepsilon_1\varepsilon_2^*) = E(\varepsilon_1\varepsilon_2) = f_1C_{xz_x}$.

Here, $f_1 = \left(\frac{1}{n} - \frac{1}{N}\right)$, $f_2 = \left(\frac{1}{nq+2p} - \frac{1}{N}\right)$, $f_3 = f_2 - f_1 = \left(\frac{1}{nq+2p} - \frac{1}{n}\right)$, $\lambda_{klm} = \frac{\mu_{klm}}{\mu_{200}^{k/2}\mu_{020}^{l/2}\mu_{002}^{m/2}}$, $\mu_{klm} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{Y})^k(x_i-\bar{X})^l(z_{x_i}-\bar{Z}_x)^m$, $C_y = \frac{S_y}{\bar{Y}}$, $C_x = \frac{S_x}{\bar{X}}$, $C_{z_x} = \frac{S_{z_x}}{\bar{X}}$, $C_{yx} = \frac{S_{yx}}{\bar{Y}\bar{X}} = \rho_{yx}C_yC_x$, $C_{yz_x} = \frac{S_{yz_x}}{\bar{Y}\bar{Z}_x} = \rho_{yz_x}C_yC_{z_x}$, $C_{xz_x} = \frac{S_{xz_x}}{\bar{X}\bar{Z}_x} = \rho_{xz_x}C_xC_{z_x}$, $R_1 = \frac{\bar{X}}{S_y^2}$ and $R_2 = \frac{\bar{Z}_x}{S_y^2}$.

3. Classical estimators available in the literature

In this part of the paper, we have discussed some already established estimation procedures of population variance S_y^2 for study variable along with their properties under the various strategies of random non-response given below.

Strategy I: When population mean $\bar{X}$ and sample mean $\bar{x}$ of auxiliary variable x are used. In this case, we assume that the information is missing at random only for y, and the population mean $\bar{X}$ is known.

Strategy II: When population mean $\bar{X}$ and respondent mean $\bar{x}^*$ of auxiliary variable x are used. In this case, we assume that the information is missing at random for y as well as the corresponding units of x, and the population mean $\bar{X}$ is known.

Strategy III: When sample mean $\bar{x}$ and respondent mean $\bar{x}^*$ of auxiliary variable x are used. In this case, we assume that the information is missing at random only for y, while the information for x is available on all the sampled units, but the population mean $\bar{X}$ is unknown.

The usual estimator of population variance in the case of random non-response, is given by

$$t_0 = s_y^{*2} \tag{2}$$

where $s_y^{*2} = \sum_{i=1}^{n-m}(y_i - \bar{y}^*)^2/(n-m-1)$ is the conditionally unbiased estimator of population variance respectively and where $\bar{y}^* = \sum_{i=1}^{n-m} y_i/(n-m)$ is the respondent mean of y (*see* Singh and Joarder, 1998).

The variance of the estimator t_0 to the first order approximation is given by

$$V(t_0) = S_y^4 f_2 \lambda_{400}^* \tag{3}$$

The classical ratio estimators on the lines of Upadhyay and Singh (2001) under the Strategies I, II and III, are respectively defined as

$$t_{m_1} = s_y^{*2}\left(\frac{\bar{X}}{\bar{x}}\right) \tag{4}$$

$$t_{m_2} = s_y^{*2}\left(\frac{\bar{X}}{\bar{x}^*}\right) \tag{5}$$

$$t_{m_3} = s_y^{*2}\left(\frac{\bar{x}}{\bar{x}^*}\right) \tag{6}$$

The MSEs of the estimators $t_{m_i}(i = 1,2,3)$ to the first order approximation are given by

$$MSE(t_{m_i}) = S_y^4[f_2\lambda_{400}^* + f_i C_x(C_x - 2\lambda_{210})] \tag{7}$$

Following Das (1978) and Upadhyay and Singh (2001), we define two sets of difference and ratio estimators under three strategies of random non-response, respectively given by

difference-type estimators:

$$t_{m_{d1}} = s_y^{*2} + k_1^*(\bar{X} - \bar{x}) \tag{8}$$

$$t_{m_{d2}} = s_y^{*2} + k_2^*(\bar{X} - \bar{x}^*) \tag{9}$$

$$t_{m_{d3}} = s_y^{*2} + k_3^*(\bar{x} - \bar{x}^*) \tag{10}$$

Ratio type estimators:

$$t_{m_{r1}} = s_y^{*2}\left(\frac{\bar{X}}{\bar{x}}\right)^{\pi_1^*} \tag{11}$$

$$t_{m_{r2}} = s_y^{*2}\left(\frac{\bar{X}}{\bar{x}^*}\right)^{\pi_2^*} \tag{12}$$

$$t_{m_{r3}} = s_y^{*2}\left(\frac{\bar{x}}{\bar{x}^*}\right)^{\pi_3^*} \tag{13}$$

where k_i^* and π_i^* (i=1,2,3) are the unknown constants which are to be chosen such that the variances of the respective estimators is minimum.

The minimum MSEs of the existing estimators $t_{md_i}(i = 1,2,3)$ and t_{mr_i}, are respectively given by

$$min.MSE(t_{md_i}) = min.MSE(t_{mr_i}) = S_y^4\left[f_2\lambda_{400}^* - f_i\lambda_{210}^2\right]. \tag{14}$$

The optimum values of k_i^* and π_i^* (i=1,2,3) are given by

$$k_{1(opt)}^* = k_{2(opt)}^* = k_{3(opt)}^* = S_y^2\frac{\lambda_{210}}{S_x} \qquad \text{and} \qquad \pi_{1(opt)}^* = \pi_{2(opt)}^* = \pi_{3(opt)}^* = \frac{\lambda_{210}}{S_x}.$$

In line with Singh et al. (1988), the optimal version of the estimators $t_{md_i}(i = 1,2,3)$ in three strategies are given by

$$t_{mD_1} = \kappa_1^* s_y^{*2} + d_1^*(\bar{X} - \bar{x}) \tag{15}$$

$$t_{mD_2} = \kappa_2^* s_y^{*2} + d_2^*(\bar{X} - \bar{x}^*) \tag{16}$$

$$t_{mD_3} = \kappa_3^* s_y^{*2} + d_3^*(\bar{x} - \bar{x}^*) \tag{17}$$

where κ_i^* and d_i^* $(i = 1,2,3)$ are the arbitrary constants to be chosen such that the MSEs of the respective estimators become minimum.

The optimum values of κ_i^* and d_i^* (i=1,2,3) are given by

$$\kappa_{i(opt)}^* = \frac{1}{\left[1 + f_2\lambda_{400}^* - f_i\lambda_{210}^2\right]} \qquad \text{and} \qquad d_{i(opt)} = S_y^2\frac{\lambda_{210}}{S_x}\kappa_{i(opt)}^*$$

The minimum MSEs of the estimators $t_{mD_i}(i = 1,2,3)$ are given by

$$min.MSE(t_{mD_i}) = \frac{S_y^4 MSE(t_{md_i})}{S_y^4 + MSE(t_{md_i})} \tag{18}$$

Thus the estimators $t_{mD_i}(i=1,2,3)$ are improvement over t_{md_i} as well as t_{mr_i} in the corresponding strategies.

4. Proposed estimators

Here, we have suggested various novel difference-type estimators of finite population variance along with their optimal variants in case of random non-response by employing the rank of an auxiliary variable. The estimators are formulated in three different cases of random non-response, which are discussed in the following three strategies.

Strategy I: When the population means $(\bar{X},\bar{Z}_x)$ and corresponding estimates $(\bar{x},\bar{z}_x)$ from the sample are used.

Strategy II: When the population means $(\bar{X},\bar{Z}_x)$ and corresponding estimates $(\bar{x}^*,\bar{z}_x^*)$ from the respondents are used.

Strategy III: When the sample means $(\bar{x},\bar{z}_x)$ and corresponding estimates $(\bar{x}^*,\bar{z}_x^*)$ from the respondents are used.

The proposed estimators under the *Strategy I*, *Strategy II* and *Strategy III* are given as

$$t_{m_{dd1}} = s_y^{*2} + \phi_1^*(\bar{X}-\bar{x}) + \varphi_1^*(\bar{Z}_x-\bar{z}_x) \tag{19}$$

$$t_{m_{dd2}} = s_y^{*2} + \phi_2^*(\bar{X}-\bar{x}^*) + \varphi_2^*(\bar{Z}_x-\bar{z}_x^*) \tag{20}$$

$$t_{m_{dd3}} = s_y^{*2} + \phi_3^*(\bar{x}-\bar{x}^*) + \varphi_3^*(\bar{z}_x-\bar{z}_x^*) \tag{21}$$

where ϕ_i^* and φ_i^* (i=1,2,3) are the unknown constants to be chosen suitably. The optimum values of these constants are given later in *Appendix*.

The optimal versions of the proposed estimators $t_{m_{dd1}}$, $t_{m_{dd2}}$ and $t_{m_{dd3}}$ are respectively given by

$$t_{m_{dD1}} = \alpha_1^* s_y^{*2} + \beta_1^*(\bar{X}-\bar{x}) + \gamma_1^*(\bar{Z}_x-\bar{z}_x) \tag{22}$$

$$t_{m_{dD2}} = \alpha_2^* s_y^{*2} + \beta_2^*(\bar{X}-\bar{x}^*) + \gamma_2^*(\bar{Z}_x-\bar{z}_x^*) \tag{23}$$

$$t_{m_{dD3}} = \alpha_3^* s_y^{*2} + \beta_3^*(\bar{x}-\bar{x}^*) + \gamma_3^*(\bar{z}_x-\bar{z}_x^*) \tag{24}$$

where $\alpha_i^*(i=1,2,3)$, β_i^* and γ_i^* are the arbitrary chosen constants. The optimum values of these constants are given later in *Appendix*.

The difference-type estimators discussed in (3.7)-(3.9) and (3.14)-(3.16), are the special cases of the proposed difference-type estimators (4.1)-(4.3) and (4.4)-(4.6) respectively in the corresponding strategies.

Theorem 1: The biases of the estimators $t_{m_{ddi}}(i=1,2,3)$ and $t_{m_{dDi}}$ to the first degree of approximations are given by

$$B(t_{m_{ddi}}) = 0 \tag{25}$$

and

$$B(t_{m_{dDi}}) = S_y^2(\alpha_i^*-1) \tag{26}$$

Proof: *See* Appendix

Theorem 2: The minimum MSEs of the estimators $t_{m_{ddi}}(i=1,2,3)$ and $t_{m_{dDi}}$ to the first degree of approximations are given by

$$min.MSE(t_{m_{ddi}}) = S_y^4(f_2\lambda_{400}^* - f_i R_{y.xz_x}^{*2}) \tag{27}$$

and

$$min.MSE(t_{m_{dDi}}) = \frac{S_y^4 min.MSE(t_{m_{ddi}})}{S_y^4 + min.MSE(t_{m_{ddi}})} \tag{28}$$

where $R_{y.xz_x}^{*2} = \frac{\lambda_{201}^2+\lambda_{210}^2-2\rho_{xz_x}\lambda_{210}\lambda_{201}}{1-\rho_{xz_x}^2}$.

Proof: *See* Appendix.

5. Optimal situations of proposed estimators

The optimal situations of the proposed estimators $t_{m_{ddi}}(i=1,2,3)$ and $t_{m_{dDi}}$ at which their variances are minimum are given as

$$t_{m_{dd1}}^* = s_y^{*2} + \phi_{1(opt)}^*(\bar{X}-\bar{x}) + \varphi_{1(opt)}^*(\bar{Z}_x - \bar{z}_x) \tag{29}$$

$$t_{m_{dd2}}^* = s_y^{*2} + \phi_{2(opt)}^*(\bar{X}-\bar{x}^*) + \varphi_{2(opt)}^*(\bar{Z}_x - \bar{z}_x^*) \tag{30}$$

$$t_{m_{dd3}}^* = s_y^{*2} + \phi_{3(opt)}^*(\bar{x}-\bar{x}^*) + \varphi_{3(opt)}^*(\bar{z}_x - \bar{z}_x^*) \tag{31}$$

$$t_{m_{dD1}}^* = \alpha_{1(opt)}^* s_y^{*2} + \beta_{1(opt)}^*(\bar{X}-\bar{x}) + \gamma_{1(opt)}^*(\bar{Z}_x - \bar{z}_x) \tag{32}$$

$$t_{m_{dD2}}^* = \alpha_{2(opt)}^* s_y^{*2} + \beta_{2(opt)}^*(\bar{X}-\bar{x}^*) + \gamma_{2(opt)}^*(\bar{Z}_x - \bar{z}_x^*) \tag{33}$$

and

$$t_{m_{dD3}}^* = \alpha_{3(opt)}^* s_y^{*2} + \beta_{3(opt)}^*(\bar{x}-\bar{x}^*) + \gamma_{3(opt)}^*(\bar{z}_x - \bar{z}_x^*) \tag{34}$$

where

$$\phi_{1(opt)}^* = \phi_{2(opt)}^* = \phi_{3(opt)}^* = \frac{\lambda_{210}-\lambda_{201}\rho_{xz_x}}{1-\rho_{xz_x}^2}\frac{S_y^2}{S_x}, \tag{35}$$

$$\varphi_{1(opt)}^* = \varphi_{2(opt)}^* = \varphi_{3(opt)}^* = \frac{\lambda_{201}-\lambda_{210}\rho_{xz_x}}{1-\rho_{xz_x}^2}\frac{S_y^2}{S_{z_x}}, \tag{36}$$

$$\alpha_{i(opt)}^* = \frac{1}{1+(f_2\lambda_{400}^* - f_i R_{y.xz_x}^{*2})}; \qquad (i=1,2,3) \tag{37}$$

$$\beta_{i(opt)}^* = \alpha_{i(opt)}^* \phi_{i(opt)}^*; \qquad (i=1,2,3) \tag{38}$$

$$\gamma_{i(opt)}^* = \alpha_{i(opt)}^* \varphi_{i(opt)}^*; \qquad (i=1,2,3) \tag{39}$$

6. Comparative study

We have judged the merits of the estimators based on real and simulated data under an empirical study and a simulation study given as follows.

6.1. Empirical study

To exhibit the performances of the estimators, we have chosen 8 populations given as follows.

Population-1: [Cochran (1977); p-182]:
Y: Number of placebo children.
X: Number of paralytic polio cases in the placebo group.
The description of the required parameters is as follows:
$N = 34$, $\bar{Y} = 2.588234$, $\bar{X} = 4.923528$, $C_y = 1.233279$, $C_x = 1.023332$, $C_z = 0.5687384$, $\rho_{yx} = 0.7328234$, $\rho_{yz_x} = 0.6571886$, $\rho_{xz_x} = 0.8165118$. Here, $n = 12$ and $m = 8$.

Population-2: [Anderson (1958); p-110]:
Y: Sepal Width of Iris flower. X: Sepal Length of Iris flower.
The description for this data is as follows:
$N = 150$, $\bar{Y} = 3.057334$, $\bar{X} = 5.843334$, $C_y = 0.1425641$, $C_x = 0.1417114$, $C_z = 0.5749112$, $\rho_{yx} = -0.1175699$, $\rho_{yz_x} = -0.1404247$, $\rho_{xz_x} = 0.9871834$. Here, $n = 50$ and $m = 12$.

Population-3: [Madala (1992); p-108]:
Y: salary (thousands of dollars)
X: years of experience (defined as years since receiving Ph D).
The description of the data is as follows:
$N = 32$, $\bar{Y} = 47.37813$, $\bar{X} = 18.376$, $C_y = 0.1819514$, $C_x = 0.4548527$, $C_z = 0.5677533$, $\rho_{yx} = 0.4245115$, $\rho_{yz_x} = 0.3367752$, $\rho_{xz_x} = 0.9447146$. Here, $n = 12$ and $m = 8$.

Population-4: [Anderson (1958); p-110]:
Y: Sepal Length. X: Petal Length.
The details of the required parameters are as follows:
$N = 150$, $\bar{Y} = 5.843334$, $\bar{X} = 3.7581$, $C_y = 0.1417112$, $C_x = 0.4697442$, $C_z = 0.5748254$, $\rho_{yx} = 0.8717537$, $\rho_{yz_x} = 0.8792952$, $\rho_{xz_x} = 0.9684332$. Here, $n = 50$ and $m = 10$.

Population-5: [Satici and Kadilar (2011)]:
Y: number of successful students. X: number of teachers.
The summary of the data is:
$N = 261$, $\bar{Y} = 222.5825$, $\bar{X} = 306.44831$, $C_y = 1.86541$, $C_x = 1.7596$, $C_z = 0.576241$, $\rho_{yx} = 0.9706$, $\rho_{yz_x} = 0.6372$, $\rho_{xz_x} = 0.6264$. Here, $n = 90$ and $m = 70$.

Population-6: [Singh (2003); p-1111]:
Y: amount (in \$000) of non-real estate farm loans in different states during 1997.
X: amount (in \$000) of real estate farm loans in different states during 1997.
The summary of the data is:
$N = 50$, $\bar{Y} = 878.1627$, $\bar{X} = 555.4346$, $C_y = 1.235166$, $C_x = 1.052917$, $C_z = 0.571663$, $\rho_{yx} = 0.8039$, $\rho_{yz_x} = 0.7462$, $\rho_{xz_x} = 0.9237$. Here, $n = 20$ and $m = 8$.

Population-7: [Anderson (1958); p-110]:
Y: Petal Length of Iris setosa. X: Sepal Length of Iris setosa.
The description of the data is given as:
$N = 50$, $\bar{Y} = 1.463$, $\bar{X} = 5.0061$, $C_y = 0.1187853$, $C_x = 0.07041345$, $C_z = 0.5683722$, $\rho_{yx} = 0.2671757$, $\rho_{yz_x} = 0.2687847$, $\rho_{xz_x} = 0.9797012$. Here, $n = 20$ and $m = 5$.

Population-8: [Mc Nill (1977)]:
Y: speed of cars. X: distances taken to stop.
The details of parameters for this data are as follows:
$N = 50$, $\bar{Y} = 15.4$, $\bar{X} = 42.981$, $C_y = 0.3433534$, $C_x = 0.5995668$, $C_z = 0.571306$, $\rho_{yx} = 0.8068948$, $\rho_{yz_x} = 0.8341367$, $\rho_{xz_x} = 0.9605414$. Here, $n = 20$ and $m = 5$.
We have estimated the percentage relative efficiencies (PREs) of the different estimators with respect to usual unbiased estimator s_y^{*2}. To compute the PREs of different estimators $(t_\bullet)$ we use the formula, given by

$$PRE(t_\bullet) = \frac{V(s_y^{*2})}{MSE(t_\bullet)} \times 100.$$

The results are shown in Table 1.

6.2. Simulation study

We have conducted a simulation study based on artificially generated data using the R programming language. To generate the data, we have considered two statistical probability distributions: (i) Gamma distribution and (ii) Normal distribution, where the performances of the estimators are appraised for different amounts of correlation, 0.6-0.9, with a step of 0.1 between the study variable and auxiliary variable. The distributions are discussed below.

Gamma distribution

Following Singh and Horn (1998), we use the transformations to generate the study and auxiliary variables, which are given as follows:

$$y_i = \mu_y + \sqrt{(1-\rho_{yx}^2)}y_i^* + \rho_{yx}\frac{S_y}{S_x}x_i^* \tag{40}$$

Table 1. *PREs* of the various estimators with respect to s_y^{*2}.

Estimators	Populations							
	1	2	3	4	5	6	7	8
	Strategy I							
t_0	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
t_{m_1}	125.37	97.34	89.07	94.72	159.79	123.98	99.26	88.34
t_{md_1}=t_{mr_1}	125.54	101.80	101.07	100.73	204.03	124.03	100.88	100.16
t_{mD_1}	157.89	106.08	116.76	103.34	222.89	142.70	113.92	106.73
t_{mdd_1}	128.33	103.10	109.51	105.07	230.60	129.88	101.06	113.04
t_{mdD_1}	160.68	107.38	125.19	107.67	249.45	148.55	114.11	119.61
	Strategy II							
t_{m_2}	160.83	96.10	81.09	92.84	217.26	151.88	98.83	82.63
t_{md_2}=t_{mr_2}	161.36	102.70	102.06	101.02	377.99	152.01	101.41	100.25
t_{mD_2}	193.71	106.98	117.75	103.62	396.85	170.69	114.45	106.82
t_{mdd_2}	170.25	104.67	119.75	107.15	546.07	168.44	101.71	122.52
t_{mdD_2}	202.60	108.95	135.44	109.76	564.93	187.12	114.75	129.09
	Strategy III							
t_{m_3}	121.34	98.69	90.05	97.90	119.84	117.40	99.56	92.74
t_{md_3}=t_{mr_3}	121.48	100.87	100.97	100.28	129.13	117.43	100.52	100.09
t_{mD_3}	153.83	105.15	116.65	102.88	147.98	136.11	113.56	106.66
t_{mdd_3}	123.74	101.48	108.47	101.89	133.43	121.40	100.63	107.35
t_{mdD_3}	156.09	105.76	124.16	104.63	152.28	140.07	113.67	113.92

and

$$x_i = \mu_x + x_i^* \tag{41}$$

where $y_i^* \sim G(a_y, b_y)$ and $x_i^* \sim G(a_x, b_x)$ are the independent gamma variables generated using R programming language. Here, (a_y, b_y) and (a_x, b_x) are the shape and scale parameters for y_i^* and x_i^*. Moreover, $\mu_y = a_y b_y$, $\mu_x = a_x b_x$, $S_y^2 = a_y b_y^2$ and $S_x^2 = a_x b_x^2$. The size of data is $N = 5000$ and sample size $n = 1500$. We have taken $m = 300$.

Normal distribution

We have considered the bivariate normal distribution as $(Y,X) \sim N(9,9,\rho_{yx},20^2,20^2)$ for the correlations (ρ_{yx}). We have chosen $N = 5000$, $n = 1500$ and $m = 300$.
The complete simulation process is as follows. Draw a sample of size n focusing on a variable of interest which is properly correlated with an auxiliary characteristic from a population of size N. Set the value of m and drop m units randomly from the sample. Now, compute the relevant statistics based on the information available on $(n-m)$ units. Repeat the whole procedure 50,000 times.

We have computed the simulated percentage relative efficiencies (PREs) of different estimators considered in this study with respect to usual estimator s_y^{*2} based on their simulated MSE values by using the formulae given as

$$V(s_y^{*2})_{simulated} = \frac{1}{50,000} \sum_{j=1}^{50,000} ((s_y^{*2})_j - S_y^2)^2;$$

$$MSE(t_\bullet)_{simulated} = \frac{1}{50,000} \sum_{j=1}^{50,000} ((t_\bullet)_j - S_y^2)^2; \quad PRE(t_\bullet)_{simulated} = \frac{V(s_y^{*2})_{simulated}}{MSE(t_\bullet)_{simulated}} \times 100.$$

The results are shown in Table 2.

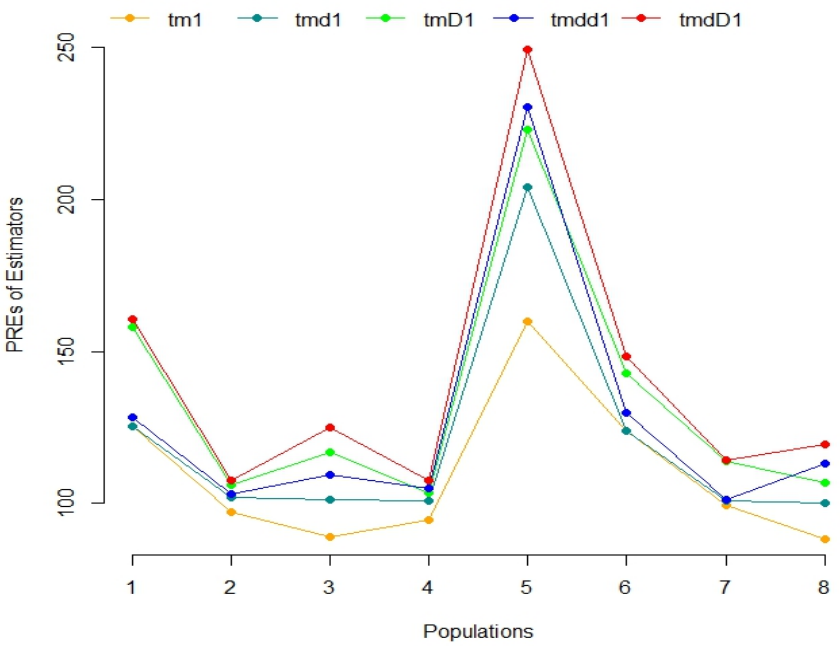

Figure 1. Comparison of PREs of different estimators for Populations 1-8 under ***Strategy-I***

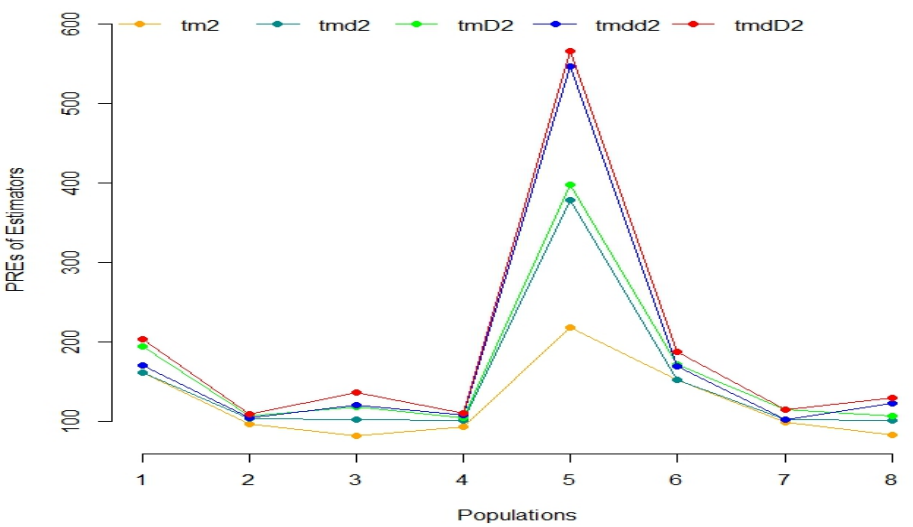

Figure 2. Comparison of PREs of different estimators for Populations 1-8 under ***Strategy-II***

Table 2. *PREs* of the different estimators with respect to s_y^{*2}.

	Gamma distribution				*Normal distribution*			
	Value of ρ_{yx}				Value of ρ_{yx}			
Estimators	0.6	0.7	0.8	0.9	0.6	0.7	0.8	0.9
Strategy I								
t_0	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
t_{m_1}	100.02	100.02	100.02	100.03	80.66	88.57	100.73	101.73
t_{md_1}	107.55	108.20	110.76	110.90	105.02	108.57	110.18	111.28
t_{mr_1}	107.56	108.23	110.77	110.89	105.03	108.59	110.19	111.29
t_{mD_1}	107.72	108.38	110.94	111.06	105.16	108.72	110.35	111.45
t_{mdd_1}	140.53	150.67	170.91	181.14	124.38	144.89	167.84	183.15
t_{mdD_1}	140.55	150.80	171.09	181.31	124.52	145.04	168.01	183.32
Strategy II								
t_{m_2}	100.03	100.03	100.03	100.03	83.79	97.62	101.68	102.32
t_{md_2}	110.50	111.41	115.22	115.38	106.94	112.00	114.34	115.96
t_{mr_2}	110.55	111.52	115.20	115.38	106.95	112.03	114.36	115.97
t_{mD_2}	110.71	111.66	115.37	115.55	107.08	112.15	114.51	116.12
t_{mdd_2}	164.09	183.90	228.93	255.13	136.26	172.59	221.56	260.62
t_{mdD_2}	165.25	184.11	229.11	255.30	136.40	172.74	221.73	260.78
Strategy III								
t_{m_3}	100.01	100.01	100.01	100.01	77.82	86.72	100.24	101.60
t_{md_3}	102.55	102.80	103.60	103.64	101.72	102.90	103.42	103.76
t_{mr_3}	102.58	102.80	103.60	103.64	101.74	102.92	103.43	103.77
t_{mD_3}	102.74	102.94	103.78	103.81	101.88	103.05	103.59	103.93
t_{mdd_3}	111.46	113.66	117.41	119.06	107.53	112.45	116.89	119.37
t_{mdD_3}	111.62	113.80	117.59	119.23	107.68	112.61	117.06	119.54

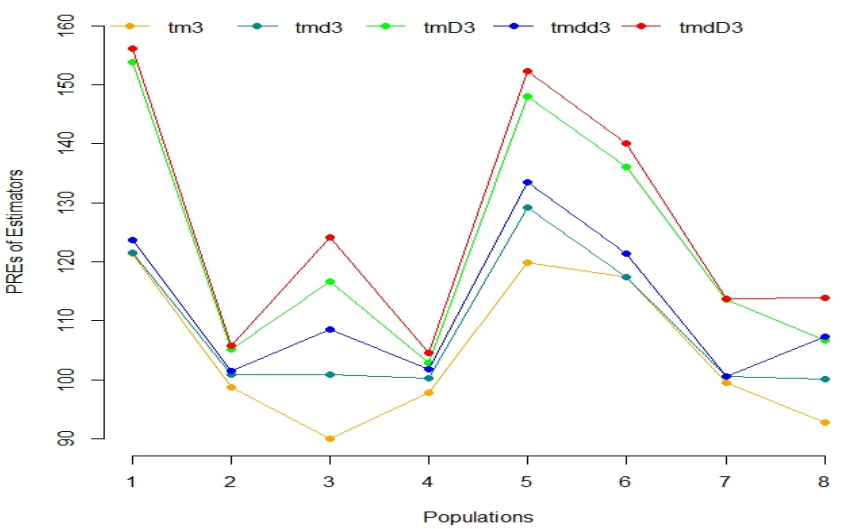

Figure 3. Comparison of PREs of different estimators for Populations 1-8 under ***Strategy-III***

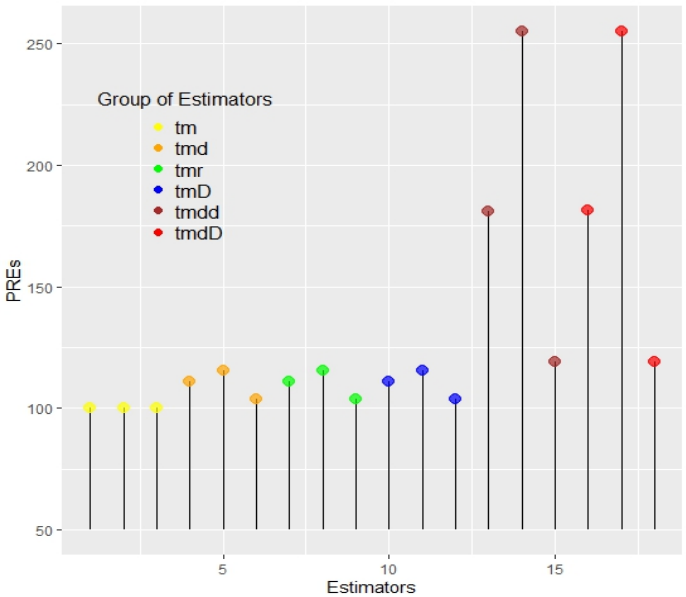

Figure 4. Comparison of PREs of different estimators in all the strategies based on Gamma distribution when $\rho_{yx} = 0.9$

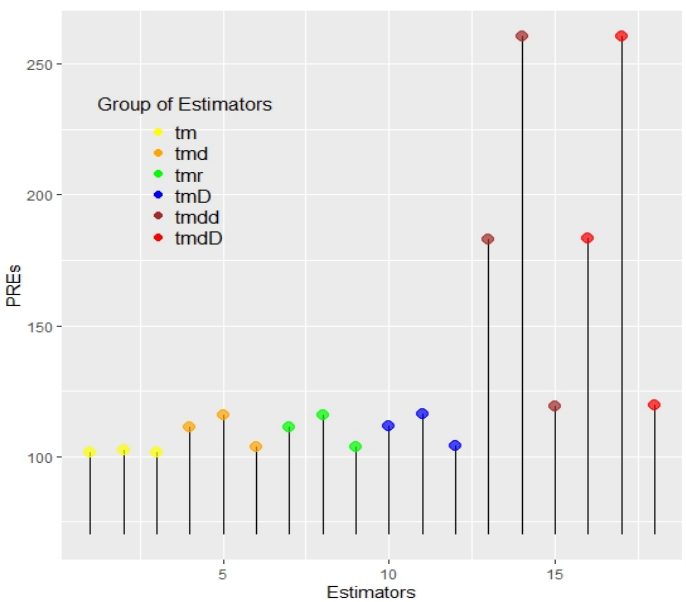

Figure 5. Comparison of PREs of different estimators in all the strategies based on Normal distribution when $\rho_{yx} = 0.9$

Interpretation of the results:
From Table 1, we report that:

(i) All the estimators, excluding the usual ratio estimators $t_{m_i}(i = 1,2,3)$, perform well in all the Populations 1-8 under each strategy. We see that the performances of usual ratio estimators t_{m_i} are good in Populations 1, 5, & 6, where the condition $\lambda^*_{210} > \frac{C_x}{2}$ holds. On the other hand, the performances of t_{m_i} are poor in Populations 2, 3, 4, 7 & 8 because the condition $\lambda^*_{210} > \frac{C_x}{2}$ does not hold in these populations.

(ii) The proposed estimators $t_{mdd_i}(i = 1,2,3)$ (constructed using the rank of an auxiliary variable) are paralleling more efficient than the existing estimators t_{md_i} or t_{mr_i}, respectively, which are formulated using the original information on an auxiliary variable. Thus, it is remarkable that the efficiency of usual difference-type estimators may be increased just by introducing the dual use of an auxiliary variable.

(iii) Similarly, the improved versions $t_{mdD_i}(i = 1,2,3)$ of the proposed estimators t_{mdd_i} also show their appreciable behaviors over the existing estimators t_{mD_i} respectively in terms of gain in percentage relative efficiencies. Thus, the efficiency of the optimal version of the usual difference estimator may also be increased using the rank of an auxiliary variable.

(iv) The proposed optimal estimators $t_{mdD_i}(i = 1,2,3)$ are the most efficient estimators among all the estimators discussed in Table 1 in the corresponding strategies.

(v) We see that the performances of all the estimators under *Strategy II* are superior to those of *Strategy I* and *Strategy III*. Thus, *Strategy II* is reasonably preferable over *Strategy I* and *Strategy III* when the information at different levels is available on an auxiliary variable.

(vi) In view of the arguments (iv) and (v), we can easily say that the efficiency of the proposed estimator t_{mdD_2} is highest among all the estimators considered in this study, which is evidently demonstrated in Table 1.

(vii) The merits of the proposed estimators based on the comparative results in Table 1 can be clearly visualized in Figures 1, 2, and 3 for Strategies I, II, and III, respectively.

Similar conclusions (as discussed above) for the proposed estimators can be drawn from the results in Table 2, which is based on the simulation study. An instant view of the results in Table 2 for gamma and normal distributions at the correlation value 0.9 can be obtained from the two scatter plots, which are displayed in Figures 4 and 5. Similar plots can be obtained for the rest of the correlation values for both distributions.
In Table 3 *(given in Appendix)*, we have demonstrated the values of estimates of all the estimators at their optimum situations considered in this study. These estimated values are based on the sample drawn from Population 6. It is observed that the estimates obtained

from proposed estimators $t_{mdd_i}(i = 1,2,3)$ and t_{mdD_i} based on the selected sample are very close to the true value of the parameter.
On the basis of the above arguments, we can easily say that the use of the rank of an auxiliary variable is capable enough to enhance the efficiency of the estimators to the next level, as the proposed estimators present considerable improvements over other existing estimators in estimating the population variance in the presence of random non-response under both empirical and simulation studies. Therefore, this comparative study may be appreciably extrapolated in general practice.

7. Conclusions

From the aforementioned results and discussions, it may be concluded that the efficiency of usual difference-type estimators may be easily increased without using any new (more than one) auxiliary variable, just by introducing the dual of an auxiliary variable. The proposed estimators, which are constructed using the rank (dual) of an auxiliary variable, are capable of providing increased efficiency in three different strategies of random non-response. The proposed difference-type estimators show better gain in terms of percentage relative efficiencies over the existing relevant estimators considered in this study for the corresponding situations of random non-response. The performances of the optimal versions of the proposed difference-type estimators are superior to all other estimators discussed in this study in respective situations at various amounts of correlations.

Hence, looking at their charming behaviors, they may be encouragingly recommended for real-life situations when faced with missing-at-random problems. The strengths of the proposed model are as follows: it may provide efficient results for both positive and negative correlations, it may be fruitfully appreciable for highly positively correlated datasets, and it may be highly preferable when the information is available only on a single auxiliary variable. On the other hand, the weakness may lie in the fact that the proposed model may not give an attractive result for the low-correlated datasets, and it may not be preferable when the information on multi-auxiliary variables is available. For future research, the present work may be extended to various sampling schemes such as successive sampling, two-phase sampling, stratified sampling, etc, for the estimation of mean, variance, and other population parameters.

Acknowledgement

Authors are highly thankful to Vellore Institute of Technology Vellore for providing 'VIT SEED GRANT (RGEMS)-Sanction Order No. SG20230035' for carrying the present research work.

Conflict of interest

There is no conflict of interest associated with the present article.

References

Ahmad, S., Adichwal, N. K., Aamir, M., Shabbir, J., Alsadat, N., Elgarhy, M. and Ahmad, H., (2023). An enhanced estimator of finite population variance using two auxiliary variables under simple random sampling. *Scientific Reports*, 13(1), 21444.

Alam, S., Shabbir, J., (2020). Calibration estimation of mean by using double use of auxiliary information. *Commun. Stat. Simul. Comput.*, pp. 1–19.

Almulhim, F. A., Aljohani, H. M., Aldallal, R., Mustafa, M. S., Alsolmi, M. M., Elshenawy, A. and Alrashidi, A., (2024). Estimation of finite population mean using dual auxiliary information under non-response with simple random sampling. *Alexandria Engineering Journal*, 100, pp. 286–299.

Anderson, T. W., (1958). An Introduction to Multivariate Statistical Analysis. New York: *Wiley Series in Probability and Statistics*.

Belili, M. C., Alshangiti, A. M., Gemeay, A. M., Zeghdoudi, H., Karakaya, K., Bakr, M. E., Balogun, O. S., Atchadé, M. N. and Hussam, E., (2023). Two-parameter family of distributions: Properties, estimation, and applications. *AIP Advances*, 13(10), Available from: https://doi.org/10.1063/5.0173532.

Bhushan S., Pandey, A. P., (2021). Optimal estimation of population variance in the presence of random non-response using simulation approach. *J Stat Comput Simul*. Available from: https://doi.org/10.1080/00949655.2021.1948547.

Bhushan, S., Pandey, S., (2025). Optimal random non-response framework for mean estimation on current occasion. Commun. *Stat. - Theory Methods*, 54(4), pp. 1205–1231.

Cochran, W. G., (1977). Sampling techniques, 3rd ed. New York: *John Wiley and Sons*.

Daraz, U., Wu, J., Agustiana and D., Emam, W., (2025). Finite population variance estimation using Monte Carlo simulation and real life application. *Symmetry*, 17(1), 84.

Das, A. K., (1978). Use of auxiliary information in estimating the finite population variance. *Sankhya*, c, 40, pp. 139–148.

Das A. K., Tripathi T. P., (1978). Use of auxiliary information in estimating the finite population variance. *Sankhya* 34, 19.

Hussain, I., Haq, A., (2019). A New Family of Estimators for Population Mean with Dual Use of the Auxiliary Information. *J. Stat. Theory Pract.*, 13(1), 23.

Irfan, M., Javed, M., Bhatti, S. H., Raza, M. A. and Ahmad, T., (2020). Almost unbiased optimum estimators for population mean using dual auxiliary information. *Journal of King Saud University-Science*, 32(6), pp. 2835–2844.

Isaki, C. T., (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78(381), pp. 117–123.

Javed, S., Masood, S. and Shokri, A., (2023). Generalized Class of Finite Population Variance in the Presence of Random Non response Using Simulation Approach. *Complexity*, 2023(1), 6643435.

Khodja, N., Gemeay, A. M., Zeghdoudi, H., Karakaya, K., Alshangiti, A. M., Bakr, M. E., Balogun, O. S., Muse, A. H. and Hussam, E., (2023). Modeling voltage real data set by a new version of Lindley distribution. *IEEE Access*, 11, pp. 67220–67229.

Kumar, S., (2014). Variance estimation in presence of random non-response. *Journal of Reliability and Statistical Studies*, pp. 65–70.

Maddala, G. S., (1992). Introduction to econometrics (2nd Edit.). New York: *Macmillan*.

McNeil, D. R., (1977) Interactive Data Analysis. *Wiley*.

Rubin, D. B., (1976). Inference and missing data. *Biometrika*, 63(3), pp. 581–592.

Satici, E., Kadilar, C., (2011). Ratio Estimator for the Population Mean at the Current Occasion in the Presence of Non-Response in Successive Sampling. *Hacettepe Journal of Mathematics and Statistics*, 40(1), pp. 115–124.

Sharma, A. K., Singh, A. K., (2020). Estimation of population variance under an imputation method in two-phase sampling. Proceedings of the National Academy of Sciences, India Section A: *Physical Sciences*, 90, pp. 185–191.

Singh, G. N., Usman, M., (2022). Some Optimal Estimators of Finite Population Variance Using Dual of Auxiliary Variable in the Presence of Random Non-Response. *Journal of Statistical Theory and Practice*, 16(4), 65.

Singh, H. P., Upadhyaya, L. N. and Namjoshi, U. D., (1988). Estimation of finite population variance. *Current Science*, pp. 1331–1334.

Singh, J., Pandey, B.N. and Hirano, K., (1973). On the utilization of a known coefficient of kurtosis in the estimation procedure of variance. *Ann Inst Stat Math*, 25, pp. 51–55.

Singh, S., (2003). Advanced Sampling Theory With Applications. (Vol. 2). *Springer Science and Business Media*.

Singh, S., Horn, S., (1998). An alternative estimator in multi-character surveys. *Metrika*, pp. 99–107.

Singh, S., Joarder, A. H., (1998). Estimation of finite population variance using random non-response in survey sampling. *Metrika*, 47(1), pp. 241–249.

Singh S., Joarder A. H. and Tracy D. S., (2000). Regression type estimators for random non-response in survey. Sampling. *Statistica LX 1*, pp. 39–43.

Upadhyaya, L. N., Singh, H. P. (2001). Estimation of the population standard deviation using auxiliary information. *Am. J. Math. Manag. Sci*, 21(3-4), pp. 345–358.

Yaqub, M., Shabbir, J., and Gupta, S. N., (2017). Estimation of population mean based on dual use of auxiliary information in non-response. *Communications in Statistics-Theory and Methods*, 46(24), pp. 12130-12151.

Yasmeen, U., Noor-ul-Amin, M. and Hanif, M., (2019). Exponential estimators of finite population variance using transformed auxiliary variables. Proceedings of the National Academy of Sciences, India Section A: *Physical Sciences*, 89(1), pp. 185–191.

Zaman, T., Bulut, H., (2022). A new class of robust ratio estimators for finite population variance. *Scientia Iranica.*

Zaman, T., Bulut, H., (2024). A simulation study: Robust ratio double sampling estimator of finite population mean in the presence of outliers. *Scientia Iranica*, 31(15), pp. 1330–1341.

Appendix

Outline of the derivations of Theorem 1 and Theorem 2.

The proposed estimators $t_{mdd_i}(i=1,2,3)$ and t_{mdD_i} under the error transformations can be written as

$$t_{mdd_1} = S_y^2(1+\varepsilon_0) - \phi_1^*\bar{X}\varepsilon_1 - \varphi_1^*\bar{Z}_x\varepsilon_2 \tag{42}$$

$$t_{mdd_2} = S_y^2(1+\varepsilon_0) - \phi_2^*\bar{X}\varepsilon_1^* - \varphi_2^*\bar{Z}_x\varepsilon_2^* \tag{43}$$

$$t_{mdd_3} = S_y^2(1+\varepsilon_0) + \phi_3^*\bar{X}(\varepsilon_1-\varepsilon_1) + \varphi_3^*\bar{Z}_x(\varepsilon_2-\varepsilon_2^*) \tag{44}$$

$$t_{mdD_1} = \alpha_1^* S_y^2(1+\varepsilon_0) - \beta_1^*\bar{X}\varepsilon_1 - \gamma_1^*\bar{Z}_x\varepsilon_2 \tag{45}$$

$$t_{mdD_2} = \alpha_2^* S_y^2(1+\varepsilon_0) - \beta_2^*\bar{X}\varepsilon_1^* - \gamma_2^*\bar{Z}_x\varepsilon_2^* \tag{46}$$

$$t_{mdD_3} = \alpha_3^* S_y^2(1+\varepsilon_0) + \beta_3^*\bar{X}(\varepsilon_1-\varepsilon_1^*) + \gamma_3^*\bar{Z}_x(\varepsilon_2-\varepsilon_2^*) \tag{47}$$

The above equations can be rewritten as

$$tm_{dd_1} - S_y^2 = S_y^2\varepsilon_0 - \phi_1^*\bar{X}\varepsilon_1 - \varphi_1^*\bar{Z}_x\varepsilon_2 \tag{48}$$

$$tm_{dd_2} - S_y^2 = S_y^2\varepsilon_0 - \phi_2^*\bar{X}\varepsilon_1^* - \varphi_2^*\bar{Z}_x\varepsilon_2^* \tag{49}$$

$$tm_{dd_3} - S_y^2 = S_y^2\varepsilon_0 + \phi_3^*\bar{X}(\varepsilon_1^*-\varepsilon_1) + \varphi_3^*\bar{Z}_x(\varepsilon_2^*-\varepsilon_2) \tag{50}$$

$$t_{mdD_1} - S_y^2 = S_y^2\{\alpha_1^*(1+\varepsilon_0)-1\} - \beta_1^*\bar{X}\varepsilon_1 - \gamma_1^*\bar{Z}_x\varepsilon_2 \tag{51}$$

$$t_{mdD_2} - S_y^2 = S_y^2\{\alpha_2^*(1+\varepsilon_0)-1\} - \beta_2^*\bar{X}\varepsilon_1^* - \gamma_2^*\bar{Z}_x\varepsilon_2^* \tag{52}$$

$$t_{mdD_3} - S_y^2 = S_y^2\{\alpha_3^*(1+\varepsilon_0)-1\} + \beta_3^*\bar{X}(\varepsilon_1^*-\varepsilon_1) + \gamma_3^*\bar{Z}_x(\varepsilon_2^*-\varepsilon_2) \tag{53}$$

Taking the expectation of both sides of equations (8.7)-(8.12), we can easily get the biases of the proposed estimators. *Hence the proof of Theorem 1.*

Now, squaring both sides of the above equations and ignoring the terms of errors having power greater than two, we get

$$\begin{aligned}(t_{mdd_1} - S_y^2)^2 =& S_y^4\left[\varepsilon_0^2 + \phi_1^{*2}R_1^2\varepsilon_1^2 + \varphi_1^{*2}R_2^2\varepsilon_2^2 + 2\phi_1^*\varphi_1^*R_1R_2\varepsilon_1\varepsilon_2 - 2\phi_1^*R_1\varepsilon_0\varepsilon_1\right.\\ &\left.-2\varphi_1^*R_2\varepsilon_0\varepsilon_2\right]\end{aligned} \tag{54}$$

$$\begin{aligned}(t_{mdd_2} - S_y^2)^2 =& S_y^4\left[\varepsilon_0^2 + \phi_2^{*2}R_1^2\varepsilon_1^{*2} + \varphi_2^{*2}R_2^2\varepsilon_2^{*2} + 2\phi_2^*\varphi_2^*R_1R_2\varepsilon_1^*\varepsilon_2^* - 2\phi_2^*R_1\varepsilon_0\varepsilon_1^*\right.\\ &\left.-2\varphi_2^*R_2\varepsilon_0\varepsilon_2^*\right]\end{aligned} \tag{55}$$

$$\begin{aligned}(t_{mdd_3} - S_y^2)^2 =& S_y^4\left[\varepsilon_0^2 + \phi_3^{*2}R_1^2(\varepsilon_1^{*2}-\varepsilon_1^2) + \varphi_3^{*2}R_2^2(\varepsilon_2^{*2}-\varepsilon_2^2) + 2\phi_3^*\varphi_3^*R_1R_2(\varepsilon_1^*\varepsilon_2^*-\varepsilon_1\varepsilon_2)\right.\\ &\left.-2\phi_3^*R_1(\varepsilon_0\varepsilon_1^*-\varepsilon_0\varepsilon_1) - 2\varphi_3^*R_2(\varepsilon_0\varepsilon_2^*\varepsilon_0\varepsilon_2)\right]\end{aligned} \tag{56}$$

$$(t_{mdD_1}-S_y^2)^2 = S_y^4\left[\alpha_1^{*2}(1+\varepsilon_0^2+2\varepsilon_0)+\beta_1^{*2}R_1^2\varepsilon_1^2+\gamma_1^{*2}R_2^2\varepsilon_2^2-2\alpha_1^*\beta_1^*R_1(\varepsilon_1+\varepsilon_0\varepsilon_1)\right.$$
$$-2\alpha_1^*\gamma_1^*R_2(\varepsilon_2+\varepsilon_0\varepsilon_2)+2\beta_1^*\gamma_1^*R_1R_2\varepsilon_1\varepsilon_2-2\alpha_1^*(1+\varepsilon_0)$$
$$\left.+2\beta_1^*R_1\varepsilon_1+2\gamma_1^*R_2\varepsilon_2+1\right] \tag{57}$$

$$(t_{mdD_2}-S_y^2)^2 = S_y^4\left[\alpha_2^{*2}(1+\varepsilon_0^2+2\varepsilon_0)+\beta_2^{*2}R_1^2\varepsilon_1^{*2}+\gamma_2^{*2}R_2^2\varepsilon_2^{*2}-2\alpha_2^*\beta_2^*R_1(\varepsilon_1^*+\varepsilon_0\varepsilon_1^*)\right.$$
$$-2\alpha_2^*\gamma_2^{*2}R_2(\varepsilon_2^*+\varepsilon_0\varepsilon_2^*)+2\beta_2^*\gamma_2^*R_1R_2\varepsilon_1^*e_2^*-2\alpha_2^*(1+\varepsilon_0)$$
$$\left.+2\beta_2^*R_1\varepsilon_1^*+2\gamma_2^*R_2\varepsilon_2^*+1\right] \tag{58}$$

$$(t_{mdD_3}-S_y^2)^2 = S_y^4\left[\alpha_3^{*2}(1+\varepsilon_0^2+2\varepsilon_0)+\beta_3^{*2}R_1^2(\varepsilon_1-\varepsilon_1^*)^2+\gamma_3^{*2}R_2^2(\varepsilon_2-\varepsilon_2^*)^2\right.$$
$$-2\alpha_3^*\beta_3^*R_1(1+\varepsilon_0)(\varepsilon_1-e_1^*)-2\alpha_3^*\gamma_3^{*2}R_2^2(1+\varepsilon_0)(\varepsilon_2-\varepsilon_2^*)$$
$$+2\beta_3^*\gamma_3^*R_1R_2(\varepsilon_1-\varepsilon_1^*)(\varepsilon_2-\varepsilon_2^*)-2\alpha_3^*(1+\varepsilon_0)$$
$$\left.+2\beta_3^*R_1(\varepsilon_1-\varepsilon_1^*)+2\gamma_3^*R_2(\varepsilon_2-\varepsilon_2^*)+1\right] \tag{59}$$

Taking expectations of both sides of equations (8.13)-(8.18), we can easily get the MSEs of the proposed estimators to the first order of approximations, are given as

$$MSE(t_{mdd_i}) = S_y^4\left[f_2\lambda_{400}^*+\phi_i^{*2}R_1^2f_iC_x^2+\varphi_i^{*2}R_2^2f_iC_{z_x}^2+2\phi_i^*\varphi_i^*R_1R_2f_iC_{xz_x}\right.$$
$$\left.-2\phi_i^*R_1f_i\lambda_{210}^*C_x-2\varphi_i^*R_2f_i\lambda_{201}^*C_{z_x}\right] \tag{60}$$

$$MSE(t_{mdD_i}) = S_y^4\left[\alpha_i^{*2}(1+f_2\lambda_{400}^*)+\beta_i^{*2}R_1^2f_iC_x^2+\gamma_1^{*2}R_2^2f_iC_{z_x}^2-2\alpha_i^*\beta_i^*R_1f_i\lambda_{210}^*C_x\right.$$
$$\left.-2\alpha_i^*\gamma_i^*R_2f_i\lambda_{201}^*C_{r_x}+2\beta_1^*\gamma_1^*R_1R_2f_iC_{xz_x}-2\alpha_1^*+1\right] \tag{61}$$

Now, differentiating partially the equations (8.19) and (8.20) with respect to the constants $\phi_i^*(i=1,2,3)$, φ_i^*, α_i^*, β_i^* and γ_i^* and equating the resultant equations to zero then solving them we can easily obtain their optimum values as given in equations (5.7)-(5.11).

Finally, by putting these optimum values in equations (8.19) and (8.20) appropriately, we can easily obtain the minimum MSEs of the proposed estimators as given in equations (4.9) and (4.10). *Hence the proof of Theorem 2.*

Table 3. Estimates of the various estimators based on a sample drawn from **Population 6** where the true value of the parameter is **1176526**

Sample		Respondents			
y	x	y	x	Estimators	Estimates
348.334	408.978	38.067	40.775	t_0	1230451
494.730	639.571	3520.361	1248.761	t_{m_1}	1027631
1692.817	413.777	57.684	139.628	t_{m_2}	1131690
43.229	42.808	440.518	323.028	t_{m_3}	1355048
298.351	756.169	571.487	114.899	t_{mr_1}	1230035
440.518	323.028	43.229	42.808	t_{mr_2}	1230258
197.244	56.908	635.774	870.720	t_{mr_3}	1230674
38.067	40.775	2610.572	2131.048	t_{md_1}	988258.4
571.487	114.899	494.730	639.571	t_{md_2}	1123362
557.656	1045.106	348.334	408.978	t_{md_3}	1365554
848.317	907.700	1372.439	1229.752	t_{mD_1}	858921.0
540.696	939.460	197.244	56.908 13	t_{mD_2}	1000448
3520.361	1248.761			t_{mD_3}	1178178
386.490	100.964			t_{mdd_1}	1161129
1372.439	1229.752			t_{mdd_2}	1186860
3585.406	1337.852			t_{mdd_3}	1256182
57.684	139.628			t_{mdD_1}	1015148
635.774	870.720			t_{mdD_2}	1068391
388.869	553.266			t_{mdD_3}	1088691
2610.572	2131.048				

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 65–84, https://doi.org/10.59139/stattrans-2026-004
Received – 27.12.2023; accepted – 14.07.2025

The impact of Cyber Supply Chain Risk Management on Supply Chain 4.0

Abdellah Sassi[1], **Mohamed Ben Ali**[2], **Oumaima Oullada**[3], **Said Rifai**[4]

Abstract

Cyber Supply Chain Risk Management (CSCRM) is a novel risk management approach with Cyber Security (CS) being its crucial component. In the age of digitalization, CS has become a major concern worldwide. This study investigates the influence of CSCRM on Supply Chain 4.0 (SC 4.0) using a causal model to evaluate the connection between CS, CSCRM, and SC 4.0. The research investigates the link between CS and CSCRM, and between CSCRM and the levers of SC 4.0. The results highlight that CSCRM significantly influences various supply chain activities. The findings show that the integration of CSCRM, supported by CS is essential for improving the performance of SC 4.0. The "Statistical Package for the Social Sciences" was employed after administering a questionnaire to stakeholders in the Moroccan automotive and aeronautic industries.

Key words: Supply Chain 4.0, Cyber Supply Chain Risk Management, Cyber Security, Causal model.

1. Introduction

The shift toward Supply Chain 4.0 requires the integration of Industry 4.0 (I4.0) technologies such as the Internet of Things (IoT), Big Data Analytics (BDA), and Artificial Intelligence (AI), thereby enhancing digitalization while simultaneously

[1] National Higher School of Electricity and Mechanics – ENSEM – Hassan II University of Casablanca – B.P: 8118 Oasis – Casablanca – Morocco. Laboratory of Process, Mechanics, Materials, and Industrial Engineering – LP2MGI – Higher School of Technology of Casablanca – EST, Hassan II University of Casablanca – B.P 8112 Oasis – Casablanca, Morocco. E-mail: abdellah.sassi@ensem.ac.ma. ORCID: https://orcid.org/0009-0000-6798-2879.

[2] Laboratory of Process, Mechanics, Materials, and Industrial Engineering – LP2MGI – Higher School of Technology of Casablanca – EST, Hassan II University of Casablanca – B.P 8112 Oasis – Casablanca, Morocco. E-mail: benali8mohamed@gmail.com. ORCID: https://orcid.org/0000-0002-8615-7935.

[3] National Higher School of Electricity and Mechanics – ENSEM – Hassan II University of Casablanca – B.P: 8118 Oasis – Casablanca – Morocco. Laboratory of Process, Mechanics, Materials, and Industrial Engineering – LP2MGI – Higher School of Technology of Casablanca – EST, Hassan II University of Casablanca – B.P 8112 Oasis – Casablanca, Morocco. E-mail: oumaimaoullada@gmail.com. ORCID: https://orcid.org/0009-0004-6313-5532.

[4] Laboratory of Process, Mechanics, Materials, and Industrial Engineering – LP2MGI – Higher School of Technology of Casablanca – EST, Hassan II University of Casablanca – B.P 8112 Oasis – Casablanca, Morocco. E-mail: dptgmp@gmail.com. ORCID: https://orcid.org/0000-0002-2813-1606.

increasing exposure to cyber risks and global supply chain uncertainties (Foli *et al.*, 2022). As stated by Muller Raschid et al. (2022), principal vulnerabilities include cyber-attacks and lack of cybersecurity awareness among decision-makers (Muller, 2022). Another gap identified in our previous studies highlights the necessity of integrating Cyber Security (CS) and Cyber Supply Chain Risk Management to address privacy and security challenges (Sassi *et al.*, 2024). Notable cyber incidents for example the 2017 WannaCry attack, which used ransomware to hit several businesses (Creazza *et al.*, 2022). This study explores how effectively CSCRM influences the advancement of SC 4.0 within Morocco's automotive and aeronautic industries. The objectives of this research paper include: (1) exploring the motivation behind practitioners' adoption of CSCRM, (2) analyzing the relationship between Cyber Security and CSCRM, and (3) investigating the connection between CSCRM and Supply Chain 4.0. This is accomplished through the utilization of a conceptual framework to evaluate the importance of linkages between CS, CSCRM, and SC 4.0. The paper presents the study's context, research questions, methods, findings, and conclusions with suggestions for future research.

2. Background of the study

Effective Cyber-risk management in Supply Chains must be integrated from the outset of strategic planning, not addressed at the end (Pandey *et al.*, 2020). Recent studies have mainly explored cyber risks in limited firm samples (Colicchia *et al.*, 2019; Creazza *et al.*, 2022). Industry 4.0 is defined as a cost-effective, data-driven, and adaptable supply network that responds dynamically to fluctuations in both demand and supply (Ivanov *et al.*, 2021).

2.1. Cyber-attacks

Cyber-attacks are breaches of IT systems that can disturb operations or compromise systems over time (Boyson, Corsi and Paraskevas, 2022). They include malicious intentional and unintentional threats such as data leakage, phishing mails or hacking (Kessler *et al.*, 2022). According to the literature, cyber-attacks contain different categories, most of which are cited in the table below:

Table 1. Examples of Cyber-attacks

Reference	Attack Types	Description
(Larriva-Novo *et al.*, 2020)	Fuzzers	Fuzz testing is an automated method that tests software by inputting random or invalid data to detect vulnerabilities.
	Backdoors	In Cyber Security, a backdoor is a way to get beyond an organization's current security measures.

Table 1. Examples of Cyber-attacks (cont.)

Reference	Attack Types	Description
(Larriva-N	DoS	Denial of Service is a cyber-attack that prevents a device or a computer's intended users from using it
	Exploits	A malicious program that exploits holes in hardware or software security
	Shellcode	Attackers use it to target vulnerable processes on local, intranet, or remote systems
(Eggers, 2021)	IP or data theft	Insider disclosure without authorization may lead to further attacks or financial losses
	''Malicious substitution''	Substitution of the entire technological infrastructure, from hardware to firmware
	Tempering, manipulating	It refers to unauthorized changes or commands aimed at manipulating a device's operation or function
(Kern and Szanto, 2022)	Cyber SC attacks	Threats called Cyber SC Attacks aim to compromise the final target particularly through operational endpoint vulnerabilities, by using rusted channels within the supply chain
(Zhang *et al.*, 2019)	Port Scan	Port scanning is a common technique used by hackers to detect unsecured entry points in a network
	DoS Hulk	One of the widely used DoS attacks tools is Hulk. It produces distinct requests in an unstable pattern

2.2. Cyber Security (CS)

Cyber Security (CS) or IT security systems is the shield that protects data and knowledge flows, and prevents data leakage. According to Aamer and *et al.*, 7.35% of the articles cited Cyber Security as a crucial element of preparation for the transition to Supply Chain 4.0 (Aamer, Sahara and Al-Awlaqi, 2023). CS in the supply chain is essential for safeguarding its operations, as it is highly susceptible to threats such as cyberterrorism and malware-malicious software intended to damage computer systems or networks without the user's knowledge (Salem and Al-Saedi, 2023), including data theft. Therefore, Cyber Security measures must be in place to minimize risks, such as buying only from reliable suppliers and disconnecting critical systems from external connectivity (Politeknik Mukah Sarawak, Sarawak, Malaysia *et al.*, 2021).

2.3. Supply Chain 4.0 (SC 4.0)

SC 4.0 denotes the evolution of traditional supply chains through the adoption of I4.0 technologies, such as IoT, CS, AI and Radio Frequency Identification (RFID). This integration facilitates digitalization, interconnectivity, and adaptability, with the

objective of addressing operational challenges, enhancing competitiveness, and optimizing business performance (Sassi *et al.*, 2021). SC 4.0 is a network of coordinated operations that integrates forecasting, production, distribution, and sales to deliver value to customers and suppliers. By aligning these activities, it aims to enhance efficiency, boost innovation, and increase revenue (Martins, Simon and Campos, 2020).

2.4. Cyber Supply Chain Risk Management (CSCRM)

CSCRM is an integrative practice, its goal is to provide strategic oversight across the end-to-end business processes of both the focal organization and its extended enterprise partners, it synthesizes principles of CS, Supply Chain Management, and enterprise risk management (Colicchia, Creazza and Menachof, 2019). According to Muller and al., CSCRM is a process that identifies, evaluates and reduces any risks related to the IT/OT (Operational Technology) goods and services (Muller, 2022). Additionally, it is regarded as a novel and potent idea that integrates business risk management supply chain management, and CS components (Tatt*, Ganesan and Fernando, 2019). Supply chain attacks are cyber threats that can bypass advanced third-party defenses. Their frequency has risen significantly since 2020. Moreover, the integration of I4.0 technologies has further increased system vulnerability (Nygård and Katsikas, 2022). Hence, governments have increasingly focused on cyber supply chain risk management to protect their supply chains at all stages, from procurement and production to distribution, sales, and after-sales, aiming to minimize cybersecurity risks.

2.5. Recommendations from previous studies to strengthen CSCRM

Table 2 presents a synthesis of key recommendations extracted from previous studies aimed at enhancing CSCRM, highlighting technological, organizational, and environmental dimensions.

Table 2. Literature review recommendations for enhancing CSCRM

Recommendations	Authors
Enhance supply chain visibility through a (Technology–Organization–Environment) TOE-based approach to move from fragmented security practices toward a cohesive and productivity-driven CSCRM strategy.	(Gani and Fernando, 2024)
Use agency theory to align roles and responsibilities among supply chain partners, enhancing accountability and reducing cyber risk through governance, risk protocols, and incentives. Support this with capacity building and transparent communication to improve cybersecurity readiness and coordination.	(Firth and Srivastava, 2024)
Embed supplier cybersecurity into logistics and procurement by using a process-driven framework, assigning clear accountability, and treating cyber risk as a core supply chain concern.	(Handfield, Earp and Sadeghi, 2025)

2.6. Overview of the Moroccan automotive and aeronautical industries

2.6.1. Automotive industry

Over the past decade, the Moroccan automotive industry has experienced remarkable and sustained growth, becoming the country's leading export sector. It has created over 147,000 new jobs, attracted more than 250 companies, and established Morocco as the continent's foremost automotive manufacturing hub (Ministry of trade and commerce, 2023).

2.6.2. Aeronautic industry

Morocco is now one of the most competitive and alluring bases on the global map of aviation construction, confirming the Kingdom's great ambition in this high added value industry, after 20 years defined by a singular technological and human journey. Morocco is becoming a prime site and location, with 140 companies (Ministry of trade and commerce, 2023).

3. Research methodology

This research paper explores various research designs, data collection methods, survey instruments, variable measurements, and data analysis techniques. It examines potential relationships, significant or not, among the tested methodologies. As noted by Sassi et al., linear regression models how a dependent variable varies in relation to one or more independent variables, through the least squares method. It includes simple regression (involving a single independent variable) and multiple regression (involving several). A p-value below 0.05 indicates a statistically significant correlation, while regression coefficients reflects the strength of each variable's impact (Sassi et al., 2025).

3.1. Problem description

The reputation of automotive and aeronautic companies for safety is critical to both the companies and their stakeholders. Global supply chain has a very large information flows through all its levers. Its digitalization and transformation will force it to deal with multiple internal and external risks that converge, threatening its stability (Azouzi, Iqqi and Amri, 2023). Therefore, these companies are very proactive about securing their supply chain which leads us to the main problem in this research paper. How to secure the supply chain after the transformation to SC 4.0?

3.2. Questions guiding the study

- Is CSCRM positively associated with SC 4.0?
- Does CS have a very important influence on CSCRM?

3.3. Research objectives

This study aims to examine how directly CSCRM affects Supply Chain 4.0 by explaining variables.

This research aims to achieve the following key objectives:

- To examine the connection between CSCRM and SC 4.0.
- To examine the connection between CS and CSCRM.

3.4. Purpose of the study

SC 4.0 is the fusion of I4.0 technologies with traditional supply chains, resulting in digitalization and increased vulnerability to cyber threats. To manage these risks, CSCRM becomes essential for ensuring the safety and performance of Supply Chains 4.0. This research paper presents a pioneering empirical study on the impact of CSCRM on SC 4.0, offering both theoretical insights and practical contributions for industry and practitioners.

3.4.1. First research construct: CS

The Cyber Security (CS) involves measures such as the trust level of the authenticity and reliability of data, information security and management, and the handling of external data from stakeholders.

3.4.2. Second research construct: CSCRM

CSCRM is measurable by determining at which level the companies respect different governmental guidelines and apply different procedures associated with risk identification, management and evaluation.

3.4.3. Third research construct: SC 4.0

The SC 4.0 could be measured throughout all its levers (PS): ''Purchase and Supply'', (PR): ''Production'', (SD): ''Storage and Distribution'', (SaS): ''Sales and after Sales'' (Sassi *et al.*, 2024).

3.4.4. Presentation of the model research

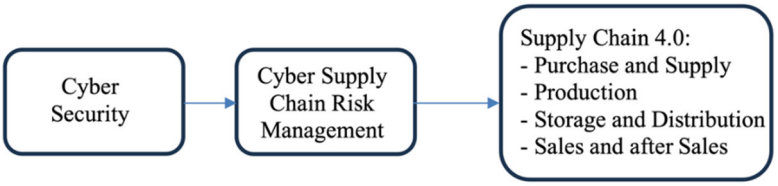

Figure 1. Theoretical Framework

Drawing on the research problem and the literature review, the proposed theoretical framework is organized into six criteria, divided across three categories: one for CS, one for CSCRM, and four for SC 4.0 (as shown in Figures 1 and 2). The model assumes causal relationships among all criteria, with five hypotheses developed to represent the five causal links in the conceptual framework.

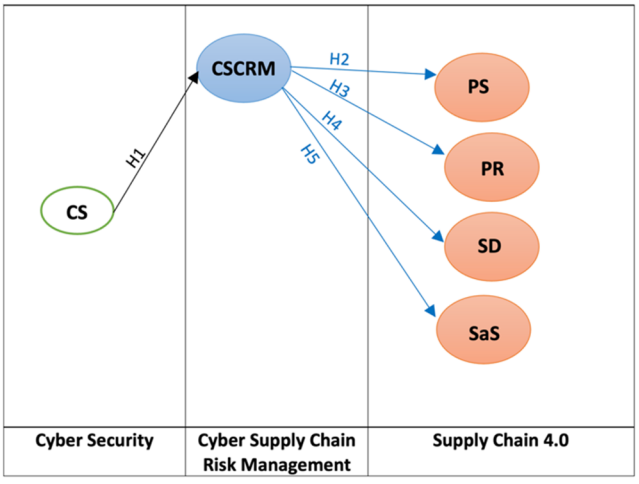

Figure 2. Detailed theoretical model

Table 3. Presenting the codes used within the proposed theoretical model (Sassi *et al.*, 2024)

Proposed Model Constructs	Codes	Titles
Cyber Security	CS	''Cyber Security''
Cyber Supply Chain Risk Management	CSCRM	''Cyber Supply Chain Risk Management''
Supply Chain 4.0	PS	''Purchase and Supply''
	PR	''Production''
	SD	''Storage and Distribution''
	SaS	''Sales and after Sales''

3.4.5. Formulation of the Hypothesis

This research seeks to confirm or refute the 5 hypotheses listed below.

Table 4. Formulated hypotheses

Hypothesis N°	Causal Connection	Formulated Hypothesis
H1	CS ⟶ CSCRM	We hypothesized that CS positively influences CSCRM
H2	CSCRM ⟶ PS	We hypothesized that CSCRM positively influences PS
H3	CSCRM ⟶ PR	We hypothesized that CSCRM positively influences PR
H4	CSCRM ⟶ SD	We hypothesized that CSCRM positively influences SD
H5	CSCRM ⟶ SaS	We hypothesized that CSCRM positively influences SaS

3.5. Preparation of the methodological framework of research:

3.5.1. Formulated hypothesis

For this study, primary data will be gathered using a structured survey, distributed either in paper format or through Google Forms, targeting industrial and supply chain managers as well as general managers in Morocco's automotive and aeronautics industries. The selection of automotive companies is based on information from the directory provided by the Moroccan Automotive Federation "The directory of the automotive sector in Morocco" (La fédération de l'automobile, 2023). For the aeronautic sector, data was sourced from the Group of Moroccan Aeronautical and Space industries. A total of 200 firms were identified across both sectors, including company names and key contact details of high-ranking representatives. The aim was to reach a diverse sample of active firms. The survey was distributed via LinkedIn in August 2023 to 190 companies (140 automotive and 50 aeronautic), and data collection was completed in November 2023. Out of 50 surveys sent to aeronautic companies in the data set, there are 33 responses, which means that the response rate of 66% is deemed representative. Out of 140 surveys distributed to automotive companies, there are 75 responses, which means that the response rate is 53.57% which is considered moderated. Notice that the total number of interviewers who answered the survey is 94 but since some companies are considered active in both sectors, the total number of responses amounts to 108, thus, the total rate of response is 56.84%. According to the classification criteria by the Moroccan High Commission for planning, the firms selected for this study represent various organizational strata (1.1% small firms, 22.5 % medium-sized firms, and the remaining 76.4% big firms). The demographics of the companies featured in the study are summarized in Table 5 (Moroccan High Commission for Planning, 2019).

Table 5. Demographics of participating firms

Characteristics	Categories	Frequency	Percentage
Year of foundations	<11 y	38	40.4%
	11 – 20 y	24	25.8%
	>20 y	32	33.7%
Region	Casablanca and region	32	33.7%
	Tanger and region	38	40.4%
	Marrakech and region	01	0.01%
	Kenitra and region	23	25.8%
Company's work force	<10	01	1.1%
	<200	21	22.5%
	>200	72	76.4%
Capital	<50 000 dhs	01	1.3%
	50 001 – 400 000 dhs	02	2.6%
	400 001 – 600 000 dhs	04	5.1%
	600 001 – 1 000 000 dhs	03	3.8%
	>1 000 000 dhs	83	87.2%

3.6. Questionnaire design

3.6.1. Questionnaire steps

This study's questionnaire consists of two primary sections, with the first focusing on respondent and company information, and the second on evaluating the study's conceptual model constructs, CS, CSCRM and SC 4.0, using 44 items (7 for CS, 8 for CSCRM, and 30 for SC 4.0). A six-point Likert scale, from 0 = Abs/No to 6 = Very High, was used for assessment (Likert, 1932). The measurement instrument was developed following Churchill's (1979) paradigm (Churchill, 1979), which involves defining the construct through an extensive literature review, generating diverse measurement items, collecting data, and refining the instrument using statistical tools like coefficient alpha and factor analysis. Reliability and validity were further tested using methods such as split-half reliability and the multirait-multimethod matrix. Norms were then established to enable meaningful interpretation and comparisons.

3.6.2. Reliability test

The reliability of the questionnaire instrument was assessed using Cronbach's Alpha. As shown in Table 6, all constructs and measurement items used in the study demonstrate acceptable reliability, with values exceeding the threshold of 0.7 (Fornell and Bookstein, 1982; Kline, 1999).

Table 6. Evaluation summary of the three constructs' reliability and validity

Constructs	Variables	Code	Number of items	Cronbach's Alpha α
Cyber Security	Cyber Security	CS	7	0.949
Cyber Supply Chain Risk Management	Cyber Supply Chain Risk Management	CSCRM	8	0.968
Supply Chain 4.0	Purchase and Supply	PS	7	0.957
	Production	PR	8	0.985
	Storage and Distribution	SD	7	0.965
	Sales and after Sales	SaS	8	0.976

4. Results of the study

In this section, we detail the findings obtained from the empirical analysis.

4.1. Examining the connection between CS and CSCRM

4.1.1. Global Model: Cyber Security (CS) – Cyber Supply Chain Management (CSCRM)

Overall, Table 7 indicates a strong correlation between CS and CSCRM (R = 0.616). The dependent variable, CSCRM, accounts for 38.0 % of the variance in the predicted variable CS, as shown by the R-squared value of 0.380. Additionally, the model is statistically significant, with a p-value of 0.00 (Table 9), which is well below the 5% threshold, confirming the model's overall validity.

Table 7. Summary of model characteristics (CS-CSCRM)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
CS-CSCRM	0.616	0.380	0.373	0.79193260

a. Predicted (Independent) value: CS.
b. Dependent variable: CSCRM.
c. Linear regression at the origin.

Table 8. Variance Analysis (CS-CSCRM)

Model	Sum of squares	ddl	Average Square	D	Sig
Regression	35.302	1	35.302	56.288	0.000
Residual	57.698	92	0.627		
Total	93.000	93			

a. Dependent variable: CSCRM.
b. Linear regression at the origin.
c. Predicted value: CS.

4.1.2. Linear model equation: Cyber Security (CS) – Cyber Supply Chain Risk Management (CSCRM)

Based on Table 9, the linear regression equation is expressed as follows:

$$CSCRM = 0.616\ CS \quad (1)$$

The causal relationship between CS and CSCRM is statistically significant (P-value = 0 < 5%), indicating that CS has a strong impact on CSCRM. Additionally the t-student exceeds |2.775| (|1.960|), confirming that the parameter estimates are significant at the 1% (or 5%) significance level (Oullada *et al.*, 2023).

Table 9. Criteria coefficients (CS-CSCRM)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
CS-CSCRM	0.616	0.082	0.616	7.503	0.000

4.2. Examining the connection between CSCRM and PS

4.2.1. Overall model: Cyber Supply Chain Risk Management (CSCRM) – Purchase and Supply (PS)

According to Table 10, we notice that the causal relationship between CSCRM and PS is positively significant (R = 0.723). The explanatory variable, PS, accounts for 52.3% of the variance in the dependent variable CSCRM, as indicated by an R-squared value of 0.523. Furthermore, the model is globally valid since its significance value is lower than 5%.

Table 10. The overall model (CSCRM-PS)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
CSCRM-PS	0.723	0.523	0.518	0.69408436

a. Predicted value: CSCRM.
b. Dependent variable: PS.
c. Linear regression at the origin.

Table 11. Variance analysis (CSCRM-PS)

Model	Sum of squares	ddl	Average Square	D	Sig
Regression	48.679	1	48.679	101.045	0.000
Residual	44.321	92	0.482		
Total	93.000	93			

a. Dependent variable: PS.
b. Linear regression at the origin.
c. Predicted value: CSCRM.

4.2.2. Linear model equation: Cyber Supply Chain Risk Management (CSCRM) – Purchase and Supply (PS)

According to Table 12, the linear regression equation is formulated as follows:

$$PS = 0.723\ CSCRM \quad (2)$$

Note that the causal relationship between CSCRM and PS is highly significant (p-value = 0 < 5%), which means that CSCRM has a strong impact on PS.

Table 12. Criteria coefficients (CSCRM-PS)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
CSCRM-PS	0.723	0.072	0.723	10.052	0.000

4.3. Examining the connection between CSCRM and PR

4.3.1. Overall model: Cyber Supply Chain Risk Management (CSCRM) – Production (PR)

There is a moderately strong connection between CSCRM and PR (R = 0.504). The dependent variable PR, accounts for 25.40 % of the variance in the predicted variable CSCRM (R-squared = 0.254). The model is considered globally reliable since the significance value (p-value) is below 5% (Table 15).

Table 13. The overall model (CSCRM-PR)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
CSCRM-PR	0.504	0.254	0.246	0.86812659

a. Predicted value: CSCRM
b. Dependent variable: PR
c. Linear regression at the origin

Table 14. Variance analysis (CSCRM-PR)

Model	Sum of squares	ddl	Average Square	D	Sig
Regression	23.665	1	23.665	31.400	0.000
Residual	69.335	92	0.754		
Total	93.000	93			

a. Dependent variable: PR.
b. Linear regression at the origin.
c. Predicted value: CSCRM.

4.3.2. Linear model equation: Cyber Supply Chain Risk Management (CSCRM) – Production (PR)

The linear regression equation is written as follows on the basis of Table 15:

$$PR = 0.504\ CSCRM \tag{3}$$

We notice that (p-value = 0 < 5%), thus, the causal relationship CSCRM-PR is considered significant, and we conclude that the CSCRM strongly impacts PR.

Table 15. Criteria Coefficients (CSCRM-PR)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
CSCRM-PR	0.504	0.090	0.504	5.604	0.000

4.4. Examining the connection between CSCRM and SD

4.4.1. Overall model: Cyber Supply Chain Risk Management (CSCRM) – Storage and Distribution (SD)

There is a strong connection between CSCRM and SD, with a correlation of R=0.454. The dependent variable SD, explains 20.6% of the variance in the independent variable CSCRM, as indicated by an R-squared value of 0.206. Additionally, the model's significance value is below the 5% threshold, confirming its overall statistical validity.

Table 16. The overall model (CSCRM-SD)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
CSCRM-SD	0.454	0.206	0.197	0.89595774

a. Predicted value: CSCRM.
b. Dependent variable: SD.
c. Linear regression at the origin.

Table 17. Variance Analysis (CSCRM-SD)

Model	Sum of squares	ddl	Average Square	D	Sig
Regression	19.148	1	19.148	23.853	0.000
Residual	73.852	92	0.803		
Total	93.000	93			

a. Dependent variable: SD.
b. Linear regression at the origin.
c. Predicted value: CSCRM.

4.4.2. Linear model equation: Cyber Supply Chain Risk Management (CSCRM) – Storage and Distribution (SD)

Based on the results presented in Table 18, the linear regression equation is expressed as follows:

$$SD = 0.454\ CSCRM \quad (4)$$

The causal relationship CSCRM-SD is considered significant since the (sig. =0<5%), thus, the CSCRM strongly impacts SD.

Table 18. Criteria coefficients (CSCRM-SD)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
CSCRM-SD	0.454	0.93	0.454	4.884	0.000

4.5. Examining the connection between CSCRM and SaS

4.5.1. Overall model: Cyber Supply Chain Risk Management (CSCRM) – Sales and after Sales (SaS)

In general, a moderate strong positive correlation exists between CSCRM and SaS criterion (R = 0.514). The model's significance level (p-value = 0.00, Table 20) is below the 5% threshold, confirming the model's reliability.

Table 19. The overall model (CSCRM-SaS)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
CSCRM-SaS	0.514	0.264	0.256	0.86246393

a. Predicted value: CSCRM.
b. Dependent variable: SaS.
c. Linear regression at the origin.

Table 20. Variance analysis (CSCRM-SaS)

Model	Sum of squares	ddl	Average Square	D	Sig
Regression	24.566	1	24.566	33.026	0.000
Residual	68.434	92	0.744		
Total	93.000	93			

a. Dependent variable: SaS.
b. Linear regression at the origin.
c. Predicted value: CSCRM.

4.5.2. Linear model equation: Cyber Supply Chain Risk Management (CSCRM) – Sales and after Sales (SaS)

The linear regression model derived from Table 21 is formulated as follows:

$$SaS = 0.514\ CSCRM \tag{5}$$

The causal relationship between CSCRM and SaS is statistically significant (p-value = 0 < 0.05), suggesting that CSCRM exerts a notable influence on SaS.

Table 21. Criteria coefficients (CSCRM-SaS)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
CSCRM-SaS	0.514	0.089	0.514	5.747	0.000

4.6. Overall model summarizing hypothesis testing outcomes

The outcomes of the hypothesis test (Table 22) and the global model (Figure 3) are detailed below:

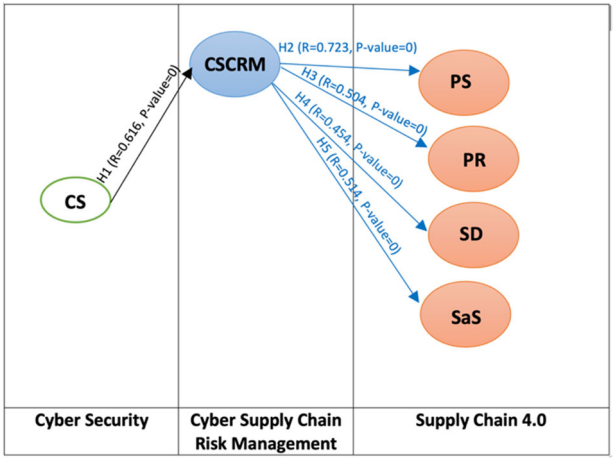

Figure 3. Overall model summarizing hypothesis test results.

Table 22. Hypothesis testing outcomes

Hypotheses	Results
H1: CS ⟶ CSCRM	Valid
H2: CSCRM ⟶ PS	Valid
H3: CSCRM ⟶ PR	Valid
H4: CSCRM ⟶ SD	Valid
H5: CSCRM ⟶ SaS	Valid

5. Discussion and conclusion

This section presents the findings of the empirical investigation, revealing growing public concern over CSCRM due to rising cyberattacks and scams. In Morocco, this concern is increasingly evident, especially in the key automotive and aeronautic industries. Serhane et al. (2023) emphasized that greater connectivity heightens industrial systems' vulnerability to cyber threats, including attacks and unauthorized access (Serhane, Hamzaoui and Ibrahimi, 2023). Industrial systems were long vulnerable to cyberattacks due to limited security in their design and their isolation from the internet. With Industry 4.0, this isolation has diminished, exposing them to a wider spectrum of cyber threats (Tamy *et al.*, 2020). Digitalizing supply chains is challenging as it demands specialized digital skills, unique organizational competencies, and a clear digital strategy (Kohnke, 2017). Hence, companies operating within Industry 4.0 need to cultivate dynamic capabilities to effectively navigate Supply Chain risks. Using the DEMATEL approach, Pandey et al. identified behavioral risk as the most critical among the various risk categories. Future research should focus on risks linked to specific I4.0 technologies within Supply Chains (Pandey, Singh and Gunasekaran, 2021). This study clarifies how CSCRM directly affects Supply Chain 4.0, linking key variables. CSCRM, supported by Cyber Security, addresses cyberattack emerging from the shift to digital supply chains.

Based on the applied research methodology, this study also finds that, Cyber Security can also affect positively CSCRM to help it influence positively the SC 4.0.

Our exploratory study's findings indicate that:

- The relationship between CS and CSCRM is both significant and positively correlated.
- The relationship between the CSCRM and Purchase and Supply (PS) is significant and positive.
- The relationship between the criterion CSCRM and Production (PR) criterion is fairly high.
- CSCRM has a positive and quite high influence on Storage and Distribution criterion.
- The relationship between CSCRM and Sales and after Sales (SaS) is fairly strong and positive.

The study shows Cyber Security positively influences CSCRM, whose success depends on strong CS. Implementing CSCRM enhances Supply Chain 4.0 by positively impacting its key levers. Therefore, supply chain managers must understand and adopt CSCRM, aligning with Tamy and al. (2020)'s strategic approach to safeguarding Industry 4.0 networks (Tamy *et al.*, 2020). These results are also in line with the outcomes of Aarland et al. (2025), who emphasized the prominence of establishing clear cybersecurity requirements and ensuring proper management within digital supply chains (Aarland, 2025). This study shows that CSCRM enhances SC 4.0 performance in Morocco's automotive and aeronautical sectors. The integration of CS, CSCRM, and SC 4.0 into one framework highlights the essential role of CS in effective CSCRM and offers practical strategies for managing digital risks in emerging industries.

6. Limitations

Although the findings are promising, the study has limitations, notably an average response rate. Larger samples should be used in future research to ensure more representative and reliable results. Moreover, the research was limited in scope to samples from only two sectors: the automotive and aeronautic industries. It is recommended that this model be applied to other fields, such as textiles, agriculture, and others. Furthermore, given that this study concentrates on one specific enabling technology, subsequent research could examine a broader spectrum of technologies to validate and refine the proposed framework.

References

Aamer, A., Sahara, C.R. and Al-Awlaqi, M. A., (2023) Digitalization of the supply chain: transformation factors. *Journal of Science and Technology Policy Management*, 14(4), pp. 713–733. Available at: https://doi.org/10.1108/JSTPM-01-2021-0001.

Aarland, M., (2025) Cybersecurity in digital supply chains in the procurement process: introducing the digital supply chain management framework, *Information & Computer Security*, 33(1), pp. 5–24. Available at: https://doi.org/10.1108/ICS-10-2023-0198.

Azouzi, M., Iqqi, I. and Amri, M., (2023) Safety and security as risk management factors in supply chains. *Journal of Operations Management*, 3(1), pp. 1-10.

Boyson, S., Corsi, T. M. and Paraskevas, J.-P., (2022) Defending digital supply chains: Evidence from a decade-long research program. *Technovation*, 118, p. 102380. Available at: https://doi.org/10.1016/j.technovation.2021.102380.

Churchill, G. A., (1979) A Paradigm for Developing Better Measures of Marketing Constructs. *Journal of Marketing Research*, 16(1), pp. 64–73. Available at: https://doi.org/10.1177/002224377901600110.

Colicchia, C., Creazza, A. and Menachof, D. A., (2019) Managing cyber and information risks in supply chains: insights from an exploratory analysis. *Supply Chain Management: An International Journal*, 24(2), pp. 215–240. Available at: https://doi.org/10.1108/SCM-09-2017-0289.

Creazza, A. *et al.*, (2022) Who cares? Supply chain managers' perceptions regarding cyber supply chain risk management in the digital transformation era. *Supply Chain Management: An International Journal*, 27(1), pp. 30–53. Available at: https://doi.org/10.1108/SCM-02-2020-0073.

Eggers, S., (2021) A novel approach for analyzing the nuclear supply chain cyber-attack surface. *Nuclear Engineering and Technology*, 53(3), pp. 879–887. Available at: https://doi.org/10.1016/j.net.2020.08.021.

Firth, R., Srivastava, M., (2024) Identifying Critical Success Factors (CSF) for Cyber Supply Chain Risk Management (CSCRM): A Qualitative Study Using Agency Theory, in K.S. Soliman (ed.) *Artificial intelligence and Machine Learning*. Cham: *Springer Nature Switzerland* (Communications in Computer and Information Science), pp. 173–186. Available at: https://doi.org/10.1007/978-3-031-62843-6_19.

Foli, S. *et al.*, (2022) Supply Chain Risk Management in Young and Mature SMEs. *Journal of Risk and Financial Management*, 15(8), p. 328. Available at: https://doi.org/10.3390/jrfm15080328.

Fornell, C., Bookstein, F. L., (1982) Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research*, 19(4), p. 440. Available at: https://doi.org/10.2307/3151718.

Gani, A. B. D., Fernando, Y., (2024) Ten-year review of cyber supply chain security: driving productivity with visibility. *International Journal of Productivity and Quality Management*, 42(2), pp. 153–169. Available at: https://doi.org/10.1504/IJPQM.2024.139156.

Handfield, R., Earp, J. and Sadeghi, A. H., (2025) Reducing cybersecurity vulnerabilities in the supply base: Insights from cyber experts, *Technology in Society*, 82, p. 102947. Available at: https://doi.org/10.1016/j.techsoc.2025.102947.

Ivanov, D. *et al.*, (2021) Researchers' perspectives on Industry 4.0: multi-disciplinary analysis and opportunities for operations management. *International Journal of Production Research*, 59(7), pp. 2055–2078. Available at: https://doi.org/10.1080/00207543.2020.1798035.

Kern, E. Szanto, A., (2022) Cyber Supply Chain Attacks, in Tim H. Stuchtey (eds). *BIGS Policy Paper Brandenburg Institute for Society and Security*, p. 10.

Kessler, M. *et al.*, (2022) Curse or Blessing? Exploring risk factors of digital technologies in industrial operations. *International Journal of Production Economics*, 243, p. 108323. Available at: https://doi.org/10.1016/j.ijpe.2021.108323.

Kline, R. B., (1999) Book Review: Psychometric theory (3rd ed.). *Journal of Psychoeducational Assessment*, 17(3), pp. 275–280. Available at: https://doi.org/10.1177/073428299901700307.

Kohnke, O., (2017) It's Not Just About Technology: The People Side of Digitization, in G. Oswald and M. Kleinemeier (eds). *Shaping the Digital Enterprise*. Cham: Springer International Publishing, pp. 69–91. Available at: https://doi.org/10.1007/978-3-319-40967-2_3.

La fédération de l'automobile, (2023) Annuaire du secteur automobile au Maroc. Available at: https://cgem.ma/structures/federations-statutaires/federation-de-lautomobile-fa/ (Accessed: 15 August 2023).

Larriva-Novo, X. A. *et al.*, (2020) Evaluation of Cybersecurity Data Set Characteristics for Their Applicability to Neural Networks Algorithms Detecting Cybersecurity Anomalies. *IEEE Access*, 8, pp. 9005–9014. Available at: https://doi.org/10.1109/ACCESS.2019.2963407.

Likert, R., (1932) *A technique for the measurement of attitudes*. Academic Dissertation. Columbia University.

Martins, F. D. C., Simon, A. T. and Campos, R. S. D., (2020) Supply Chain 4.0 challenges. *Gestão & Produção*, 27(3), p. e5427. Available at: https://doi.org/10.1590/0104-530x5427-20.

Ministry of Industry and Trade, (2023) Aeronautic. [Online] Available at: https://www.mcinet.gov.ma/fr/content/aeronautique. Consulted: May 2023.

Ministry of Industry and Trade, (2023) Automobile. [Online] Available at: https://www.mcinet.gov.ma/fr/content/aeronautique. Consulted: May 2023.

Moroccan High Commission for Planning, (2019) Classification criteria of Moroccan companies. *19 Novembre 2019*, 19 November, p. 28.

Muller, S. R., (2022) Analyzing Deficits in Awareness Among Chief Supply Chain Officers Who Have Not Adopted Cybersecurity as a Threat to Supply Chains, in *2022 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT). 2022 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, Soyapango. El Salvador: *IEEE*, pp. 1–6. Available at: https://doi.org/10.1109/ICMLANT56191.2022.9996456.

Nygård, A. R., Katsikas, S., (2022) SoK: Combating threats in the digital supply chain, in *Proceedings of the 17th International Conference on Availability, Reliability and Security. ARES 2022: The 17th International Conference on Availability, Reliability and Security*, Vienna Austria: *ACM*, pp. 1–8. Available at: https://doi.org/10.1145/3538969.3544421.

Oullada, O. *et al.* ,(2023) Model for measuring the impact of good pharmacovigilance practices of COVID-19 patients on hcp reactivity: Morocco case study. *Statistics in Transition new series*, 24(5), pp. 63–88. Available at: https://doi.org/10.59170/stattrans-2023-064.

Pandey, S. *et al.*, (2020) Cyber security risks in globalized supply chains: conceptual framework. *Journal of Global Operations and Strategic Sourcing*, 13(1), pp. 103–128. Available at: https://doi.org/10.1108/JGOSS-05-2019-0042.

Pandey, S., Singh, R. K. and Gunasekaran, A., (2021) Supply chain risks in Industry 4.0 environment: review and analysis framework. *Production Planning & Control*, pp. 1–28. Available at: https://doi.org/10.1080/09537287.2021.2005173.

Politeknik Mukah Sarawak, Sarawak, Malaysia *et al.*, (2021) Cyber security in supply chain management: A systematic review. *Logforum*, 17(1), pp. 49–57. Available at: https://doi.org/10.17270/J.LOG.2021555.

Salem, I. E., Al-Saedi, K. H., (2023) Intensive Malware Detection Approach based on Data Mining. *Journal of Applied Engineering and Technological Science (JAETS)*, 5(1), pp. 414–424. Available at: https://doi.org/10.37385/jaets.v5i1.2865.

Sassi, A. *et al.*, (2021) The relation between Industry 4.0 and Supply Chain 4.0 and the impact of their implementation on companies. *International Journal of Innovation and Applied Studies*. performance: State of the Art', 31(4), p. 820-828.

Sassi, A. *et al.*, (2024) Model for Assessing the Impact of Internet of Things on Supply Chain 4.0: Moroccan Case, in Y. Mejdoub and A. Elamri (eds) *Proceeding of the International Conference on Connected Objects and Artificial Intelligence (COCIA2024)*. Cham: Springer Nature Switzerland (Lecture Notes in Networks and Systems), pp. 252–258. Available at: https://doi.org/10.1007/978-3-031-70411-6_39.

Sassi, A. *et al.*, (2025) A causal model to assess the influence of supply chain 4.0 on Moroccan companies' performance. *International Journal of Advances in Applied Sciences*, 14(1), p. 111. Available at: https://doi.org/10.11591/ijaas.v14.i1.pp111-122.

Serhane, A., Hamzaoui, E.-M. and Ibrahimi, K., (2023) IA Applied to IIoT Intrusion Detection: An Overview, in *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM). 2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Istanbul. Turkiye: *IEEE*, pp. 1–6. Available at: https://doi.org/10.1109/WINCOM59760.2023.10323032.

Tamy, S. *et al.*, (2020) Cyber security based machine learning algorithms applied to industry 4.0 application case: Development of network intrusion detection system using hybrid method. *Journal of Theoretical and Applied Information Technology*, 98(12), pp. 2078–2091.

Tatt*, D. Y. B., Ganesan, Y. and Fernando, Y., (2019) 'The Effects of Cyber Supply Chain Risk Management in Financial Industry'. *ICBSI 2018 - International Conference on Business Sustainability and Innovation*, pp. 512–521. Available at: https://doi.org/10.15405/epsbs.2019.08.51.

Zhang, Y. *et al.*, (2019) 'PCCN: Parallel Cross Convolutional Neural Network for Abnormal Network Traffic Flows Detection in Multi-Class Imbalanced Network Traffic Flows'. *IEEE Access*, 7, pp. 119904–119916. Available at: https://doi.org/10.1109/ACCESS.2019.2933165.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 85–103, https://doi.org/10.59139/stattrans-2026-005
Received – 06.03.2024; accepted – 24.09.2025

An extended odd log-logistic-Lindley distribution with properties, applications and Bayesian estimation

Abbas Eftekharian[1], **Morad Alizadeh**[2], **Vahid Ranjbar**[3], **Omid Kharazmi**[4], **Gholamhossein Hamedani**[5]

Abstract

This paper introduces a four-parameter extended odd log-logistic-Lindley distribution from which moments, hazard, and quantile functions are then obtained. The statistical properties of this distribution show the high flexibility of the proposed distribution. The maximum likelihood and least-squares estimators of the extended odd log-logistic-Lindley parameters are studied. Moreover, a simulation study is carried out for evaluating the performance of the estimation methods, and the usefulness of the new distribution is illustrated using two real data sets. Finally, Bayesian analysis and efficiency of Gibbs sampling are provided on the basis of two real data sets.

Key words: Bayesian estimation, Gibbs sampling, Lindley distribution, moment, odd log-logistic, simulation.

1. Introduction

Modelling and analysing real lifetime data are widely used in many applied fields such as finance, reliability, engineering, medicine. In practice, researchers dealt with different types of survival data and they proposed various lifetime models for modelling such data. The statistical analysis depends on the procedure used by the researcher and the generated family of distributions. Recently, new families of distributions have been introduced in the literature that could considerably help to analyse complex real data. However, it is necessary to find more efficient statistical models; since there are many real data sets in practice that need to be investigated with statistical models that are more flexible. Therefore, the researchers have had many attempts to extend distributions theory by adding new shape parameters to different families of distribution to introduce new families. In particular, some extended distributions demonstrate high flexibility in hazard rate function (hrf) such as increasing, decreasing and bathtub shapes even though the baseline hazard rate function may not have these shapes.

[1]University of Hormozgan, Iran. E-mail: ab2005eftekharian@gmail.com.
ORCID: https://orcid.org/0000-0002-5343-8597.
[2]Persian Gulf University, Iran. E-mail: m.alizadeh@pgu.ac.ir.
ORCID: https://orcid.org/0000-0001-6638-2185.
[3]Corresponding author. Golestan University, Iran. E-mail: vahidranjbar@gmail.com.
ORCID: https://orcid.org/0000-0003-3743-0330.
[4]Vali-e-Asr University of Rafsanjan, Iran. E-mail: omidkharazmi14@gmail.com.
ORCID: https://orcid.org/0000-0003-4176-9708.
[5]Marquette University, United States, E-mail: gholamhoss.hamedani@marquette.edu.
ORCID: https://orcid.org/0000-0003-3216-0511.

Most of the new generators of G family can be obtained using T-X class, which is proposed by Alzaatreh et al. (2013). For example, Kumaraswamy generated, odd log-logistic-G, Exponentiated-G (Exp-G), gamma generated, proportional odds and generalized beta generated. Recently, the extended exponentiated-G (EE-G) family was defined by Alizadeh et al. (2018a).

In this paper, we introduce a new generator of G family using T-X class, which is called the extended odd log-logistic-G (EOLL-G) family, and study some of its mathematical properties. The main idea of the EOLL-G family is based on a contribution presented by Gleaton and Lynch (2010). They introduced an extended generalized log-logistic family for lifetime distribution. Following their idea and using T-X class the cumulative distribution function (cdf) of the EOLL-G family with parameters $\alpha > 0,\ \beta > 0$ and $\gamma > 0$ as three additional shape parameters is defined by

$$F(x;\xi) = \int_0^{\frac{G(x;\theta)^\alpha}{[1-G(x;\theta)]^\beta}} \frac{\gamma}{(1+\gamma t)^2} dt = \frac{G(x;\theta)^\alpha}{G(x;\theta)^\alpha + \gamma[1-G(x;\theta)]^\beta}, \tag{1}$$

where $G(x;\theta)$ is the baseline cdf with the parameter vector θ and $\xi = (\alpha, \beta, \gamma, \theta)$.

It is clear that in the special case, the EOLL-G family reduces to EE-G family when $\beta = 1$. For $\alpha = \gamma = 1$, it transforms into Marshal-Olkin family. If $\beta = 1$ and $\alpha = \gamma$, then it reduces to Exp-G family. By considering $\alpha = \beta = \gamma = 1$, we obtain the baseline distribution G.

Gleaton and Lynch (2010) showed that the extended generalized log-logistic family has appropriate performance for lifetime data. Therefore, we can use the EOLL-G family for lifetime data by choosing a lifetime distribution as $G(\cdot)$ in (1). Although, there are several lifetime distributions that we can use, which is due to the fact that the proposed family has three parameters, it is better to select a lifetime distribution with only one parameter, for example, exponential or Lindley. It should be noted that hrf of the exponential is constant while the hrf of the Lindley distribution has different shapes as increasing, decreasing, unimodal and bathtub. Moreover, the Lindley distribution is a well-known distribution that is employed widely in different fields such as lifetime and reliability, medical, finance, engineering and insurance. These reasons motivate the use of this distribution for modeling real lifetime data. Therefore, we consider the Lindley distribution as the baseline distribution in this paper.

The Lindley distribution was originally proposed by Lindley (1958) in the Bayesian statistical context. Some properties of this distribution such as moments, failure rate function, characteristic function, mean residual life function, mean deviations, Lorenz curve, stochastic ordering, entropies, asymptotic distribution of the extreme order statistics have been studied by Ghitany et al. (2008). The cdf of the Lindley distribution with scale parameter $\lambda > 0$ is

$$G(x;\lambda) = 1 - \left(1 + \frac{\lambda x}{1+\lambda}\right) \mathrm{e}^{-\lambda x},\ x > 0, \tag{2}$$

and its corresponding probability density function (pdf) is given by

$$g(x;\lambda)=\frac{\lambda^2}{1+\lambda}(1+x)\mathrm{e}^{-\lambda x}. \tag{3}$$

Many authors have published various extensions of the Lindley distribution recently. For example, a three-parameter generalization of the Lindley distribution proposed by Zakerzadeh and Dolati (2009), Nadarajah et al. (2011) defined a generalized Lindley distribution, a new generalized Lindley distribution based on the weighted mixture of two gamma distributions was studied by Abouammoh et al. (2015), Asgharzadeh et al. (2016, 2018) introduced a weighted Lindley distribution and Weibull Lindley distribution, respectively, and Alizadeh et al. (2017a,b,2018b) proposed several generalizations of the Lindley distribution based on the odd log-logistic model. Given the vast amount of papers published recently, we can only mention a few of the most recent contributions: Gomes-Silva et al. (2017), Afify et al. (2019), Alizadeh et al. (2019) and Alizadeh et al. (2025).

In the present paper, we introduce a new generalization of the Lindley distribution using the EOLL-G family. To this end, it is enough to choose the Lindley distribution as the baseline $G(x;\theta)$ in (1). By substituting (2) in (1), we get

$$F(x;\alpha,\beta,\gamma,\lambda)=\frac{\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\alpha}}{\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\alpha}+\gamma\left[\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\beta}},\quad x\geq 0, \tag{4}$$

and its corresponding pdf is given by

$$\begin{aligned}f(x;\alpha,\beta,\gamma,\lambda)=&\frac{\gamma\lambda^2(1+x)\left(1+\frac{\lambda}{1+\lambda}x\right)^{\beta-1}\mathrm{e}^{-\beta\lambda x}\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\alpha-1}}{(1+\lambda)\left\{\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\alpha}+\gamma\left[\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\beta}\right\}^2}\\&\times\left\{\alpha+(\beta-\alpha)\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]\right\}.\end{aligned} \tag{5}$$

A random variable X with pdf (5) has extended odd log-logistic-Lindley (EOLL-L) distribution and is denoted by $X\sim$ EOLL-L$(\alpha,\beta,\gamma,\lambda)$. The EOLL-L distribution is more flexible than the Lindley distribution and allows for greater flexibility of the tails.

Special cases: Let $X\sim EOLL-L(\alpha,\beta,\gamma,\lambda)$.

- If $\alpha=\beta,\gamma=1$, then EOLL-L reduces to the Odd Log-Logistic Lindley (OLL-L) Ozel et al. (2017).
- For $\alpha=\beta$, EOLL-L coincides with OLL-Marshall- Olkin Lindley (OLL-MOL) Alizadeh et al. (2017b).
- If $\alpha=\beta=1$, then XEOLL-L reduces to Marshall- Olkin Lindley (MOL).

- By taking $\gamma = 1$, EOLL-L coincides with the new OLL-Lindley (NOLLL) Alizadeh et al. (2018b).

- For $\alpha = \beta = \gamma = 1$, EOLL-L is ordinary Lindley.

The different shapes of the pdf such as unimodal, symmetric, skewed, and monotonically decreasing are shown in Figure 1 (left plot). As seen, the density of the EOLL-L model can be right-skewed density with one peak and heavy tail to the right, right-skewed density without a peak and with heavy tail to the right, bimodal and unimodal density with different shapes.

The point that catches our attention in this graph is that, for gamma values greater than 1, the density curve is symmetrical, and for less than 1, it is skewed to the right, and also for larger beta values, the tails of the distribution will become heavier.

The rest of the paper is organized as follows. In Section 2, some mathematical properties of the EOLL-L distribution are obtained. Certain characterizations are presented in Section 3. The estimations of the unknown parameters based on different methods are investigated in Section 4. A simulation study is reported in Section 5. In Section 6, the performance and application of the EOLL-L distribution are evaluated using two real data sets. Bayesian inference and Gibbs sampling procedure for the considered data sets are investigated in Section 7. Finally, some conclusions are stated in Section 8.

2. Main properties

2.1. Hazard rate function

In reliability studies, the hrf is an important characteristic and fundamental to the design of safe systems in a wide variety of applications. Using equations (4) and (5) the hrf of the EOLL-L distribution takes the form

$$\begin{aligned} h(x;\alpha,\beta,\lambda) &= \left(\frac{\lambda^2(1+x)\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\alpha-1}}{(1+\lambda+\lambda x)\left\{\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\alpha}+\gamma\left[\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\beta}\right\}} \right) \\ &\times \left\{\alpha+(\beta-\alpha)\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]\right\}. \end{aligned} \tag{6}$$

Plots for the hrfs for selected parameter values are displayed in Figure 1(right plot). As seen in Figure 1, the hrf of the EOLL-L distribution has very flexible shapes such as increasing, decreasing, upside-down, bathtub and upside-down-bathtub. It is evident that the EOLL-L distribution is more flexible than the Lindley distribution, in other words, the additional parameters $\alpha > 0, \beta > 0$ allow for a high degree of flexibility of the EOLL-L distribution. This attractive flexibility implies that the hrf of the EOLL-L is useful for non-monotone empirical hazard behaviour, which is more likely observed in real life situations.

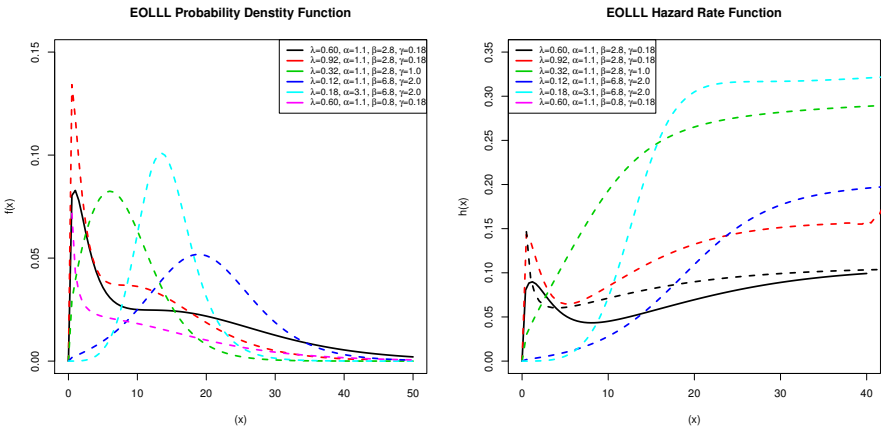

Figure 1. Plots of the density and hazard function for the EOLL-L distribution for selected parameter values

2.2. Quantile function

Quantile function is generally used to find representations in terms of lookup tables for key percentiles. Let X be an EOLL-L distributed random variable with parameters $\alpha, \beta, \lambda, \gamma$. The quantile function, $Q(p)$, defined by $F[Q(p)] = p$ is the root of the equation as

$$p = \frac{\left[1 - \left(1 + \frac{\lambda}{1+\lambda} Q(p)\right) \mathrm{e}^{-\lambda Q(p)}\right]^{\alpha}}{\left[1 - \left(1 + \frac{\lambda}{1+\lambda} Q(p)\right) \mathrm{e}^{-\lambda Q(p)}\right]^{\alpha} + \gamma \left[\left(1 + \frac{\lambda}{1+\lambda} Q(p)\right) \mathrm{e}^{-\lambda Q(p)}\right]^{\beta}}. \tag{7}$$

A closed form of quantile function is available when $\alpha = \beta$. For this purpose, we define

$$[1 + \lambda + \lambda Q(p)] \mathrm{e}^{-\lambda Q(p)} = \frac{(1+\lambda)(1-p)^{\frac{1}{\alpha}}}{(\gamma p)^{\frac{1}{\alpha}} + (1-p)^{\frac{1}{\alpha}}}, \tag{8}$$

for $0 < p < 1$. After some simple algebraic manipulation one can obtain

$$Q(p) = -1 - \frac{1}{\lambda} - \frac{1}{\lambda} W_{-1}\left[\frac{-(1+\lambda)(1-p)^{\frac{1}{\alpha}} \mathrm{e}^{-1-\lambda}}{(\gamma p)^{\frac{1}{\alpha}} + (1-p)^{\frac{1}{\alpha}}}\right]. \tag{9}$$

where $W_{-1}[.]$ is the negative branch of the Lambert function (Corless et al. 1996). Note that the particular case of (9) for $\alpha = \beta = \gamma = 1$ is derived by Jodr (2010).

Now, we propose the following algorithm for generating random data from the EOLL-L distribution for the case $\alpha = \beta$.

Algorithm 1 (Inverse cdf)

- Generate $U_i \sim \text{Uniform}(0,1), \quad i = 1,\ldots,n;$
- Set

$$X_i = \left\{ -1 - \frac{1}{\lambda} - \frac{1}{\lambda} W_{-1}\left[\frac{-(1+\lambda)(1-U_i)^{\frac{1}{\alpha}} \mathrm{e}^{-1-\lambda}}{(\gamma U_i)^{\frac{1}{\alpha}} + (1-U_i)^{\frac{1}{\alpha}}} \right] \right\}, \quad i = 1,\ldots,n.$$

For $\alpha \neq \beta$, we applied the following algorithm for generating random data:

- Step 1. Generate random numbers u_i from $U \sim U(0,1)$ for $i = 1,\cdots,n$.
- Step 2. Select arbitrary values for parameters of EOLL-L distribution, i.e. α, β, γ and λ.
- Step 3. Solve numerically the non-linear equation

$$u_i = \frac{\left[1-(1+\frac{\lambda x_i}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x_i}}\right]^{\alpha}}{\left[1-(1+\frac{\lambda x_i}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x_i}}\right]^{\alpha} + \gamma\left[(1+\frac{\lambda x_i}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x_i}}\right]^{\beta}}, \tag{10}$$

and compute values of x_i for $i = 1,\cdots,n$.

2.3. Expansions for the density and cumulative distribution functions

In this subsection, two mixture representations of the pdf and cdf for EOLL-L are proposed. Despite the fact that the pdf and cdf of EOLL-L require mathematical functions that are widely available in modern statistical packages, frequently analytical and numerical derivations take advantage of power series representations for the pdf. Therefore, we use the concept of power series to calculate the useful expansions. Accordingly, the pdf of the EOLL-L distribution is given by

$$\left\{1-\left(1+\frac{\lambda x}{1+\lambda}\right)\mathrm{e}^{-\lambda x}\right\}^{\alpha} = \sum_{k=0}^{\infty} a_k \left\{1-\left(1+\frac{\lambda x}{1+\lambda}\right)\mathrm{e}^{-\lambda x}\right\}^{k}, \tag{11}$$

where $a_k = \sum_{i=k}^{\infty}(-1)^{i+k}\binom{\alpha}{i}\binom{i}{k}$ and

$$\left[1-\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\alpha} + \gamma\left[\left(1+\frac{\lambda}{1+\lambda}x\right)\mathrm{e}^{-\lambda x}\right]^{\beta} = \sum_{k=0}^{\infty} b_k \left\{1-\left(1+\frac{\lambda x}{1+\lambda}\right)\mathrm{e}^{-\lambda x}\right\}^{k}, \tag{12}$$

where $b_k = a_k + \gamma(-1)^k\binom{\beta}{k}$. Then, we can write

$$F(x) = \frac{\left\{1-\left(1+\frac{\lambda x}{1+\lambda}\right)\mathrm{e}^{-\lambda x}\right\}^{\alpha}}{\sum_{k=0}^{\infty} b_k \left\{1-\left(1+\frac{\lambda x}{1+\lambda}\right)\mathrm{e}^{-\lambda x}\right\}^{k}} = \sum_{k=0}^{\infty} c_k \left\{1-\left(1+\frac{\lambda x}{1+\lambda}\right)\mathrm{e}^{-\lambda x}\right\}^{k+\alpha}. \tag{13}$$

where $c_0 = \frac{1}{b_0}$ and for $k \geq 1$,

$$c_k = -b_0^{-1} \sum_{r=1}^{k} b_r c_{k-r}. \tag{14}$$

Hence, the cdf of the EOLL-L distribution can be written as

$$F(x) \quad = \quad \sum_{k=0}^{\infty} c_k G_{k+\alpha}(x). \tag{15}$$

where $G_{k+\alpha}(x)$ denotes the cdf of the generalized Lindley (exponentiated Lindley) distribution with parameters λ and $k+\alpha$.

Moreover, by differentiating from (15), the pdf of X can be expressed as

$$f(x) \quad = \quad \sum_{k=0}^{\infty} c_k g_{k+\alpha}(x). \tag{16}$$

where $g_{k+\alpha}(x)$ is the pdf of the generalized Lindley distribution with parameters λ and $k+\alpha$. Several properties of the EOLL-L distribution can be available from the cdf and pdf expansions, given in (15) and (16), respectively.

2.4. Moments and moment generating function

Some of the most important features and characteristics of a distribution can be investigated through moments (e.g., central tendency, dispersion, skewness, and kurtosis). In what follows, we present ordinary moments and the moment generating function (mgf) of the EOLL-L distribution. To find the ordinary moments (μ'_r), we use the following equation, which is introduced by Nadarajah et al. (2011) as

$$A(a,b,c,\delta) = \int_0^{\infty} x^c (1+x) \left[1 - \left(1 + \frac{bx}{b+1}\right) \mathrm{e}^{-bx}\right]^{a-1} \mathrm{e}^{-\delta x} dx. \tag{17}$$

From (17), we have

$$A(a,b,c,\delta) = \sum_{l=0}^{\infty} \sum_{r=0}^{l} \sum_{s=0}^{r+1} \binom{a-1}{l} \binom{l}{r} \binom{r+1}{s} \frac{(-1)^l b^r \Gamma(s+c+1)}{(1+b)^l (bl+\delta)^{c+s+1}}. \tag{18}$$

Using equations (15) and (16), we get the ordinary moments of the EOLL-L distribution as

$$\mu'_r = E[X^r] = \frac{\lambda^2}{1+\lambda} \sum_{k=0}^{\infty} (k+\alpha) c_k A(k+\alpha, \lambda, r, \lambda). \tag{19}$$

We now provide a formula for the conditional moments of the EOLL-L distribution. To this end, we use the following equation, which is introduced by Nadarajah et al. (2011) as

$$L(a,b,c,\delta,t) = \int_t^{\infty} x^c (1+x) \left[1 - \left(1 + \frac{bx}{b+1}\right) \mathrm{e}^{-bx}\right] \mathrm{e}^{-\delta x} dx. \tag{20}$$

Using the generalized binomial expansion, we have

$$L(a,b,c,\delta,t) = \sum_{l=0}^{\infty} \sum_{r=0}^{l} \sum_{s=0}^{r+1} \binom{a-1}{l} \binom{l}{r} \binom{r+1}{s} \frac{(-1)^l b^r \Gamma(s+c+1,(bl+\delta)t)}{(1+b)^l (bl+\delta)^{c+s+1}}, \tag{21}$$

where

$$\Gamma(a,x) = \int_x^\infty t^{a-1}\,\mathrm{e}^{-t}\,dt, \tag{22}$$

denotes the incomplete gamma function. From equations (16) and (21), we obtain the conditional moments of the EOLL-L distribution as

$$\mu'_r(t) = E\left[X^r | X > t\right] = \frac{\lambda^2}{1+\lambda}\sum_{k=0}^{\infty}(k+\alpha)\,c_k\,L(k+\alpha,\lambda,r,\lambda,t). \tag{23}$$

Moreover, the incomplete moments of the EOLL-L distribution can be obtained directly from (23).

Using (16) and (18), we can derive the mgf as follows:

$$M_X(t) = E\left[\mathrm{e}^{\mathrm{t}X}\right] = \frac{\lambda^2}{1+\lambda}\sum_{k=0}^{\infty}(k+\alpha)\,c_k\,A(k+\alpha,\lambda,0,\lambda-t).$$

Remark 1 *The central moments (μ_n) and cumulants (κ_n) of X are easily calculated from (19) (e.g. see Arellano-Valle et al., 2017, and Contreras-Reyes et al., 2021) as*

$$\mu_n = \sum_{k=0}^{n}(-1)^k\binom{n}{k}\mu_1'^k\,\mu'_{n-k} \qquad \text{and} \qquad \kappa_n = \mu'_n - \sum_{k=1}^{n-1}\binom{n-1}{k-1}\kappa_k\,\mu'_{n-k},$$

respectively, where $\kappa_1 = \mu'_1$. Thus, $\kappa_2 = \mu'_2 - \mu'^2_1$, $\kappa_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1$, etc.

From the ordinary moments and using (19) the mean, variance, skewness and kurtosis are calculated for different values of parameters in Table 1. From Table 1 it can be seen that skewness and kurtosis are very sensitive to changes in the shape parameters so, the importance of the proposed distribution can be concluded.

Table 1. Moments, skewness, and kurtosis of EOLL-L distribution for some parameters values

α	γ	β	λ	μ'_1	μ'_2	μ'_3	μ'_4	*Skewness*	*Kurtosis*
0.5	0.5	1.0	0.5	1.903213	9.702343	72.93571	707.2085	2.089507	53.199973
0.5	0.5	1.0	2.0	0.371702	0.428410	0.756338	1.754821	2.438598	3.1978419
0.5	0.5	2.0	0.5	1.248766	3.631016	15.00114	78.75589	1.775172	14.724114
0.5	0.5	2.0	2.0	0.225271	0.136348	0.124155	0.148577	2.190991	0.8235115
0.5	2.0	1.0	0.5	4.006446	27.10376	236.9726	2507.292	1.083841	49.493653
0.5	2.0	2.0	2.0	0.473993	0.383384	0.407392	0.531980	1.189489	0.7872935
3.0	1.5	1.0	0.5	4.985129	31.22020	242.6686	2311.769	1.464291	43.219819
1.5	0.5	2.0	1.5	0.535931	0.438362	0.494825	0.729813	1.666026	1.1710862
2.0	2.5	1.5	3.0	0.530031	0.370096	0.323693	0.345280	1.240077	0.5167241
2.0	0.5	0.5	1.0	1.409837	3.076611	9.717392	41.62731	2.03217	10.713407

3. Estimation

Point estimation is the first step of statistical inference on the unknown parameters of the underlying population. In order to find point estimations, there are different methods such as maximum likelihood estimation (MLE), least square and moment method. In the

present paper, we obtain the maximum likelihood, least square and weighted least-square estimations for the parameters of the EOLL-L distribution.

3.1. Maximum likelihood estimation

Let $X_1,\ldots,X_n$ be a random sample of size n from the EOLL-L$(\alpha,\beta,\gamma,\lambda)$ distribution. The log-likelihood function for the vector of parameters $\theta=(\alpha,\beta,\gamma,\lambda)^T$ can be written as

$$\begin{aligned} l(\theta) &= n\log\left(\frac{\gamma\lambda^2}{1+\lambda}\right)+\sum_{i=1}^{n}\log(1+x_i)+(\beta-1)\sum_{i=1}^{n}\log(1+\frac{\lambda\, x_i}{1+\lambda})-\beta\lambda\sum_{i=1}^{n}x_i \\ &+ (\alpha-1)\sum_{i=1}^{n}\log(q_i)+\sum_{i=1}^{n}\log\left[\alpha+(\beta-\alpha)q_i\right]-2\sum_{i=1}^{n}\log\left[q_i^{\alpha}+\gamma(1-q_i)^{\beta}\right] \quad (24) \end{aligned}$$

where $q_i=1-(1+\frac{\lambda}{1+\lambda}x_i)\mathrm{e}^{-\lambda x_i}$ is a transformed observation. The log-likelihood can be maximized by differentiating (24) and solving the nonlinear likelihood equations. The components of the score vector $U(\theta)$ are given by

$$\begin{aligned} U_{\lambda}(\theta) &= \frac{2n}{\lambda}-\frac{n}{1+\lambda}-\beta\sum_{i=1}^{n}x_i+(\beta-1)\sum_{i=1}^{n}\frac{x_i}{(1+\lambda)(1+\lambda+\lambda\, x_i)} \\ &+ (\alpha-1)\sum_{i=1}^{n}\frac{q_i^{(\lambda)}}{q_i}+(\beta-\alpha)\sum_{i=1}^{n}\frac{q_i^{(\lambda)}}{\alpha+(\beta-\alpha)q_i} \\ &-2\sum_{i=1}^{n}q_i^{(\lambda)}\ \frac{\alpha\, q_i^{\alpha-1}-\gamma\beta\,(1-q_i)^{\beta-1}}{q_i^{\alpha}+\gamma(1-q_i)^{\beta}}, \\ U_{\alpha}(\theta) &= \sum_{i=1}^{n}\log(q_i)+\sum_{i=1}^{n}\frac{1-q_i}{\alpha+(\beta-\alpha)q_i}-2\sum_{i=1}^{n}\frac{q_i^{\alpha}\log(q_i)}{q_i^{\alpha}+\gamma(1-q_i)^{\beta}}, \\ U_{\gamma}(\theta) &= \frac{n}{\gamma}-2\sum_{i=1}^{n}\frac{(1-q_i)^{\beta}}{q_i^{\alpha}+\gamma(1-q_i)^{\beta}}, \end{aligned}$$

$$\begin{aligned} U_{\beta}(\theta) &= \sum_{i=1}^{n}\log(1+\frac{\lambda\, x_i}{1+\lambda})-\lambda\sum_{i=1}^{n}x_i \\ &+ \sum_{i=1}^{n}\frac{q_i}{\alpha+(\beta-\alpha)q_i}-2\sum_{i=1}^{n}\frac{(1-q_i)^{\beta}\log(1-q_i)}{q_i^{\alpha}+\gamma(1-q_i)^{\beta}}. \end{aligned}$$

To construct a confidence interval and find test statistic for testing hypothesis on the parameters, the 4×4 observed information matrix $J=J(\theta)$ is required.

Under conditions that are fulfilled for parameters in the interior of the parameter space but not on the boundary, the asymptotic distribution of $\sqrt{n}(\hat{\theta}-\theta)$ is $N_4(0,I(\theta)^{-1})$, where $I(\theta)$ is the expected information matrix. In practice, we can replace $I(\theta)$ by the observed information matrix evaluated at $\hat{\theta}$ (say $J(\hat{\theta})$). We can construct approximate confidence intervals and confidence regions for the individual parameters and for the hazard and survival functions based on the multivariate normal $N_4(0,J(\hat{\theta})^{-1})$ distribution.

3.2. Ordinary and weighted least-square estimators

Let $x_{(1)}, x_{(2)}, x_{(3)}, \cdots, x_{(n)}$ denote the ordered sample of the random observations of size n from the EOLL-L distribution. By minimizing the following equation

$$\ell(\theta) = \sum_{i=1}^{n} \left(\frac{\left[1 - \left(1 + \frac{\lambda}{1+\lambda} x_{(i)}\right) e^{-\lambda x_{(i)}}\right]^{\alpha}}{\left[1 - \left(1 + \frac{\lambda}{1+\lambda} x_{(i)}\right) e^{-\lambda x_{(i)}}\right]^{\alpha} + \gamma\left[\left(1 + \frac{\lambda}{1+\lambda} x_{(i)}\right) e^{-\lambda x_{(i)}}\right]^{\beta}} - \frac{i}{n+1} \right)^2, \quad (25)$$

the least-square estimations (LSEs) of the EOLL-L distribution can be computed. Moreover, the weighted least square estimators (WLSEs) of the EOLL-L distribution can be derived by minimizing the following equation

$$\ell(\theta) = \sum_{i=1}^{n} \frac{(n+1)^2 (n+2)}{i(n-i+1)} \left(\frac{\left[1 - \left(1 + \frac{\lambda}{1+\lambda} x_{(i)}\right) e^{-\lambda x_{(i)}}\right]^{\alpha}}{\left[1 - \left(1 + \frac{\lambda}{1+\lambda} x_{(i)}\right) e^{-\lambda x_{(i)}}\right]^{\alpha} + \gamma\left[\left(1 + \frac{\lambda}{1+\lambda} x_{(i)}\right) e^{-\lambda x_{(i)}}\right]^{\beta}} - \frac{i}{n+1} \right)^2. \quad (26)$$

One can use the **optim** function in R software to minimize the (25) and (26). The partial derivatives of (25) and (26) with respect to α, β, γ and λ can be obtained from the authors upon request.

4. Simulation

In this section, a simulation study on the model parameters is investigated. The MLE, LSE, and WLSE methods are used for estimating the unknown parameters of the EOLL-L distribution and the performance of the methods are compared. The simulation procedure has been performed according to the following steps:

1. Set the sample size n and the vector of parameters $\theta = (\alpha, \beta, \gamma, \lambda)$.
2. Generate random observations from the $EOLL-L(\alpha, \beta, \gamma, \lambda)$ distribution with size n using Algorithm 1 in subsection 2.2.
3. Apply the generated random observations in Step 2 and estimate $\hat{\theta}$ by means of MLE, LSE and WLSE methods.
4. Repeat Steps 2 and 3 for N times.
5. Compute the mean relative estimates (MREs) and mean square errors (MSEs) using $\hat{\theta}$ and θ on the basis of the following equations:

$$MRE = \sum_{j=1}^{N} \frac{\hat{\theta}_{i,j}/\theta_i}{N}, \qquad MSE = \sum_{j=1}^{N} \frac{\left(\hat{\theta}_{i,j} - \theta_i\right)^2}{N},$$

where $\hat{\theta}_{i,j}$ for $i = 1, .., 4$ and $j = 1, ..., N$, is the estimation of ith element of parameter vector in jth iteration. The simulation results are obtained with R software. The chosen parameters of the simulation study are $\theta = (\alpha = 0.5, \beta = 2, \gamma = 1.5, \lambda = 2.5)$, $N = 1000$ and

$n = (50, 55, 60, ..., 500)$. We expect that MREs are closer to one when the MSEs are near zero. Figure 2 represents estimated MSEs and MREs based on the MLE, LSE and WLSE methods. As expected, MSEs and MREs of all estimates tend to zero and one for large n, respectively. Furthermore, it is deduced generally that the LSE method has better performance than the MLE method as well as the WLSE method to estimate EOLL-L parameters based on both MSE and MRE criteria even for the small sample size.

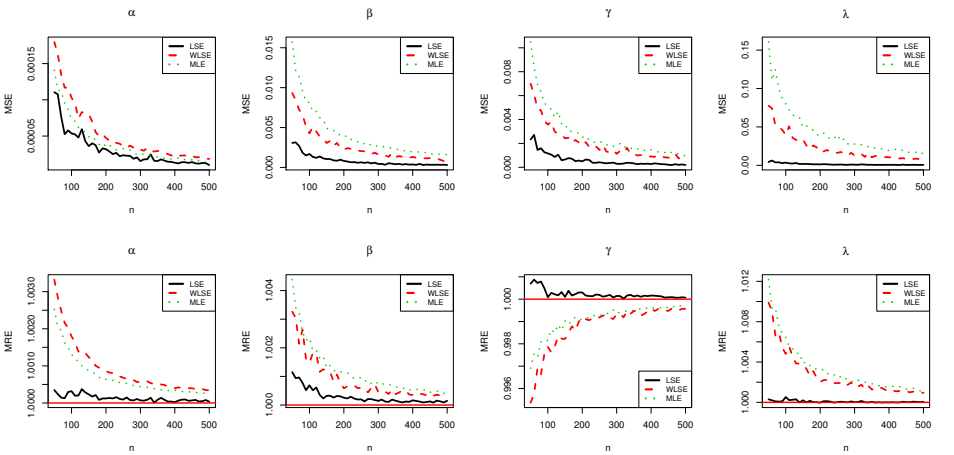

Figure 2. The behavior of MSEs and MREs of MLE, LSE and WLSE methods for different values of sample size.

5. Applications

In this section, we illustrate the fitting performance of the EOLL-L distribution using a real data sets. To evaluate the performance of the EOLL-L distribution, we recall a few extended of Lindley distribution such as: Power Lindley distribution, $PL(\beta,\lambda)$ (Ghitany et al. (2013)), Generalized Lindley, $GL(\alpha,\lambda)$, (Nadarajah et al. (2011)), Beta Lindley, $BL(\alpha,\beta,\lambda)$, (Merovci and Sharma (2014)), Exponentiated power Lindley distribution, $EPL(\alpha,\beta,\lambda)$, (Ashour and Eltehiwy (2015)), Odd log-logistic power Lindley distribution, $OLL-PL(\alpha,\beta,\lambda)$, (Alizadeh et al. (2017a)), Kumaraswamy Power Lindley, $Kw(\alpha,\beta,\gamma,\lambda)$, (Oluyede et al. (2016)), Odd Burr- Lindley, $OBu-L(\alpha,\beta,\lambda)$ (Altun et al. (2017)), Extended generalized Lindley, $EGL(\alpha,\gamma,\lambda)$, (Ranjbar et al. (2019)), Marshal-Olkin Lindley, $MOL(\gamma,\lambda)$, (Marshall and Olkin (1997)), New odd-log logistic Lindley, $NOLLL(\alpha,\beta,\lambda)$, (Alizadeh et al. (2018b)), Odd-log logistic Marshal-Olkin Lindley, $OLL-MOL(\alpha,\gamma,\lambda)$, (Alizadeh et al. (2017b)).

To compare the EOLL-L distribution with the above-mentioned distributions we consider several well-known criteria such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Cramer Von Mises (W^*) and Anderson-Darling (A^*) statistics. In addition, Kolmogorov-Smirnov (K-S) statistic with its corresponding p-value and minimum value of minus log-likelihood function (-Log(L)) are investigated for all distributions.

Furthermore, the likelihood ratio (LR) tests apply for evaluating the EOLL-L distribution with its sub-models. For example, the test of $H_0 : \beta = 1$ against $H_1 : \beta \neq 1$ is equivalent to comparing the EOLL-L with EGL, and the LR test statistic is given by

$$LR = 2\left[l(\hat{\alpha},\hat{\beta},\hat{\gamma},\hat{\lambda}) - l(\hat{\alpha^*},1,\hat{\gamma^*},\hat{\lambda^*})\right],$$

where $\hat{\alpha^*}, \hat{\gamma^*}$ and $\hat{\lambda^*}$ are the ML estimators under H_0 of α, γ and λ, respectively. It should be highlighted that the initial values of the parameters are quite important to obtain the correct MLEs of parameters. To avoid the local minima problem, we first obtain the parameter estimation of the Lindley distribution. Then, the estimated parameter of the Lindley distribution is used as the initial value of the parameter in all the mentioned extended of the Lindley distribution as well as the EOLL-L distribution. This approach is quite useful to obtain correct parameter estimates of extended distributions.

The data are the exceedances of flood peaks (in m^3/s) of the Wheaton River near Carcross in Yukon Territory, Canada. The data consist of 72 exceedances for the years 1958-1984 rounded to one decimal place. These data were analyzed by Akinsete et al. (2008). Throughout this subsection, we present the obtained results using the exceedances of the flood peaks data set.

Table 2. The exceedances of flood peaks data set

1.70	2.20	14.4	1.10	0.40	20.6	5.30	0.70	1.90	13.0	12.0	9.30	1.40
18.7	8.50	25.5	11.6	14.1	22.1	1.10	2.50	14.4	1.70	37.6	0.60	2.20
39.0	0.30	15.0	11.0	7.30	22.9	1.70	0.10	1.10	0.60	9.00	1.70	7.00
20.1	0.40	2.80	14.1	9.90	10.4	10.7	30.0	3.60	5.60	30.8	13.3	4.20
25.5	3.40	11.9	21.5	27.6	36.4	2.70	64.0	1.50	2.50	27.4	1.00	27.1
20.2	16.8	5.30	9.70	27.5	2.50	27.0						

The ML estimates and the goodness-of-fit test statistics are presented in Tables 3 and 4, respectively. From Table 4, the smallest values of AIC, A^*, W^* and $-l$ statistics and the largest p-value belong to the EOLL-L distribution. Although, the BIC of OLLPL is less than that of EOLLL, in general, the EOLL-L distribution outperforms the other competitive considered distributions on the basis of the criteria. The values of LR test statistics and

Table 3. The ML estimates and their standard errors (in parentheses) for first data set

Model	α	β	γ	λ
Lindley(λ)	–	–	–	0.153 (0.0128)
GL(α,λ)	0.508 (0.0767)	–	–	0.104 (0.01491)
PL(β,λ)	–	0.700 (0.0570)	–	0.338 (0.0559)
BL(α,β,λ)	0.555 (0.0983)	0.274 (0.2397)	–	0.333 (0.2723)
EPL(α,β,λ)	0.730 (0.2351)	0.915 (0.5956)	–	0.300 (0.2791)
OLLPL(α,β,λ)	0.183 (0.0222)	–	–	0.612 (0.0660)
KwL($\alpha,\beta,\gamma,\lambda$)	1.675 (2.4335)	0.453 (0.4323)	7.563 (11.7366)	0.279 (0.5225)
OBuL(α,β,λ)	24.91 (25.654)	0.024 (0.0326)	–	0.984 (0.1496)
EGL(α,γ,λ)	0.618 (0.1018)	–	2.770 (1.7047)	0.169 (0.0288)
MOL(γ,λ)	–	–	0.215 (0.1276)	0.090 (0.0246)
NOLLL(α,β,λ)	1.1735(0.1917)	–	0.171 (0.0238)	0.547 (0.0262)
OLLMOL(α,γ,λ)	0.6165(0.0880)	–	0.965 (0.4366)	0.180 (0.0470)
EOLLL($\alpha,\beta,\gamma,\lambda$)	1.113 (0.2132)	1.775 (0.4509)	0.176 (0.0244)	0.618 (0.0026)

Table 4. Goodness-of-fit test statistics for the data set

Model	AIC	BIC	p-value	W^*	A^*	$-l$
Lindley(λ)	530.423	532.700	0.001	0.139	0.852	264.211
GL(α,λ)	509.349	513.902	0.276	0.132	0.822	252.674
PL(β,λ)	508.443	512.996	0.405	0.123	0.766	252.103
BL(α,β,λ)	510.206	517.036	0.297	0.150	0.866	252.221
EPL(α,β,λ)	510.425	517.255	0.395	0.147	0.854	252.212
OLLPL(α,β,λ)	506.029	510.582	0.501	0.100	0.621	251.015
KwL($\alpha,\beta,\gamma,\lambda$)	512.221	521.328	0.371	0.152	0.866	252.110
OBuL(α,β,λ)	511.212	520.319	0.401	0.140	0.799	251.606
EGL(α,γ,λ)	508.931	515.761	0.174	0.101	0.662	251.465
MOL(γ,λ)	522.570	527.124	0.024	0.214	1.208	259.285
NOLLL(α,β,λ)	506.505	513.335	0.035	0.095	0.517	250.252
OLLMOL(α,γ,λ)	508.023	514.853	0.517	0.101	0.623	251.011
EOLLL($\alpha,\beta,\gamma,\lambda$)	502.327	511.433	0.958	0.041	0.249	247.163

Table 5. The LR test results for the data set

	Hypotheses	LR	p-value
EOLL-L versus Lindley	$H_0:\alpha=\beta=\gamma=1$	34.0966	<0.0001
EOLL-L versus OLL-L	$H_0:\alpha=\beta,\ \gamma=1$	7.7022	0.0212
EOLL-L versus MOL	$H_0:\alpha=\beta=1$	24.2439	<0.0001
EOLL-L versus NOLLL	$H_0:\gamma=1$	6.1782	0.01293
EOLL-L versus OLL-MOL	$H_0:\alpha=\beta$	7.6962	0.00553

their corresponding p-values are exhibited in Table 5. From Table 5, we observe that the computed p-values are too small so we reject all the null hypotheses and conclude that the EOLL-L fits the data set better than the considered sub-models according to the LR criterion.

We also plotted the fitted pdfs, cdfs and P-P plots of the considered models for the sake of visual comparison, in Figures 4 and 5, respectively. Figure 4 suggests that the EOLL-L fits the skewed data very well. Figures 5 shows that the plotted points for the EOLL-L distribution best capture the diagonal line in the probability plots. Therefore, the EOLL-L distribution can be considered as an appropriate model for fitting the first data set.

6. Bayesian estimation

The Bayesian inference procedure has been taken into consideration by many statistical researchers, especially researchers in the field of survival analysis and reliability engineering. In this section, a complete sample data is analysed through the Bayesian point of view. We assume that the parameters α, β, γ and λ of the $EOLL-L$ distribution have independent prior distributions as

$$\alpha \sim Gamma(a,b), \gamma \sim Gamma(c,d), \lambda \sim Gamma(e,f), \beta \sim Gamma(g,h)$$

where a,b,c,d,e, f, g and h are positive. Hence, the joint prior density function is formulated as follows:

$$\pi(\alpha,\beta,\gamma,\lambda) = \frac{b^a d^c f^e h^g}{\Gamma(a)\Gamma(c)\Gamma(e)\Gamma(g)} \alpha^{a-1}\beta^{h-1}\gamma^{c-1}\lambda^{e-1}e^{-(b\alpha+h\beta+d\gamma+f\lambda)}. \quad (27)$$

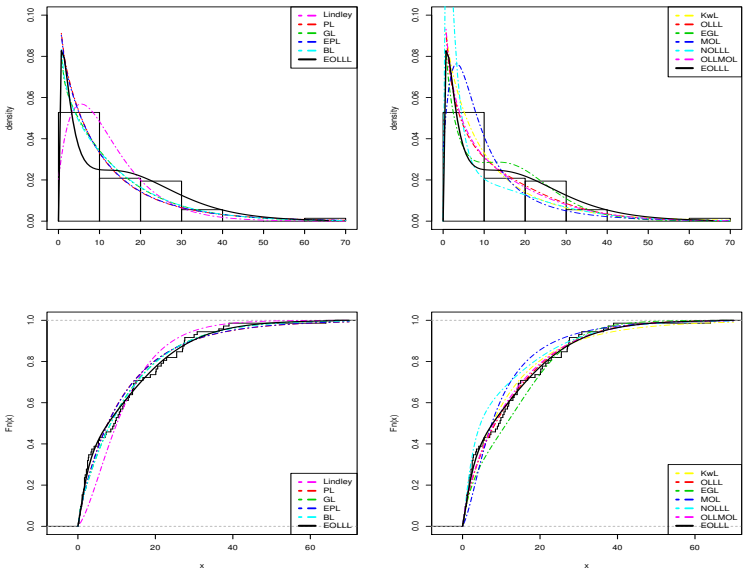

Figure 3. Fitted pdfs and cdfs of the distributions for the data set

In the Bayesian estimation, according to which we do not know the actual value of the parameter, we may be adversely affected by loss when we choose an estimator. This loss can be measured by a function of the parameter and the corresponding estimator.

Five well-known loss functions and associated Bayesian estimators and corresponding posterior risks are presented in Table 6. For more details, the reader can refer to Calabria and Pulcini (1996). Next, we provide the posterior probability distributions for a complete

Table 6. Bayes estimator and posterior risk under different loss functions

Loss function	Bayes estimator	Posterior risk
$L_1 = SELF = (\theta - d)^2$	$E(\theta\|x)$	$Var(\theta\|x)$
$L_2 = WSELF = \frac{(\theta-d)^2}{\theta}$	$(E(\theta^{-1}\|x))^{-1}$	$E(\theta\|x) - (E(\theta^{-1}\|x))^{-1}$
$L_3 = MSELF = \left(1 - \frac{d}{\theta}\right)^2$	$\frac{E(\theta^{-1}\|x)}{E(\theta^{-2}\|x)}$	$1 - \frac{E(\theta^{-1}\|x)^2}{E(\theta^{-2}\|x)}$
$L_4 = PLF = \frac{(\theta-d)^2}{d}$	$\sqrt{E(\theta^2\|x)}$	$2\left(\sqrt{E(\theta^2\|x)} - E(\theta\|x)\right)$
$L_5 = KLF = \left(\sqrt{\frac{d}{\theta}} - \sqrt{\frac{\theta}{d}}\right)$	$\sqrt{\frac{E(\theta\|x)}{E(\theta^{-1}\|x)}}$	$2\left(\sqrt{E(\theta\|x)E(\theta^{-1}\|x)} - 1\right)$

data set. Let us we define the function φ as

$$\varphi(\alpha,\beta,\gamma,\lambda) = \alpha^{a-1}\beta^{h-1}\gamma^{c-1}\lambda^{e-1}e^{-(b\alpha+h\beta+d\gamma+f\lambda)},\ \alpha>0,\ \beta>0,\ \gamma>0,\ \lambda>0.$$

The joint posterior distribution in terms of a given likelihood function $L(data)$ and joint prior distribution $\pi(\alpha,\beta,\gamma,\lambda)$ is defined as

$$\pi^*(\alpha,\beta,\gamma,\lambda|data) \propto \pi(\alpha,\beta,\gamma,\lambda)L(data). \tag{28}$$

Table 7. Bayesian estimates $\widehat{\theta}$ and their posterior risks $r_{\widehat{\theta}}$ of the parameters under different loss functions based on the flood peaks data.

Data	Flood peaks			
Bayesian estimation				
Loss function	$\widehat{\alpha}\ (r_{\widehat{\alpha}})$	$\widehat{\beta}\ (r_{\widehat{\beta}})$	$\widehat{\gamma}\ (r_{\widehat{\gamma}})$	$\widehat{\lambda}\ (r_{\widehat{\lambda}})$
SELF	1.5331 (0.0863)	0.1852 (0.0012)	1.3004 (0.0386)	0.5993 (0.0076)
WSELF	1.4771 (0.0561)	0.1791 (0.0061)	1.2702 (0.0302)	0.5858 (0.0135)
MSELF	1.4217 (0.0375)	0.1735 (0.0312)	1.2396 (0.0241)	0.6056 (0.0247)
PLF	1.5610 (0.0557)	0.1886 (0.0067)	1.3152 (0.0295)	0.6056 (0.0126)
KLF	1.5049 (0.0376)	0.1821 (0.0338)	1.2852 (0.0236)	0.5925 (0.0229)

Table 8. Credible and *HPD* intervals of the parameters α, β, γ and λ for the flood peaks.

	Credible interval	HPD interval
α	(1.329, 1.737)	(0.957, 2.068)
β	(1.161, 1.429)	(0.949, 1.703)
γ	(0.160, 0.205)	(0.124, 0.254)
λ	(0.542, 0.662)	(0.428, 0.760)

Hence, we get joint posterior density of parameters α, β, γ and λ for complete sample data by combining the likelihood function and joint prior density (27). Therefore, the joint posterior density function is given by

$$\pi^*(\alpha,\beta,\gamma,\lambda|\underline{x}) \quad = \quad K\varphi(\alpha,\beta,\gamma,\lambda)L(\underline{x},\xi) \tag{29}$$

where

$$L(\underline{x};\xi)=\prod_{i=1}^{n}\frac{\gamma\lambda^2(1+x_i)\mathrm{e}^{-\beta\lambda\,\mathrm{x_i}}\left[1-(1+\frac{\lambda\mathrm{x_i}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x_i}}\right]^{\alpha-1}\left\{\alpha+(\beta-\alpha)\left[1-(1+\frac{\lambda\mathrm{x_i}}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x_i}}\right]\right\}}{(1+\lambda)\left\{\left[1-(1+\frac{\lambda\,x_i}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x_i}}\right]^{\alpha}+\gamma\left[(1-(1+\frac{\lambda\,x_i}{1+\lambda})\mathrm{e}^{-\lambda\,\mathrm{x_i}})\beta\right]\right\}^2}. \tag{30}$$

and K is given as

$$K^{-1} \quad = \quad \int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty \varphi(\alpha,\beta,\gamma,\lambda)L(\underline{x},\xi)d\alpha d\beta d\gamma d\lambda.$$

It is clear from equation (29) that there is no closed form for the Bayesian estimators under the five loss functions described in Table 6, so we suggest using an *MCMC* procedure based on 10000 replicates to compute Bayesian estimators. The corresponding Bayesian point and interval estimation and posterior risk are provided in Tables 7 and 8 for the flood peaks data set. Table 8 provides 95% credible and *HPD* intervals for each parameter of the $EOLL-L$ distribution. The posterior samples are extracted using Gibbs sampling technique. Moreover, we provide the posterior summary plots in Figure 4. These plots confirm that the convergence of Gibbs sampling process occurred.

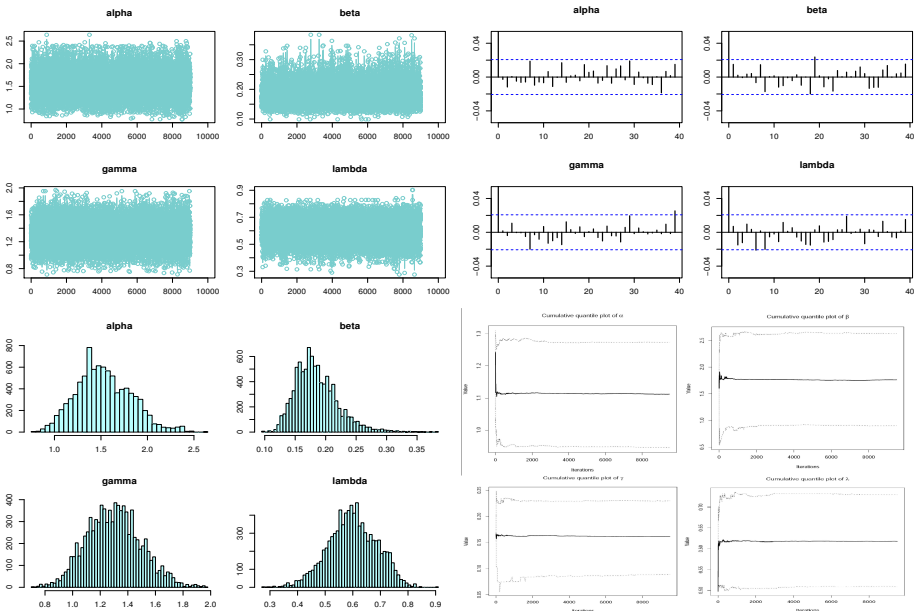

Figure 4. Plots of Bayesian analysis and performance of Gibbs sampling for the flood peaks data set.

7. Conclusion

In this paper, a new distribution which is called extended odd log-logistic-Lindley (EOLL-L) distribution was introduced. The statistical properties of the EOLL-L distribution including the hazard function, quantile function, moments, incomplete moments and generating functions and maximum likelihood estimation for the model parameters were given. Simulation studies were conducted to examine the performance of this distribution. We also presented applications of this new distribution for two real-life data sets in order to illustrate the usefulness of the distribution. Finally, the Bayesian estimation and Gibbs sampling procedure for the considered data sets were discussed.

References

Abouammoh, A., Alshangiti, A. M. and Ragab, I., (2015). A new generalized Lindley distribution. *Journal of Statistical Computation and Simulation*, 85(18), pp. 3662–3678.

Afify, A. Z., Cordeiro, G. M., Maed, M. E., Alizadeh, M., Al-Mofleh, H. and Nofal, Z. M., (2019). The generalized odd lindley-g family: properties and applications. *Anais da Academia Brasileira de Ciencias*, 91(3).

Akinsete, A., Famoye, F. and Lee, C., (2008). The beta-pareto distribution. *Statistics*, 42(6), pp. 547–563.

Alizadeh, M., Afshari, M., Hosseini, B. and Ramires, T. G., (2018a). Extended exp-g family of distributions: Properties and applications. *Communication in Statistics-Simulation and Computation*, accepted.

Alizadeh, M., Afify, A. Z., Eliwa, M. and Ali, S., (2019). The odd log-logistic lindley-g family of distributions:properties, bayesian and non-bayesian estimation with applications. *Computational Statistics*, pp. 1–28.

Alizadeh, M., Altun, E., Ozel, G., Afshari, M., and Eftekharian, A., (2018b). A new odd log-logistic lindley distribution with properties and applications. *Sankhya A*, 81(2), pp. 323–346.

Alizadeh, M., K MirMostafaee, S., Altun, E., Ozel, G. and Khan Ahmadi, M., (2017a). The odd log-logistic marshall-olkin power lindley distribution: Properties and applications. *Journal of Statistics and Management Systems*, 20(6), pp. 1065–1093.

Alizadeh, M., Ozel, G., Altun, E., Abdi, M. and Hamedani, G. (2017b). The odd log-logistic marshall-olkin lindley model for lifetime data. *Journal of Statistical Theory and Applications*, 16(3), pp. 382–400.

Alizadeh, M., Afshari, M., Cordeiro, G. M., Ramaki, Z., Contreras-Reyes, J. E., Dirnik, F. and Yousof, H. M., (2025). A Weighted Lindley Claims Model with Applications to Extreme Historical Insurance Claims. *Stats*, 8(1), p. 8.

Altun, G., Alizadeh, M., Altun, E. and Ozel, G., (2017). Odd burr lindley distribution with properties and applications. *Hacettepe Journal of Mathematics and Statistics*, 46(2), pp. 255–276.

Alzaatreh, A., Lee, C. and Famoye, F., (2013). A new method for generating families of continuous distributions. *Metron*, 71(1), pp. 63–79.

Arellano-Valle, R. B.; Contreras-Reyes, J. E.; Stehlík, M., (2017). Generalized skew-normal negentropy and its application to fish condition factor time series. *Entropy*, 19, p. 528.

Asgharzadeh, A., Bakouch, H. S., Nadarajah, S., Sharafi, F., et al., (2016). A new weighted lindley distribution with application. *Brazilian Journal of Probability and Statistics*, 30(1), pp. 1–27.

Asgharzadeh, A., Nadarajah, S. and Sharafi, F., (2018). Weibull lindley distribution. *REVSTAT–Statistical Journal*, 16(1), pp. 87–113.

Ashour, S. K., Eltehiwy, M. A., (2015). Exponentiated power lindley distribution. *Journal of advanced research*, 6(6), pp. 895–905.

Calabria, R., Pulcini, G., (1996). Point estimation under asymmetric loss functions for left-truncated exponential samples. *Communications in Statistics-Theory and Methods*, 25(3), pp. 585–600.

Contreras-Reyes, J. E., Kahrari, F. and Cortés, D. D., (2021). On the modified skew-normal-Cauchy distribution: Properties, inference and applications. *Communications in Statistics-Theory and Methods*, 50(15), pp. 3615-3631.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J. and Knuth, D. E., (1996). On the lambertw function. *Advances in Computational mathematics*, 5(1), pp. 329–359.

Galambos, J., Kotz, S., (2006). *Characterizations of Probability Distributions.: A Unified Approach with an Emphasis on Exponential and Related Models*, vol. 675. Springer.

Ghitany, M., Al-Mutairi, D. K., Balakrishnan, N. and Al-Enezi, L., (2013). Power lindley distribution and associated inference. *Computational Statistics Data Analysis*, 64, pp. 20–33.

Ghitany, M., Atieh, B. and Nadarajah, S., (2008). Lindley distribution and its application. *Mathematics and computers in simulation*, 78(4), pp. 493–506.

Gleaton, J. U., Lynch, J. D., (2010). Extended generalized log-logistic families of lifetime distributions with an application. *J. Probab. Stat. Sci*, 8, pp. 1–17.

Gomes-Silva, F. S., Percontini, A., de Brito, E., Ramos, M. W., Venancio, R., and Cordeiro, G. M., (2017). The odd lindley-g family of distributions. *Austrian Journal of Statistics*, 46(1), pp. 65–87.

Jodr, P., (2010). Computer generation of random variables with lindley or poisson–lindley distribution via the lambert w function. *Mathematics and Computers in Simulation*, 81(4), pp. 851–859.

Kim, J. H., Jeon, Y., (2013). Credibility theory based on trimming. *Insurance: Mathematics and Economics*, 53(1), pp. 36–47.

Kotz, S., Shanbhag, D., (1980). Some new approaches to probability distributions. *Advances in Applied Probability*, pp. 903–921.

Lindley, D. V., (1958). Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(1), pp. 102–107.

Marshall, A. W., Olkin, I., (1997). A new method for adding a parameter to a family of distributions with application to the exponential and weibull families. *Biometrika*, 84(3), pp. 641–652.

Merovci, F., Sharma, V. K., (2014). The beta-lindley distribution: properties and applications. *Journal of Applied Mathematics*.

Murthy, D. P., Xie, M. and Jiang, R., (2004). *Weibull models*, vol. 505, John Wiley Sons.

Nadarajah, S., Bakouch, H. S. and Tahmasbi, R., (2011). A generalized lindley distribution. *Sankhya B*, 73(2), pp. 331–359.

Oluyede, B. O., Yang, T. and Makubate, B., (2016). A new class of generalized power lindley distribution with applications to lifetime data. *Asian Journal of Mathematics and Applications*, 2016, p.1.

Ozel, G., Alizadeh, M., Cakmakyapan, S., Hamedani, G., Ortega, E. M. and Cancho, V. G., (2017). The odd log- logistic lindley poisson model for lifetime data. *Communications in Statistics-Simulation and Computation*, 46(8), pp. 6513–6537.

Ranjbar, V., Alizadeh, M. and Altun, E., (2019). Extended generalized lindley distribution: Properties and applications. *Journal of Mathematical Extension*, 13(1), pp. 117–142.

Zakerzadeh, H., Dolati, A., (2009). Generalized lindley distribution. *Journal of Mathematical Extension*, 3(2), pp. 1–17.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 105–118, https://doi.org/10.59139/stattrans-2026-006
Received – 24.02.2025; accepted – 27.08.2025

Modeling the impact of demand, supply, and budget constraints on consumer preferences

Julia Fidler[1], **Łukasz Matysiak**[2]

Abstract

Mathematical risk modeling in a market economy has become a key tool for analyzing consumer behavior under conditions of unstable prices and shifting supply. In this study, we combine Paul Samuelson's classical theory of revealed preferences with dynamic demand and supply mechanisms, using Afriat's theorem and extensions by Varian and Mas-Colell to construct utility functions without survey data. Critical voices (e.g. Dryzek) prompt a reexamination of the assumptions of full information and fixed preferences, inspiring our proposal of the $F(w,z)$ function that accounts for the strength of market fluctuations. Empirical simulations and an analysis of market equilibrium stability yield new insights for economic policy and marketing strategies.

Key words: budget, demand, preference, price, supply.

1. Introduction

In the face of sudden commodity price swings and growing market uncertainty, integrating risk modeling with demand and supply mechanisms allows for more accurate predictions of consumer behavior. Policymakers and practitioners demand tools that combine quantitative risk analysis with classical economic assumptions to respond effectively to sudden price shocks and income changes.

The classical foundations of revealed preference theory trace back to Samuelson's paper (1938), who based assumptions about consumer rationality and consistency on observed choices. In Afriat's article (1967) the author demonstrated that rationality conditions enable the construction of a utility function without survey data, and Varian and Mas-Colell extended this framework to accommodate more complex preference structures in their articles (1978, 1982). Critics such as Dryzek, in his book (2014), and other behavioral economists highlight unrealistic assumptions of full information, fixed preferences, and the neglect of socio-cultural factors. Recent attempts to merge demand and supply analysis with revealed preference theory—particularly in the papers of Fidler & Matysiak (2024, 2025) - have not yet produced a cohesive model that integrates dynamic market parameters in a single framework.

[1]Nicolaus Copernicus University, Torun. Poland. E-mail: juliafidler@mat.umk.pl.
ORCID: https://orcid.org/0009-0003-4895-4663.
[2]Military University of Technology, Warsaw. Poland. E-mail: lukasz.matysiak@wat.edu.pl.
ORCID: https://orcid.org/0000-0002-0819-1468.

Despite numerous extensions to the classical approach, existing models have not formally integrated the strength of demand and supply with income variability and market risk. To address this gap, this paper:

1. Extends the classical utility-maximization model under a budget constraint to incorporate dynamic demand and supply parameters.

2. Introduces the function $F(w,z)$, where w denotes the baseline preference scale and z measures the strength of market changes (demand/supply).

3. Analyzes market equilibrium stability and consumer sensitivity to price and income fluctuations.

In Section 2 we present a detailed literature review focusing on the evolution of revealed preference theory and critiques of the full-information assumption. Section 4 develops the mathematical model, including the definition of the function $F(w,z)$. Section 5 presents simulation results and a stability analysis, while Section 6 reviews classical consumer theory through indifference curves and budget constraints.

2. The impact of supply and demand

First, let us define a utility function $U(x_1,x_2)$, where x_1 and x_2 are the quantities of two goods. A common choice may be the Cobb-Douglass function:

$$U(x_1,x_2) = x_1^{\alpha} \cdot x_2^{\beta}. \tag{1}$$

We denote the consumer's income by M and the prices of goods are p_1 and p_2. The budget constraint reads:

$$p_1x_1 + p_2x_2 \leq M \tag{2}$$

The relationship between demand and prices and income can be expressed by the demand function:

$$x_1 = D_1(p_1,p_2,M) \tag{3}$$

$$x_2 = D_2(p_1,p_2,M) \tag{4}$$

Market equilibrium occurs when supply S equals demand D. For each good we have:

$$S_1 = D_1 \tag{5}$$

$$S_2 = D_2 \tag{6}$$

Given the above, we can formulate preferences as a function of demand, supply, and market equilibrium.

Suppose that preferences P depend on the demand/supply ratio $\frac{D}{S}$:

$$P = f\left(\frac{D_1(p_1,p_2,M)}{S_1}, \frac{D_2(p_1,p_2,M)}{S_2}\right) \tag{7}$$

We can also include additional factors, such as the price elasticity of demand, to more precisely model consumer preferences.

This is just a simplified model that can be expanded with additional variables and more complex features to better reflect economic reality.

From the article by Fidler and Matysiak (2025), we assumed that we can study consumer preferences with one of four functions F depending on the situation. And we can assume this here as well, which we will discuss at the end of this section.

$$F(w,z) = \begin{cases} (1)\ w+az, \\ (2)\ w+a\ln(1+z), \\ (3)\ w+az^2, \\ (4)\ w+ae^z. \end{cases} \tag{8}$$

where w is the initial preference value, a, z are some indicators under consideration.

Example 2.1 (Preferences via D/S Ratios). Let $p_1 = 2$, $p_2 = 3$, $M = 30$, and $U(x_1,x_2) = x_1^{0.5}x_2^{0.5}$. Solving yields $D_1 = 7.5$, $D_2 = 5$. With $S_1 = 15$, $S_2 = 10$:

$$z_1 = \frac{D_1}{S_1} = 0.5, \quad z_2 = \frac{D_2}{S_2} = 0.5. \tag{9}$$

Set the baseline $w = 1$ and weights $a = b = 1$. Then:

(a) Linear:

$$P = w + z_1 + z_2 = 1 + 0.5 + 0.5 = 2. \tag{10}$$

(additive weighting)

(b) Logarithmic:

$$P = w + \ln(1+z_1) + \ln(1+z_2) = 1 + 2\ln(1.5) \approx 1.81. \tag{11}$$

(dampens extremes)

(c) Exponential:

$$P = w + e^{z_1} + e^{z_2} = 1 + 2e^{0.5} \approx 4.30. \tag{12}$$

(exaggerates moderate changes)

(d) Cobb–Douglas:

$$P = w \cdot z_1 \cdot z_2 = 1 \cdot 0.5 \cdot 0.5 = 0.25. \tag{13}$$

(multiplicative interaction)

Remark 2.2. The function f is just a placeholder for any mapping of the two demand–supply ratios into a single preference value. In our setup

$$P \;=\; f\big(z_1,z_2\big) \;=\; f\Big(\tfrac{D_1}{S_1},\tfrac{D_2}{S_2}\Big), \tag{14}$$

where

$$z_1 = \frac{D_1}{S_1}, \quad z_2 = \frac{D_2}{S_2}. \tag{15}$$

In the numerical example we have

- $\frac{D_1}{S_1} = 0.5$: demand for good 1 is 50% of its supply,
- $\frac{D_2}{S_2} = 0.5$: demand for good 2 is 50% of its supply.

Depending on how sensitively we want to weight those ratios, f could be an arithmetic mean, a weighted sum, a nonlinear mapping, etc. For instance, the simple arithmetic-mean choice is

$$P = \frac{\frac{D_1}{S_1} + \frac{D_2}{S_2}}{2}, \tag{16}$$

which with $z_1 = z_2 = 0.5$ yields $P = 0.5$.

Remark 2.3. Below are the extreme values for each of the four canonical forms $F(w, z_1, z_2)$, given nonnegative parameters a, b, k, m. We write

$$z_1 = k, \quad z_2 = m \tag{17}$$

for the "max-ratio" scenario.

- **Linear** $P = w + a\,z_1 + b\,z_2$
  $P_{\min} = w$ at $z_1 = z_2 = 0$, $P_{\max} = w + a\,k + b\,m$.
- **Logarithmic** $P = w + a\ln(1+z_1) + b\ln(1+z_2)$
  domain requires $z_i > -1$, $P_{\min} \to -\infty$ as $z_i \to -1^+$, $P_{\max} = w + a\ln(1+k) + b\ln(1+m)$.
- **Exponential** $P = w + a\,e^{z_1} + b\,e^{z_2}$
  $P_{\min} = w + a + b$ at $z_1 = z_2 = 0$, $P_{\max} = w + a\,e^{k} + b\,e^{m}$.
- **Cobb–Douglas** $P = w \cdot z_1^{a} \cdot z_2^{b}$
  $P_{\min} = 0$ if either $z_1 = 0$ or $z_2 = 0$, $P_{\max} = w \cdot k^{a}\,m^{b}$.

Each form highlights a different behavior: the linear case grows proportionally, the logarithmic can diverge to $-\infty$ near its lower domain limit, the exponential has a positive floor and rapid growth, and Cobb–Douglas vanishes when either ratio is zero but rises multiplicatively otherwise.

Studying consumer preferences through shifts in supply and demand yields key insights into price sensitivity and purchasing behavior; however, a truly comprehensive analysis must also incorporate income effects that determine consumers' purchasing power, cultural and social influences that shape preferences beyond pure economic criteria, the availability of substitutes and market competition influencing perceived value, market equilibrium conditions (such as surplus or shortage) that alter purchasing dynamics, and heterogeneity across social and demographic groups that affects behavior. Historical sales and price data

allow us to reconstruct how changes in supply and demand translated into actual consumer choices, while incorporating income variability enables us to model the impact of budget shifts on the composition of consumers' baskets. Finally, examining the interaction between economic variables and socio-cultural factors—such as emerging trends or segment-specific tastes—completes the full picture of the drivers behind purchase decisions. This expanded approach enhances our ability to capture the complexity of the mechanisms shaping consumer preferences and produces more reliable predictions of their responses to price shocks or income changes.

Example 2.4. Assume the consumer's utility function is $U(x_1,x_2) = x_1^{0.5}x_2^{0.5}$ and the budget constraint is $2x_1 + 3x_2 \leq 30$. The Marshallian demand functions then read

$$x_1 = \frac{0.5M}{p_1} = \frac{0.5 \cdot 30}{2} = 7.5, \qquad x_2 = \frac{0.5M}{p_2} = \frac{0.5 \cdot 30}{3} = 5. \tag{18}$$

Suppose a shift in preferences raises the price of apples to $p_1 = 3$. The new demands become

$$x_1 = \frac{0.5M}{p_1} = \frac{0.5 \cdot 30}{3} = 5, \qquad x_2 = \frac{0.5M}{p_2} = \frac{0.5 \cdot 30}{3} = 5. \tag{19}$$

At the higher apple price, the consumer buys fewer apples and reallocates expenditure toward oranges. We can then model preferences directly as a function of demands:

$$P = f(D_1, D_2), \tag{20}$$

so that in this case $P = f(5,5)$. Finally, to obtain the explicit form of the preference function in the spirit of Remark 2.3, replace the ratios D_1/S_1 and D_2/S_2 with D_1 and D_2 respectively.

When we study the effect of supply itself on consumer preferences, we can choose how that supply is modeled in relation to preferences.

Direct Supply Relationship:

$$W = f(S_1, S_2) \tag{21}$$

In this case, greater availability of goods may increase consumer preferences, since more goods on the market mean more choices and potentially higher satisfaction.

Inverse Supply Relationship:

$$W = f\left(\tfrac{1}{S_1}, \tfrac{1}{S_2}\right) \tag{22}$$

Here, scarcity drives up the valuation: lower S_1 or S_2 makes each unit more coveted, raising consumer preference for the rarer good.

The choice of formula hinges on your research context and the behavioral assumptions you adopt. To derive a concrete preference function, replace $\frac{D_1}{S_1}$ by S_1, and $\frac{D_2}{S_2}$ by S_2 (or by $\frac{1}{S_1}$ and $\frac{1}{S_2}$ for the inverse relationship) in the functional forms of Remark 2.3.

At the end of this section, we discuss modeling changes in consumer preferences under the joint influence of demand and supply. This framework is motivated by the article by Fidler and Matysiak (2025).

Let us use our 4 formulas introduced at the beginning of this section (see the article Fidler and Matysiak (2025)). We define

$$F(w,z) \tag{23}$$

where w denotes baseline preference and z captures the demand–supply influence. The parameter a measures the sensitivity of preferences to z.

Formulas:

1. $F(w,z) = w + az$
2. $F(w,z) = w + a\ln(1+z)$
3. $F(w,z) = w + az^2$
4. $F(w,z) = w + ae^z$

Let us look at the example below:

Example 2.5. Suppose the consumer prefers product P at $w = 5$. Assume that the demand for P is $D = 7$ and the supply is $S = 10$. Let $a = 1$ and let z represent the influence of demand and supply, i.e. $z = \frac{D}{S} = \frac{7}{10} = 0.7$.

From the first formula, $F(w,z) = w + az$:

$$F(5,0.7) = 5 + 1 \cdot 0.7 = 5.7. \tag{24}$$

From the second formula, $F(w,z) = w + a\ln(1+z)$:

$$F(5,0.7) = 5 + 1 \cdot \ln(1.7) \approx 5 + 0.531 = 5.531. \tag{25}$$

From the third formula, $F(w,z) = w + az^2$:

$$F(5,0.7) = 5 + 1 \cdot (0.7)^2 = 5 + 0.49 = 5.49. \tag{26}$$

From the fourth formula, $F(w,z) = w + ae^z$:

$$F(5,0.7) = 5 + 1 \cdot e^{0.7} \approx 5 + 2.014 = 7.014. \tag{27}$$

Different models produce different results, illustrating how nonlinear transformations of z (logarithmic, quadratic, exponential) affect the strength of preference change. For more details, see the article by Fidler and Matysiak (2025).

Remark 2.6. In the linear, logarithmic, and quadratic cases, a fixed point $w^* = F(w^*, z^*)$ occurs only when $z = 0$ (assuming $a \neq 0$). For the exponential model, no such fixed point exists, so preferences always shift under the influence of z.

3. Empirical example with real demand–supply volume data

To validate our dynamic preference-update model on real figures, we use annual statistics (Statistics Poland - from Poland, Eurostat) converted into monthly volumes for 2023.

We assume:

- Annual domestic apple production: 3 600 000 t; annual apple consumption (after exports): 2 700 000 t.
- Annual banana imports: 494 000 t; annual banana consumption: 460 000 t.
- Monthly volumes are seasonally adjusted and averaged.

For each month t, compute the demand–supply ratios

$$z_{1,t} = \frac{D_{\text{apples},t}}{S_{\text{apples},t}}, \quad z_{2,t} = \frac{D_{\text{bananas},t}}{S_{\text{bananas},t}}, \quad \bar{z}_t = \frac{z_{1,t} + z_{2,t}}{2}. \tag{28}$$

We then update the preference index linearly:

$$w_0 = 5, \qquad w_t = w_{t-1} + a\,\bar{z}_t, \quad a = 1. \tag{29}$$

Table 1. Monthly demand–supply ratios and w_t (2023)

Month	$S_{\text{apples},t}$ [t]	$D_{\text{apples},t}$ [t]	$z_{1,t}$	$S_{\text{bananas},t}$ [t]	$D_{\text{bananas},t}$ [t]	$z_{2,t}$	$\bar{z}_t$	w_t
Jan	280 000	210 000	0.750	42 000	38 000	0.905	0.828	5.828
Feb	270 000	200 000	0.741	41 000	37 000	0.902	0.822	6.650
Mar	260 000	205 000	0.788	42 000	38 500	0.917	0.853	7.503
Apr	250 000	200 000	0.800	42 000	38 000	0.905	0.853	8.356
May	240 000	195 000	0.813	43 000	39 000	0.907	0.860	9.216
Jun	230 000	190 000	0.826	44 000	40 000	0.909	0.868	10.084
Jul	220 000	185 000	0.841	44 000	41 000	0.932	0.887	10.972
Aug	230 000	190 000	0.826	45 000	41 000	0.911	0.869	11.841
Sep	250 000	200 000	0.800	44 000	40 000	0.909	0.854	12.695
Oct	300 000	220 000	0.733	42 000	38 000	0.905	0.819	13.514
Nov	310 000	215 000	0.694	41 000	37 000	0.902	0.798	14.312
Dec	320 000	210 000	0.656	42 000	38 000	0.905	0.781	15.093

We observe that w_t grows from 5.00 to 15.09 over the year, driven by sustained demand pressure ($\bar{z}_t > 0.8$). Key insights:

- The largest monthly jump occurs in March ($\bar{z}_3 = 0.853$).
- Comparing this volume-based index with the price-based version shows whether real consumption intensity induces larger swings than price shifts alone.
- To capture extreme fluctuations more faithfully, one can test nonlinear forms $F(w,z)$ (logarithmic, exponential).
- Calibrating a to match the empirical variance of w_t will align model sensitivity with observed consumer behavior.

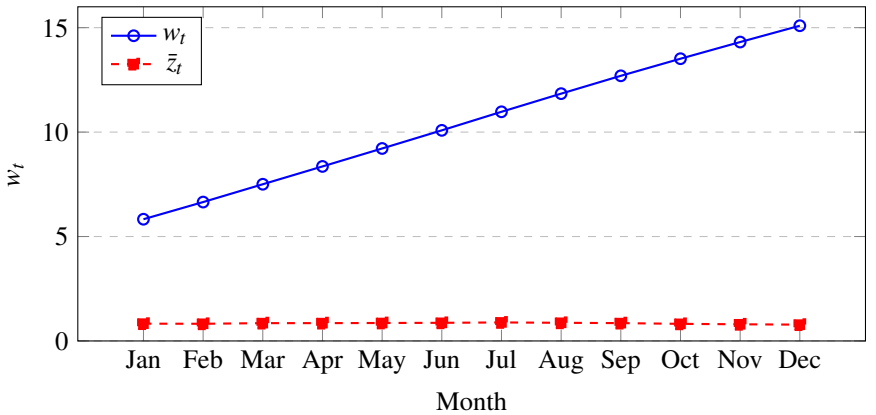

Figure 1. Dynamics of the preference index w_t and average demand–supply ratio $\bar{z}_t$ in 2023.

4. The impact of the budget constraint on consumer preferences

Focusing solely on the budget constraint isolates the pure effect of prices and income on consumer choices. Consider two goods, apples x_1 and oranges x_2, with prices p_1, p_2 and income M. The consumer solves

$$\max_{x_1,x_2} U(x_1,x_2) \quad \text{s.t.} \quad p_1x_1 + p_2x_2 \le M. \tag{30}$$

Example 4.1. Let $p_1 = 2$, $p_2 = 3$, $M = 30$ and

$$U(x_1,x_2) = x_1^{0.5}x_2^{0.5}. \tag{31}$$

Form the Lagrangian

$$\mathscr{L}(x_1,x_2,\lambda) = x_1^{0.5}x_2^{0.5} - \lambda\,(2x_1 + 3x_2 - 30). \tag{32}$$

First-order conditions:

$$\begin{aligned} \mathscr{L}_{x_1}: &\quad 0.5\,x_1^{-0.5}x_2^{0.5} - 2\lambda = 0, \\ \mathscr{L}_{x_2}: &\quad 0.5\,x_1^{0.5}x_2^{-0.5} - 3\lambda = 0, \\ \mathscr{L}_{\lambda}: &\quad 2x_1 + 3x_2 - 30 = 0. \end{aligned} \tag{33}$$

Divide the first eq. by the second:

$$\frac{x_2}{x_1} = \frac{2}{3} \implies x_2 = \tfrac{2}{3}x_1. \tag{34}$$

Substitute into the budget line:

$$2x_1 + 3\left(\tfrac{2}{3}x_1\right) = 30 \implies 4x_1 = 30 \implies x_1 = 7.5, \quad x_2 = 5. \tag{35}$$

Hence the optimum under the budget constraint is $(x_1, x_2) = (7.5, 5)$.

Remark 4.2. Example 4.1 shows that at the optimal bundle, the marginal rate of substitution equals the price ratio:

$$\frac{MU_{x_1}}{MU_{x_2}} = \frac{0.5\,x_1^{-0.5}x_2^{0.5}}{0.5\,x_1^{0.5}x_2^{-0.5}} = \frac{x_2}{x_1} = \frac{p_1}{p_2}. \tag{36}$$

Substituting back into the budget equation pins down the exact quantities. Thus, prices and income jointly determine the consumer's preferred mix of goods.

5. Dynamics of preferences under supply and demand in discrete time

We study how the ratio of demand to supply affects the evolution of a consumer's preference index w_t over discrete periods $t = 0, 1, \ldots, T$. Let

$$z_t = \frac{D_t}{S_t} \tag{37}$$

denote the demand-to-supply ratio in period t, and let $a > 0$ be a sensitivity parameter. We assume a linear update rule:

$$w_{t+1} = F(w_t, z_t) = w_t + a\,z_t, \tag{38}$$

with initial preference w_0.

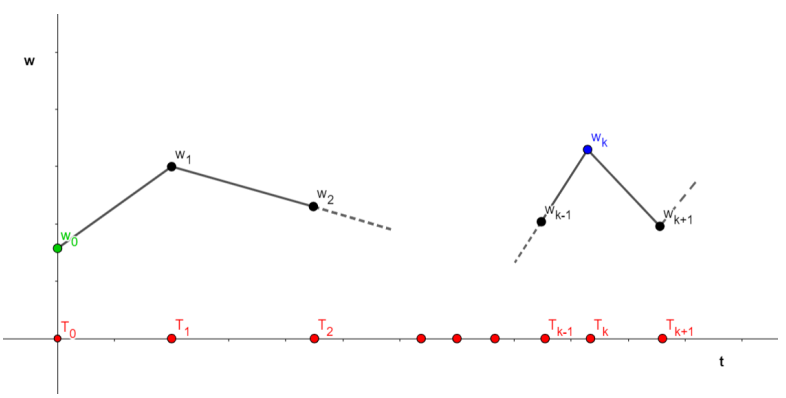

Figure 2. Evolution of preferences w_t under varying demand–supply ratios z_t.

From the recurrence we get the closed-form expression:

$$w_t = w_0 + a\sum_{i=0}^{t-1} z_i. \tag{39}$$

Hence:

1. The preference returns to its initial level w_0 after T periods if and only if

$$w_T = w_0 \quad \Longleftrightarrow \quad \sum_{i=0}^{T-1} z_i = 0. \tag{40}$$

2. We define a *critical threshold* w_k (e.g. the point of complete preference reversal or resource exhaustion). Two regimes arise:

$$\begin{cases} w_t > w_k, & \text{(safe region)}, \\ w_t \leq w_k, & \text{(critical region)}. \end{cases} \tag{41}$$

To link this to a budget-type constraint (cf. Sec. 3), suppose the total resources available to influence preferences cannot exceed M. Interpreting $a z_t$ as the resource expenditure in period t, the critical threshold is reached when

$$\sum_{i=0}^{k-1} a z_i = M. \tag{42}$$

Solving for k gives the first period at which resources are fully utilized. At that moment,

$$w_k = w_0 + M, \tag{43}$$

and any further demand–supply shock z_t would push the system beyond the consumer's capacity.

Table 2. Values of $P = F(w,z)$ for $w = 5$, $a = 1$ and selected z

Model	Formula $F(w,z)$	$z = 0.2$	$z = 0.5$	$z = 1.0$	$z = 2.0$
Linear	$5+z$	5.20	5.50	6.00	7.00
Logarithmic	$5+\ln(1+z)$	5.18	5.41	5.69	6.10
Quadratic	$5+z^2$	5.04	5.25	6.00	9.00
Exponential	$5+e^z$	5.22	5.65	7.72	12.39

Remark 5.1. In this linear framework the cumulative effect of supply–demand imbalances is transparent. One sees immediately how alternating positive and negative ratios can cancel out, returning w_t to baseline, or else drive it past a critical threshold if the aggregate sum exceeds M.

6. Indifference curves and consumer preferences

Indifference curves are a cornerstone of consumer choice theory. For a given utility level U, the indifference curve

$$I_U \;=\; \{(x_1,x_2)\colon U(x_1,x_2) = U\} \tag{44}$$

collects all bundles of goods that yield the same satisfaction.

Key properties of indifference curves:

- They slope downward: to keep utility constant, an increase in x_1 must be offset by a decrease in x_2.
- They are typically convex to the origin, reflecting a diminishing marginal rate of substitution (MRS).

The marginal rate of substitution at any point measures the rate at which the consumer is willing to exchange good 2 for good 1 while remaining on the same curve:

$$\text{MRS} = \frac{MU_{x_1}}{MU_{x_2}} = -\frac{dx_2}{dx_1}\Big|_{U=\text{const}}, \tag{45}$$

where $MU_{x_i} = \frac{\partial U}{\partial x_i}$.

When we overlay the budget line

$$p_1 x_1 + p_2 x_2 = M, \tag{46}$$

the optimal consumption bundle is found at the tangency point satisfying

$$\text{MRS} = \frac{p_1}{p_2}. \tag{47}$$

Illustration. Let I_{U_1} and I_{U_2} be two indifference curves with $U_2 > U_1$. If the higher curve I_{U_2} lies outside the budget set, the consumer's utility-maximizing choice is the tangency point on I_{U_1}. Otherwise, they reach I_{U_2} and attain greater utility.

Indifference curves themselves do not change underlying preferences; rather, they provide a graphical tool to analyze how budget constraints, price changes, and income shifts guide consumer choices and substitution patterns.

7. Changes in income, prices, and consumer choices

When a consumer's income or the prices of goods change, the budget set and the optimal consumption bundle both shift. We distinguish two main effects:

1. Income Changes
   - Income increase: the budget line $p_1 x_1 + p_2 x_2 = M$ shifts outward.
   - Income decrease: the same line shifts inward.
   - The new optimum traces out an Engel curve showing how x_i varies with M.
2. Price Changes
   - Substitution effect: consumer moves along the original indifference curve to a tangency with a hypothetical "compensated" budget line.

– Income effect: the compensation restores purchasing power, yielding the final bundle.

– Slutsky decomposition separates these two responses.

7.1. Marshallian demand for Cobb–Douglas preferences

For

$$U(x_1,x_2) = x_1^{\alpha} x_2^{1-\alpha}, \tag{48}$$

the Marshallian (uncompensated) demands are

$$x_1(M,p) = \frac{\alpha M}{p_1}, \quad x_2(M,p) = \frac{(1-\alpha)M}{p_2}. \tag{49}$$

Example 7.1 (Income Shock). Let $\alpha = 0.5$, $p_1 = 2$, $p_2 = 3$, and initial income $M = 30$. Then

$$x_1 = \frac{0.5 \cdot 30}{2} = 7.5, \qquad x_2 = \frac{0.5 \cdot 30}{3} = 5. \tag{50}$$

If M rises to 45, the new demands become

$$x_1 = \frac{0.5 \cdot 45}{2} = 11.25, \qquad x_2 = \frac{0.5 \cdot 45}{3} = 7.5. \tag{51}$$

Thus, higher income yields strictly larger consumption of both goods.

Example 7.2 (Price Shock). With the original income $M = 30$ and $\alpha = 0.5$, let p_1 increase from 2 to 3 while $p_2 = 3$. Then

$$x_1 = \frac{0.5 \cdot 30}{3} = 5, \qquad x_2 = \frac{0.5 \cdot 30}{3} = 5. \tag{52}$$

The rise in p_1 reduces apples from 7.5 to 5 units, and the consumer reallocates expenditure toward oranges.

7.2. Graphical interpretation

- An outward shift of the budget line (via higher M) lets the consumer reach a higher indifference curve.
- A pivot of the budget line (via a price change) causes a rotation around the intercept on the axis of the non-stochastically priced good.
- The tangency condition

$$\frac{MU_1}{MU_2} = \frac{p_1}{p_2} \tag{53}$$

still determines the optimal bundle after any shift.

In summary, income variations slide the chosen point along Engel curves, while price changes combine substitution along one indifference curve with an income effect that shifts to another. Both mechanisms alter the mix of goods that maximizes consumer utility under the new budget.

8. Conclusions

Building on our earlier studies by Fidler and Matysiak (2024, 2025), this paper has broadened the mathematical modeling of consumer preferences by explicitly integrating demand–supply dynamics, budget constraints, and income variations into a single framework. Our main contributions are:

- **Unified quantitative–qualitative modeling.** We contrasted multiple functional forms (linear, logarithmic, quadratic, exponential, Cobb–Douglas) for the preference–shock mapping $F(w,z)$, capturing both smooth and extreme market responses.
- **Dynamic stability analysis.** Introducing a critical threshold and fixed-point condition for the preference index w_t under discrete demand–supply shocks reveals when and how aggregate imbalances return to—or deviate from—baseline levels.
- **Endogenous feedback between preferences and demand.** By allowing shifts in preferences to feed back into demand (and hence prices), we moved beyond the one-way causality of standard models and showed how sentiment shifts can amplify market movements.
- **Rigorous incorporation of revealed-preference tools.** Embedding Afriat – Varian – Mas – Colell constructions and classical indifference-curve analysis within our dynamic setting provides a tighter link between theoretical consistency and empirical price–income observations.

Taken together, these advances deepen our understanding of how real-world shocks - whether from price spikes, income changes, or supply disruptions—propagate through individual utility maximization and aggregate market behavior. They also offer a versatile toolkit for policymakers and marketers to simulate consumer responses under varying risk and uncertainty conditions.

Looking ahead, we plan to enrich the model by:

1. Introducing heterogeneous consumer types and segment-specific functional forms.
2. Endogenizing income processes (e.g. stochastic wages, credit constraints).
3. Incorporating social network effects and habit formation into the preference-update rule.
4. Validating the framework on high-frequency transaction data to calibrate sensitivity parameters a and critical thresholds.

These extensions will help bridge the gap between theoretical preference representations and observed purchasing patterns in dynamic, uncertain markets.

References

Afriat, S., (1967). The construction of utility functions from expenditure data. *International Economic Review*, 8(1), pp. 67–77.

Asano, A., (2013). *An Introduction to Mathematics for Economics*. Cambridge: Cambridge University Press.

Chambers, C., Echenique, F., (2016). *Revealed Preference Theory*. Cambridge: Cambridge University Press.

Dryzek, J. S., (2014). Revealed preference and the pluralism of rationality. *Economics and Philosophy*, 30(3), pp. 383–403.

Fidler, J., Matysiak, Ł., (2024). A Consumer Preferences. An Influence of Society on the Individual. *Preprint*, 2024090998. Available at:https://doi.org/10.20944/preprints 202409.0998.v1

Fidler, J., Matysiak, Ł., (2024). Risk, uncertainty, psychological factors and individual preferences. *SSRN Working Paper*. Available at:http://dx.doi.org/10.2139/ ssrn.4983662

Fidler, J., Matysiak, Ł., (2025). The impact of selected psychological factors on consumer preferences. *SSRN Working Paper*. Available at: http://dx.doi.org/10.2139/ssrn. 5084459

Manzini, P., Mariotti, M., (2014). Revealed preference and behavioral economics. In: Zamir, E., and Teichman, D., (eds.), *The Oxford Handbook of Behavioral Economics and the Law*. Oxford: Oxford University Press, pp. 3–21.

Mas-Colell, A., (1978). On revealed preference analysis. *The Review of Economic Studies*, 45(1), pp. 121–131.

Samuelson, P. A., (1938). A note on the pure theory of consumers' behaviour. *Economica*, 5(17), pp. 61–71.

Varian, H., (1982). Nonparametric analysis of optimizing behavior. *Econometrica*, 50(4), pp. 945–973.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 119–135, https://doi.org/10.59139/stattrans-2026-007
Received – 08.12.2024; accepted – 22.09.2025

The comparison of the hidden Markov model with machine learning techniques in agricultural prediction

Muraleedharan Vyshnavi[1], **Madaswamy Muthukumar**[2]

Abstract

This study compares hidden Markov models (HMMs) with various machine learning approaches to assess their effectiveness in forecasting agricultural data based on Python. Accurate forecasts are essential to promote sustainability and increase agricultural productivity. Through the use of an extensive dataset of agricultural parameters, specifically the cultivated area of oilseeds, the study explores historical trends and correlations. Model performance is evaluated using the Mean Absolute Error (MAE) along with the R-squared, and residual analysis is used to analyse how well the models represent the underlying trends. The findings demonstrate that HMMs are able to predict agricultural trends with higher accuracy than their other counterparts, thereby providing useful information for improved agricultural planning and decision-making. Future studies should concentrate on improving forecast accuracy and resolving any issues associated with agricultural data prediction.

Key words: machine learning, mean absolute error, R-squared, hidden Markov model, residual analysis.

1. Introduction

Technological developments in data science and machine learning (ML) have revolutionized the agriculture industry in the last few decades. With these advanced tools, farmers, agronomists, and policymakers can now maximize productivity, control risks, and improve decision-making. Due to their capacity to manage the intricate temporal and spatial interactions present in agricultural systems, HMM and other ML techniques have become essential tools for assessing agricultural data among the variety of methodologies available. The agricultural sector faces distinct difficulties that are defined by unpredictability and are impacted by market forces, environmental variables, and technological advancements. While traditional statistical techniques have yielded valuable

[1] Department of Statistics, PSG College of Arts & Science, Coimbatore-641014, Tamil Nadu, India. E-mail: vyshnavimp@gmail.com. ORCID: https://orcid.org/0009-0000-1098-8161.
[2] Department of Statistics, PSG College of Arts & Science, Coimbatore-641014, Tamil Nadu, India. E-mail: muthukumar@psgcas.ac.in.

insights, they are still limited in their ability to capture complex patterns and interconnections seen in dynamic agricultural systems. This constraint has prompted the use of new computational techniques that can better describe uncertainties, nonlinear relationships, and stochastic processes.

A strong foundation for simulating temporal sequences in which underlying states are deduced from observable data is provided by HMMs. HMMs, which were first created in the fields of speech recognition and computational linguistics, have found extensive use in agriculture for tasks like predicting crop yield, identifying disease outbreaks, and evaluating the effects of climate change. Because HMMs are probabilistic, state transitions and emissions can be flexibly modeled to account for the inherent complexities and uncertainties seen in agricultural datasets. Techniques for ML include a broad range of algorithms that can recognize patterns in data and forecast outcomes without the need for explicit programming. In agriculture, ML models such as SVMs, NNs, LR, and KNN, have been extensively applied across various domains including crop classification, soil fertility prediction, yield estimation, and pest management. These techniques leverage large datasets to identify patterns and relationships, offering valuable insights for optimizing agricultural practices and resource allocation.

Poonam Somani, Shreyas Talele, and Suraj Sawant (2014) investigated the application of SVMs, NNs, and HMM to stock market forecasting. It emphasizes how crucial these methods are for reporting on changes in stock prices and suggests an HMM to increase precision. Choukri Djellali and Mehdi Adda (2020) presented the RHMM deep learning model, which leverages ANN and the HMM for recommender systems. Through experimentation, the model outperforms testing models by optimizing the bias-variance conflicts and enhancing training stability and accuracy. Pasak Senawongse, Andrew R. Dalby, and Zheng Rong Yang (2005) predicted phosphorylation sites in amino acid sequences, the study used HMMs which outperform decision trees and neural network methods and produce predictions with greater certainty than HMMs. Lefteris Benos, et al. (2008) examined the potential of machine learning in agriculture, emphasizing managing crops, water, soil, and livestock. It emphasizes the effectiveness of algorithms using Artificial Neural Networks, with the most researched crops being sheep, cattle, wheat, and maize. To summarize, this study aims to extend our understanding of how machine learning approaches, such as HMMs, might improve the analysis and prediction of long-term oilseed area patterns. This research aims to give practical insights to stakeholders in the agricultural industry by assessing the strengths and limitations of various models, ultimately contributing to more informed decision-making.

2. Methodology

2.1. Hidden Markov Model

A Hidden Markov Model is a statistical model describing systems with unobserved (hidden) states. It works particularly well with time series data, where the system is considered to follow a Markov process with unknown parameters. It is utilized in this study to model and forecast the actual annual oilseed area values. The model begins by defining hidden states S = $\{S_1, S_2, \ldots, S_N\}$ that represent various underlying conditions impacting oilseed areas, such as different weather patterns or economic factors. These hidden states correspond to observable data O = $\{O_1, O_2, \ldots, O_T\}$, which in this case are the measurements of the oilseed area. The HMM (Leite et al. 2008) parameters are the following,

- Transition Probabilities(A): A matrix A = $[a_{ij}]$ where a_{ij}= P $(q_{t+1} = S_j / q_t = S_i)$ indicates the probability of transitioning from state $S_i to\ S_j$.
- Emission Probabilities (B): A matrix B = $[b_j(k)]$ where b_j(k) = P $(o_t = O_k / q_t = S_j)$ specifies the probability of observing an oilseed area value O_k given that the system is in state S_j.
- Initial State Distribution (π): A vector π = $[\pi_i]$ where $\pi_i = P(q_1 = S_i)$ denotes the probability of starting in each hidden state.

The model is trained using the Baum-Welch algorithm, which is an expectation-maximization method that iteratively estimates these parameters based on the observed oilseed area data. This algorithm enhances the model's comprehension of state transitions and emissions to maximize the likelihood P(O/ʎ) of the observed data, where ʎ represents the model parameters A, B, and π. After, training HMM can predict the actual oilseed area value for the years 1992 to 2022 by identifying the sequence of hidden states $q_{1:T}$ = arg $max_{q1:T}$ P($q_{1:T}$ / $O_{1:T}$, ʎ) that best explains the observed data. This method utilizes the most probable sequence of hidden states and their corresponding emissions, providing insights into historical trends and fluctuations in oilseed areas.

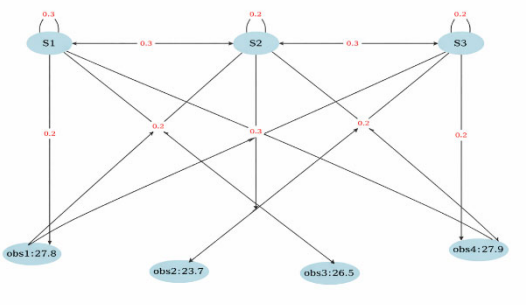

Figure 1. Structural diagram of HMM

2.2. Linear regression

A modeling approach called linear regression (Tranmer and Elliot 2008) is used to analyze data and generate predictions. To predict a response variable (Y) from an explanatory variable (X), a bivariate model is constructed using simple linear regression. The scatterplot's points are fitted with a straight line using linear regression, and predictions are made using the line's equation. The equation for the line in regression modeling is as follows:

$$Y = \beta_0 + \beta_1 X + e_i \tag{1}$$

where the slope of the regression line is the regression parameter β_1 and the intercept is the regression parameter β_0. With a mean of zero and a constant variance, it is assumed that the random error term e_i is uncorrelated. The assumption that the mistakes are distributed regularly is frequently added to analyses to make inference easier and enhance estimation efficiency. The data may be transformed to approach normality (Zou et al. 2003).

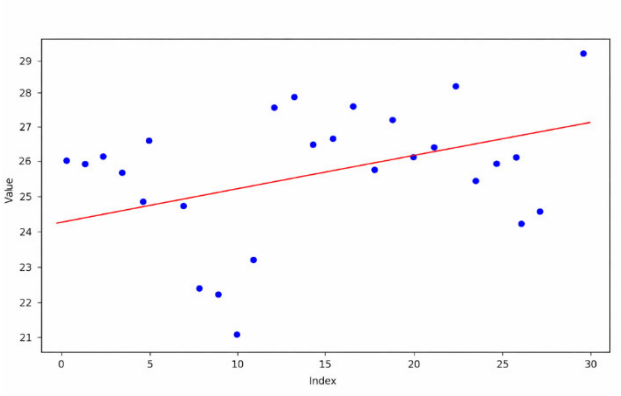

Figure 2. Visualization of linear regression model

2.3. Support vector machines

SVMs (Meyer and Wien 2001) for binary classification were created by Cortes and Vapnik (1995). In machine learning, support vector machines are an effective technique that may be applied to both regression and classification problems. The first step in using SVMs for prediction is gathering and preparing the dataset to ensure it is appropriately scaled for analysis. The data type is then considered while choosing an appropriate kernel function, such as radial basis function (RBF), polynomial, or linear. Additionally, the regularization parameter (C) is initialized. It regulates the trade-off between maximizing the margin and decreasing classification errors. The data is then transformed into a higher-dimensional space where it is simpler to separate using a hyperplane by computing the kernel matrix using the chosen kernel function. When

input data points x_i and x_j are used, the kernel matrix element K_{ij} is calculated as follows (Jakkula 2006):

$$K_{ij} = K(x_i, x_j) \tag{2}$$

The next step is to formulate a quadratic optimization problem, whose goal is to identify the best hyperplane that maximizes the margin between classes while adhering to the restrictions. For the quadratic optimization problem, the objective function (Jakkula 2006) is:

$$\min_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{3}$$

Based on the limitations;

- $\sum_{i=1}^{n} \alpha_i y_i = 0$
- $0 \leq \alpha_i \leq C$, for all i.

The solution function for regression problems is:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \, K(x_i, x_j) + b \tag{4}$$

SVMs are certain to generate reliable and accurate predictions on fresh data through this systematic technique.

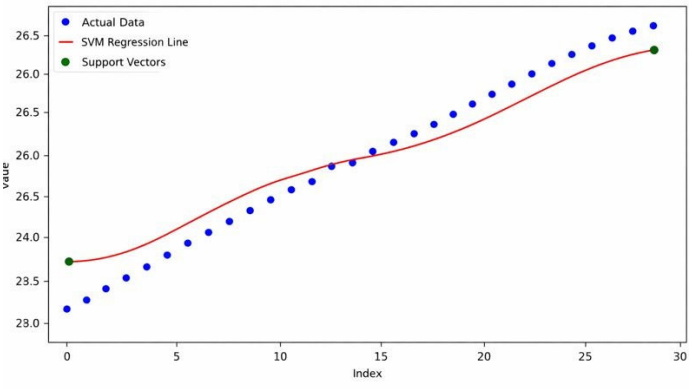

Figure 3. Support vector machine structure

2.4. Neural networks

The neural network (Mark et al. 1992) was designed with three layers: input, hidden, and output, as illustrated in the following figure. Layers consist of many nodes and are connected by correlation weights. Nodes accept input from either outside the model or interconnections. Nodes process the input to produce an analogue output, determining the firing rate. The weight function multiplies the incoming firing rate

before it arrives at the next layer. Each node's transformation is a sigmoid function as specified below:

$$f(x) = \frac{1}{1+ exp(-a(x-b))} \tag{5}$$

When the node's input is denoted by x, its output, or firing rate, is represented by f(x), the gain is 'a', and the bias is 'b'. The values of a and b are calculated using in the below equation:

$$(H_{input}) = f\left(\sum_{j-1}^{n} w_{ij}x_j + b_i\right) \tag{6}$$

where w_{ij} are the weights connecting the input x_j.

$$(O_{output}) = \sum_{k=1}^{m} w_k h_k + b_o \tag{7}$$

In this neural network model, raw data is processed by the input layer, non-linear changes are applied by the hidden layer using the sigmoid function, and predictions are produced by the output layer. The network gains strong predictive capabilities through training, which teaches it to modify its weights and biases to minimize the error between real and anticipated values.

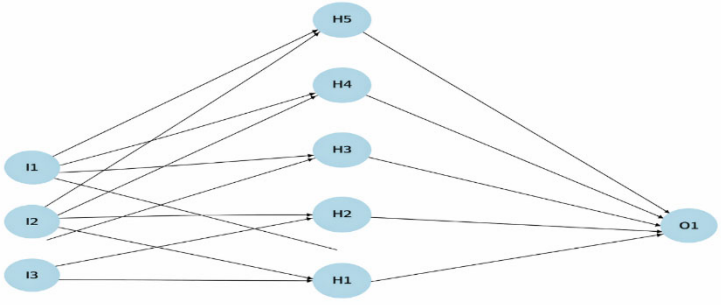

Figure 4. Structural overview of neural networks: input, hidden, and output layers

2.5. K-Nearest Neighbors

The K-Nearest Neighbors (Sadegh and Bolandraftar 2013) technique predicts the outcome for a query point X based on the average of its K nearest neighbors. To improve this, one can weigh the contributions of these neighbors so that closer neighbors have a bigger impact on the forecast than those further away. Regression problems involve predicting the value of a dependent variable (y) using independent variables (x). Consider a scheme where a series of points reflects the relationship between the independent variable and the dependent variable. KNN uses a set of known examples to predict the outcome of a query point (X). The k-nearest neighbor approach predicts the value of the nearest neighbor. If x_4 is the nearest neighbor to X, then:

$$\bar{y}_X = y_4 \tag{8}$$

In the 2-nearest neighbor approach, we average the results of the two nearest neighbors x_3 *and* x_4.

$$\bar{y}_X = \frac{y_3 + y_4}{2} \tag{9}$$

This strategy can be applied to K nearest neighbors. The predicted value is obtained as the average of the outcomes of the K nearest neighbors.

$$\bar{y}_X = \frac{1}{k}\sum_{i=1}^{k} y_i \tag{10}$$

where y_i denotes the outcome of the i^{th} nearest neighbor. The distance between the query points can be measured using the Euclidean distance, Squared Euclidean distance, Manhattan distance, and Chebyshev distance.

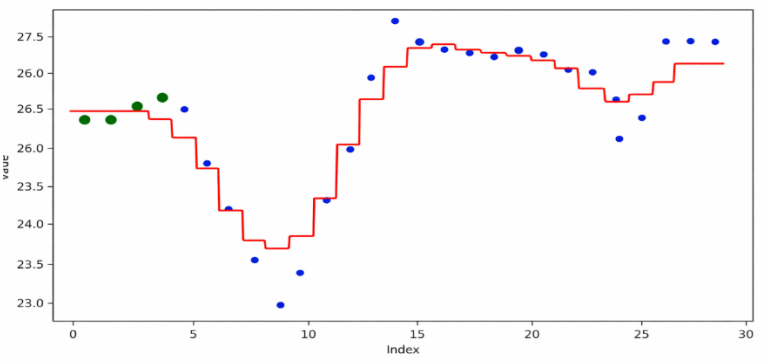

Figure 5. Schematic diagram of K-Nearest Neighbors Model

2.6. Residual analysis

Residual analysis (Rebekka and Gomez 2004) is a valuable set of tools for evaluating the goodness of a fitted model. The purpose of residual analysis is to verify if statistical models applied to data satisfy the underlying assumptions of the models. These presumptions include independent and identically distributed model errors, as well as a regression function that is appropriately defined. Additionally, since most regression estimators only maintain consistency under the assumption that the error distribution is true, residual analysis is particularly crucial. The usage of residual plots, or plots of residuals against explanatory variables or the corresponding fitted values, has become commonplace in the detection of regression model deficiencies evaluation. Residuals are the variations between observed values (y_i) and the predicted values ($\hat{y}_i$) from the models. The mathematical representation is as follows:

$$\text{Residual} = y_i - \hat{y}_i \tag{11}$$

To evaluate how well the model's predictions match the actual data, residuals are utilized. The residuals, when they are randomly distributed around zero, show that there are no systematic errors in the model's representation of the underlying structure of the data. Several methods are used in residual analysis to validate a model's key assumptions, which include independence, normality and homoscedasticity.

A statistical test called the Durbin-Watson (DW) test is used to determine whether the residuals from a regression analysis contain autocorrelation, which is a link between values that are separated from one another by a certain amount of time. The DW (Walter 2011) is calculated as:

$$\mathrm{DW} = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{12}$$

The Shapiro-Wilk (W) test is a statistical tool for evaluating the normality of residuals in regression analysis. Residuals must be normally distributed because this validates the use of statistical tests and confidence intervals in the model. By running this test, we can find potential problems with the assumptions in the model and fix them to increase the model's accuracy and reliability. The test statistic (Zofia et al. 2016) is calculated as:

$$\mathrm{W} = \frac{(\sum_{i=1}^{n} a_i e_i)^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2} \tag{13}$$

The Breusch-Pagan (BP) test is a statistical method for detecting homoscedasticity in regression model residuals. Homoscedasticity happens when the variance of the residuals fluctuates between observations, which contradicts the regression analysis assumption that there is constant variance. The Breusch-Pagan test statistic is given by:

$$\mathrm{BP} = \mathrm{n}\ \mathrm{R}^2 \tag{14}$$

where n is the number of values and R^2 is the coefficient of determination from the auxiliary regression of e_i^2 on the independent variables. The computed test statistic BP follows a chi-squared distribution with k degrees of freedom, where k represents the number of independent variables in the supplementary regression.

3. Results and discussions

The dataset used in this study comprises annual historical records of oilseed cultivation area in India, covering the period from 1992 to 2022. Each data point represents the total cultivated area measured in million hectares for the respective agricultural year. This data was sourced from the Economics & Statistics Division of the

Ministry of Agriculture & Farmers Welfare and is publicly accessible via their official website: http://desagri.gov.in.

Table 1. Effectiveness of various models in predicting values in given- years

Year	Area-Actual Value	Predicted Value				
		HMM	KNN	LR	NN	SVMs
1992-93	24.26	25.85	25.764	24.431	24.836	25.940
1993-94	26.20	25.85	25.764	24.538	24.915	25.965
1994-95	26.09	25.85	25.764	24.648	24.993	25.990
1995-96	26.31	25.85	25.972	24.762	25.072	26.014
1996-97	25.96	25.85	26.112	24.877	25.151	26.039
1997-98	25.30	25.85	25.942	24.986	25.229	26.064
1998-99	26.90	25.85	25.234	25.087	25.308	26.088
1999-00	25.24	25.85	24.570	25.189	25.386	26.113
2000-01	22.77	22.93	23.808	25.290	25.465	26.138
2001-02	22.64	23.70	23.160	25.391	25.544	26.162
2002-03	21.49	22.93	23.616	25.493	25.622	26.187
2003-04	23.66	23.70	24.634	25.594	25.701	26.212
2004-05	27.52	27.30	25.408	25.696	25.779	26.237
2005-06	27.86	27.30	26.448	25.798	25.858	26.261
2006-07	26.51	27.30	27.228	25.896	25.937	26.286
2007-08	26.69	27.30	26.916	25.994	26.015	26.311
2008-09	27.56	27.30	26.788	26.092	26.094	26.335
2009-10	25.96	27.30	26.748	26.190	26.172	26.360
2010-11	27.22	27.30	26.706	26.288	26.251	26.385
2011-12	26.31	27.30	26.804	26.386	26.329	26.410
2012-13	26.48	27.30	26.732	26.483	26.408	26.434
2013-14	28.05	27.30	26.506	26.581	26.487	26.459
2014-15	25.60	25.85	26.480	26.679	26.565	26.484
2015-16	26.09	25.85	26.086	26.777	26.644	26.508
2016-17	26.18	25.85	25.434	26.875	26.722	26.533
2017-18	24.51	22.93	25.742	26.973	26.801	26.558
2018-19	24.79	23.70	26.290	27.084	26.880	26.582
2019-20	27.14	27.30	26.888	27.196	26.958	26.607
2020-21	28.83	27.30	26.888	27.307	27.037	26.632
2021-22	29.17	27.30	26.888	27.419	27.115	26.657

Model evaluation

R-squared measures how well a model explains the variability in the observed data, making it sensitive to how accurately the model captures overall trends and fluctuations. Even small changes in the pattern of prediction errors can significantly impact its value. A higher R-squared indicates a stronger fit, meaning the model accounts for more of the

data's variability, while a low or negative R-squared suggests a poor fit. In contrast, Mean Absolute Error (MAE) represents the average size of prediction errors regardless of their direction, offering a clear measure of accuracy without reflecting how well the model explains the variance in the data. Together, a high R-squared and low MAE provide complementary insights into a model's explanatory power and predictive accuracy.

Table 2. Performance comparison of predictive models

Models	MAE	R^2- Valued
Hidden Markov Model	0.7041	0.7449
K- Nearest Neighbors	0.9353	0.5760
Linear Regression	1.307	0.14
Neural Networks	1.2783	0.1529
Support Vector Machines	1.246	0.0466

With the greatest R-squared of 0.7449 and the lowest MAE of 0.7041 when compared to other models, Hidden Markov Models proved to have greater prediction accuracy. These results highlight how well they predict agricultural oilseed areas.

Table 3. The Durbin-Watson test results for the independence of residuals

Model	Durbin-Watson Statistic	Autocorrelation Interpretation
Residual-HMM	1.324	Mild positive autocorrelation
Residual-KNN	1.536	Mild positive autocorrelation
Residual-NN	0.817	Significant positive autocorrelation
Residual-LR	0.832	Significant positive autocorrelation
Residual-SVMs	0.744	Significant positive autocorrelation

These results show variable degrees of positive autocorrelation in the residuals among models. A Durbin-Watson statistic closer to 2 suggests less autocorrelation, whereas values that deviate significantly from 2 imply more autocorrelation.

Table 4. The normality test results for residual errors

Model	Statistic	p-value	Conclusion
Residual-HMM	0.978	0.767	Fail to reject H_0
Residual-KNN	0.973	0.621	Fail to reject H_0
Residual-NN	0.908	0.013	Reject H_0
Residual-LR	0.921	0.028	Reject H_0
Residual-SVMs	0.953	0.199	Fail to reject H_0

Shapiro-Wilk tests show that residuals from HMM, KNN, and SVMs follow a normal distribution ($p > 0.05$), however, residuals from NN and LR do not ($p < 0.05$).

Table 5. The results of the Breusch-Pagan test for homoscedasticity

Model	Breusch-Pagan Test	Homoscedasticity (Interpretation)
Residual-HMM	0.128	Likely homoscedastic
Residual-KNN	0.231	Likely homoscedastic
Residual-NN	0.615	Likely homoscedastic
Residual-LR	0.801	Likely homoscedastic
Residual-SVMs	0.790	Likely homoscedastic

This assumption indicates that the residuals from each model have a constant variance. The p-values (all greater than 0.05) indicate that there is no meaningful evidence to challenge this assumption for any of the models. As a result, each model's residuals are likely to be homoscedastic, which supports their dependability in producing consistent predictions.

3.1. Model fitting and sequence estimation

This study uses the hidden state space S = {Low, Medium, High} and observable state space O = {Decrease, Neutral, Increase}. The hidden states reflect categorized levels of the target agricultural variable (area). Specifically, the Low state corresponds to area values ranging from 21 to 23.83, indicating the lower end of the distribution. The Medium state includes values between 23.83 and 26.66, reflecting the mid-range or average levels. The High state covers values from 26.66 to 29.5, representing the upper portion of the dataset. The observable states describe the direction of year-to-year changes: Decrease indicates a drop in value, Neutral reflects minimal or no change, and Increase denotes a rise in value. These classifications support the model in linking hidden state transitions with observable patterns, enhancing both estimation and interpretability.

The model was trained after determining that HMM was the most suitable based on performance metrics. During training, the following parameters were estimated:

1. Transition Probability Matrix (A): It defines the likelihood of moving from one hidden state to another over time. In this study, the hidden states are categorized as Low, Medium, and High, representing different levels of agricultural area. Each element in the matrix indicates the probability of transitioning from a current state to a future state. It is calculated by analyzing how often the model transitions from one hidden state to another in a given sequence. Each value in the matrix represents the probability of moving from one state (e.g., Low) to another (e.g., Medium or High). To compute it, the number of transitions between states is counted and then normalized by dividing each row by the total transitions from that state. This provides a probability-based understanding of how the system evolves over time.

$$A = \begin{array}{c|ccc} & \text{Low} & \text{Medium} & \text{High} \\ \text{Low} & 0.75 & 0 & 0.25 \\ \text{Medium} & 0.062 & 0.625 & 0.312 \\ \text{High} & 0 & 0.55 & 0.44 \end{array}$$

2. Emission Probability Matrix (B): It shows the likelihood of observing a particular output given a specific hidden state. It connects the hidden states (e.g., Low, Medium, High) to the observable outcomes (e.g., Decrease, Neutral, Increase). To calculate it, the number of times each observation occurs while the system is in a specific hidden state is counted. These counts are then normalized by dividing each by the total number of observations for that hidden state, resulting in probabilities that reflect how likely each observation is under each state.

$$B = \begin{array}{r} \\ \text{Low} \\ \text{Medium} \\ \text{High} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Decrease} & \text{Neutral} & \text{Increase} \end{array} \\ \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.30 & 0.70 \end{bmatrix} \end{array}$$

3. Initial Probability Matrix (π): It represents the probability distribution over hidden states at the starting point (time t = 0). It indicates how likely the system is to begin in each state (e.g., Low, Medium, or High). To calculate it, the number of times each state appears as the first state in multiple sequences is counted and then divided by the total number of sequences.

$$\pi = [0.133, 0.533, 0.333]$$

Using the initial state distribution, transition matrix, and emission matrix, a 10-year forecast can be made by repeatedly updating the state probabilities through the transition matrix. Expected values are then calculated by combining these state probabilities with the emission matrix. This method captures the system's time-dependent behavior and can be efficiently performed using MATLAB's built-in HMM tools.

3.2. Graphical representation

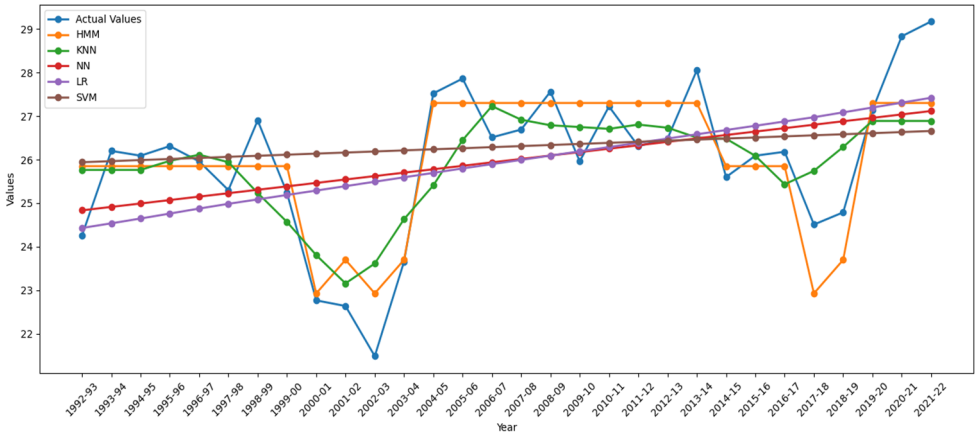

Figure 6. The actual and predicted values of different models

The graph presents a comparison between actual oilseed area data and the predictions made by five models over the period from 1992 to 2022. The actual data shows significant variability, with a noticeable decline during the early 2000s. Among all models, the HMM closely tracks the real data, effectively reflecting both declines and recoveries. KNN and NN provide reasonably accurate predictions but exhibit some inconsistencies. In contrast, SVM and Linear Regression tend to produce smoother or more linear trends, overlooking key changes in the actual data. Overall, HMM provides the most reliable fit to the values.

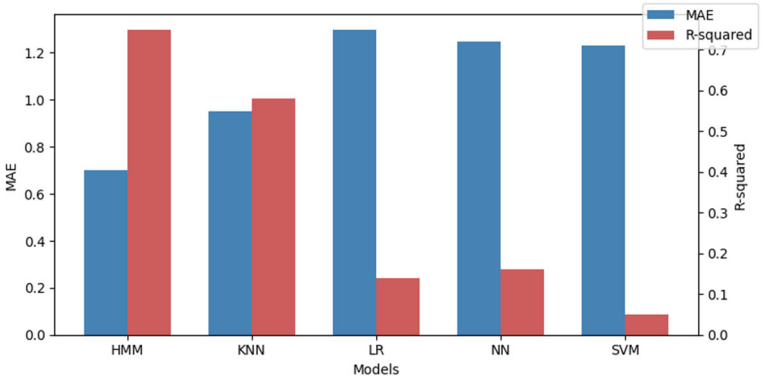

Figure 7. Comparison of MAE and R-squared Value

The bar chart presents a comparison of predictive performance based on MAE and R-squared values. One model achieves the lowest error and highest fit, indicating strong accuracy and reliability. Others show moderate to poor performance, either with higher errors or weaker data alignment. Overall, the best-performing model stands out in both accuracy and consistency.

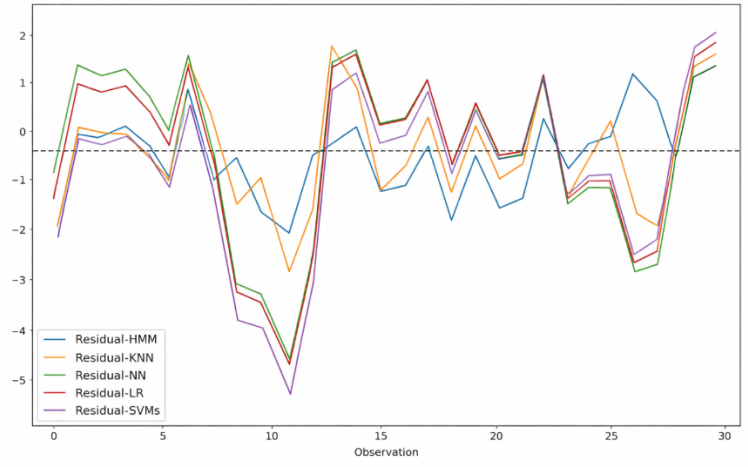

Figure 8. The residuals for different models across the observations

The line graph illustrates residual variations across different prediction models. Residuals close to zero indicate higher prediction accuracy. The HMM model shows relatively smaller and more stable residuals, especially after the 10th observation, indicating better consistency. In contrast, other models exhibit larger fluctuations, particularly around the 10th and 30th observations, suggesting less reliable predictions. Overall, the HMM model demonstrates more controlled and balanced residual behavior.

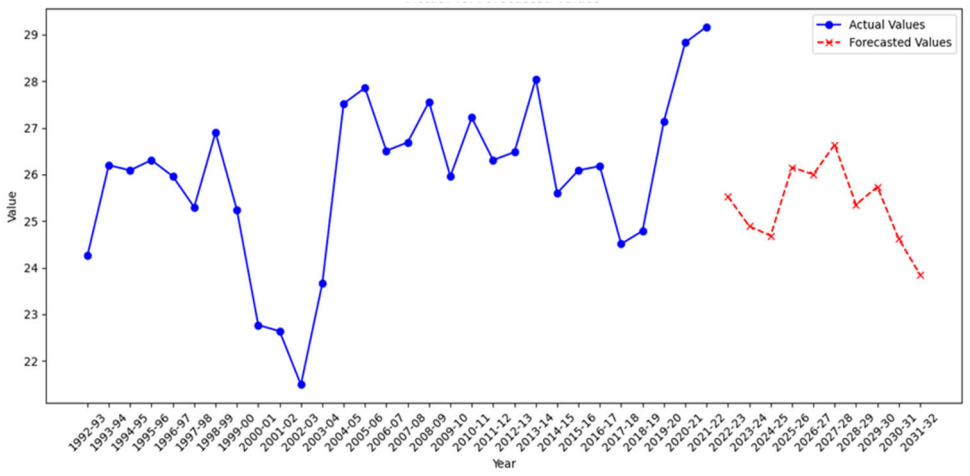

Figure 9. Comparison of actual and forecasted values over time using HMM parameters

The line chart presents both the actual and forecasted oilseed area values over time. The solid blue line represents the observed data, showing noticeable fluctuations with periods of sharp declines followed by recoveries. The forecasted values, indicated by a red dashed line, generally follow the overall trend of the historical data while displaying some variability. This suggests that the model effectively captures the main movement of the data and adapts to changing patterns, providing a useful projection for future values.

4. Conclusions

The findings of this study demonstrate that Hidden Markov Models (HMMs) significantly outperform other predictive methods in forecasting agricultural data. HMMs exhibit the lowest Mean Absolute Error (MAE) and the highest R-squared values, signifying their superior accuracy and reliability. The residual analysis further confirms the model's robustness, as the residuals are randomly distributed with no identifiable patterns, indicating a strong model fit. HMMs have proven to be highly effective in capturing the temporal and stochastic characteristics of agricultural data,

surpassing traditional machine learning techniques in performance. These results underline the potential of HMMs as a reliable tool for agricultural forecasting, enabling stakeholders to make data-driven decisions and develop strategic plans based on accurate predictions.

Additionally, this study utilizes HMM parameters to forecast agricultural trends for the next 10 years, providing valuable insights for long-term planning. Future research can expand on these findings by applying HMMs to different agricultural domains and exploring further enhancements to improve their predictive capabilities.

References

Agarwal, S., Tarar, S., (2021). A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. In Journal of *Physics*. Conference Series, Vol. 1714, page 012012.IOP Publishing.

Bafandeh, I. S., Bolandraftar, M., (2013). Application of K-Nearest Neighbor (KNN) approach for predicting economic events: Theoretical Background. *International Journal of Engineering Research and Applications*, Vol. 3, issue 5, pp. 605–610.

Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D. and Bochtis, D., (2021). Machine Learning in Agriculture: A Comprehensive updated review. *Sensors*, 21(11), 3758.

Chai, T., Ren, H., (2023). Risk prediction of financial management in agricultural companies based on RBF neural network and Markov. *Pakistan Journal of Agricultural Sciences*, 60(4), pp. 739–749.

Choukri, D., Adda, M., (2020). A new hybrid deep learning model-based recommender system using Artificial Neural Network and Hidden Markov model. *Procedia Computer Science*, 175, pp. 214–220.

Elavarasan, D., Vincent, P. M. D., (2020). Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. In *IEEE* Access, Vol. 8, pp. 86886–86901.

Erivwo, O., Makis, V. and Kwon, R., (2024). Bayesian change point prediction for downhole drilling pressures with hidden Markov models. *Applied Stochastic Models in Business and Industry*, 40(3), pp. 772–790.

French, M. N., Krajewski, W. F. and Cuykendall, R. R., (1992). Rainfall forecasting in space and time using a neural network. *Journal of Hydrology*, 137, pp. 1–31.

Hanusz, Z., Tarasinska, J. and Zielinski, W., (2016). Shapiro-Wilk test with known mean. *Revstat-Statistical Journal*, Vol. 14, No. 1, pp. 89–100.

Jakkula, V., (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5), 3.

Khosla, E., Dharavath, R. and Priya, R., (2020). Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environment, Development and Sustainability*, 22, pp. 5687–5708.

Kramer, W., (2011). Durbin-Watson test. In International Encyclopedia of Statistical Science. *Springer*, pp. 408–409.

Leite, P. B. C., Feitosa, R. Q., Formaggio, A. R., Da Costa, G. A. O. P., Pakzad, K. and Sanches, I. D. A., (2008). Hidden Markov models applied in agricultural crops classification, proceeding of GEOBIA (GEOgraphic Object-Based Image Analysis for the 21St Century).

Meyer, D. and Wien, F. T., (2001). Support vector machines. *R News*, 1(3), pp. 23–26.

Mithra, C. and Suhasini, (2023). Machine Learning based oilseed crop yield forecasting with recommendation system using organic manures in Tamil Nadu. *Journal of Survey in Fisheries Sciences*, 10(2S), pp. 1601–1620.

Senawongse, P., Dalby, A. R. and Yang, Z. R., (2005). Predicting the phosphorylation sites using Hidden Markov Models and machine learning methods. *Journal of chemical information and modelling*, 45(4), pp. 1147–1152.

Sharma, P., Dadheech, P., Aneja, N. and Aneja, S., (2023). Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning, in *IEEE Access*, Vol. 11, pp. 111255–111264.

Somani, P., Talele, S. and Sawant, S., (2014). Stock market prediction using Hidden Markov Model. *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*, pp. 89-92, Chongqing, China.

Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T. and Rahimi, J., (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Sci Rep*, 12, 3215.

Topp, R., Gomez, G., (2004). Residual analysis in linear regression models with an interval-censored covariate. *Statistics in medicine*, 23, pp. 3377–3391.

Tranmer, M., Elliot, M., (2008). Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research*, 5(5), pp. 1–5.

Yuan, H., Yang, G., Li, C., Wang, Y., Liu, J., Yu, H., Feng, H., Xu, B., Zhao, X. and Yang, X., (2017). Retrieving soybean leaf area index from unmanned aerial vehicle hyperspectral remote sensing: Analysis of RF, ANN, and SVM regression models. *Remote Sensing*, 9(4), 309.

Zhu, D., Ding, Z. and Huang, X., (2024). Probabilistic model for fatigue damage estimation of wind turbines with hidden Markov model and neural network. *Ocean Engineering*, 310, 118663.

Zou, K. H., Tuncali, K. and Silverman, S. G., (2003). Correlation and simple linear regression. *Radiology*, 227(3), pp. 617–628.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 137–153, https://doi.org/10.59139/stattrans-2026-008
Received: 12.08.2024; accepted: 30.07.2025

Jackknife-based diagnostics for non-monotonic hazard survival model with interval-censored data

Jayanthi Arasan[1]

Abstract

This study focuses on jackknife-based model diagnostics for a non-monotonic two-parameter hazard survival regression model (TBPR) when data is interval and right-censored. This distribution is very flexible, because it accommodates both monotonic and bathtub-shaped hazard rates. This research proposes a bias-corrected jackknife harmonic mean and a ran-dom imputation technique to obtain the altered Cox-Snell (r^*_{iC}), adjusted Martingale (r^*_{iM}) and Schoenfeld ($r^*_{S_i}$) residuals. Two simulation studies were conducted to assess the perfor-mances of the altered residuals and their ability to detect extreme observations and outliers at various censoring proportions (cp) and sample sizes (n) for this model. The results indicated that the altered residuals based on jackknife outperformed other residuals at cp and n levels. The proposed methods are then illustrated using a real dataset on Hodgkin's Disease with the prior treatment group as the covariate. The results showed that the altered residuals work well to address model adequacy and identify potential outliers in the dataset.

Keywords: Jackknife, interval-censored, outliers, covariate.

1. Introduction

Survival data with non-monotonic or bathtub hazard rates is commonly encountered in medical research. Some examples include lifetimes of kidney or heart transplant patients, lifetime of curability of breast cancer and lung cancer patients. The two-parameter distribution with bathtub shape (TPB) model was proposed by (Chen 2000) and extended by (Ismail, Arasan, Safie & Mohd Safari 2022) to incorporate covariates, resulting in what is known as the TPB regression (TPBR) model. This model is very flexible compared to other survival models as it accommodates both monotonic and non-monotonic, namely bathtub shaped hazard rates, see (Chen 2000).

This research focuses on the model diagnostics for the TBPR model when data is both right and interval-censored. Although residual analysis plays a central role in model diagnostics, traditional approaches such as the Cox-Snell residual often fail to perform well under right or interval censoring, particularly when the underlying hazard is non-monotonic. Existing adjustments like midpoint imputation or bootstrap methods have been explored primarily for simpler monotonic hazard models. However, real-world survival data often involve more complex hazard shapes (e.g. bathtub) and censoring types. This study addresses this gap by proposing a jackknife-based adjustment to residuals specifically for the

[1]Department of Mathematics and Statistics, Universiti Putra Malaysia, Malaysia.
E-mail: jayanthi@upm.edu.my. ORCID: https://orcid.org/0000-0003-1805-9601.

TBPR model, a flexible model that accommodates both monotonic and non-monotonic hazard functions.

Interval-censored data is prevalent in many clinical and longitudinal studies, primarily due to constraints such as time, cost, and the necessity for periodic inspections conducted at varying intervals. Data is interval-censored when the lifetime of the i^{th} patient lies within an interval, $t_{L_i} < t_i < t_{R_i}$, where t_{L_i} and t_{R_i} denote the left and right endpoints of the observed interval, respectively.

A special case of the interval-censored data, where $t_{L_i} < t_i < \infty$, gives us the right-censored data, see (Sun 2006), who provided a detailed overview of statistical methods for analyzing interval-censored failure time data, covering techniques like maximum likelihood, nonparametric, semiparametric, and Bayesian methods. (Lawless 1982) discussed statistical methods for analyzing interval-censored data, including current status data as a special case.

The use of computer intensive techniques such as the jackknife and bootstrap can be found in (Arasan & Lunn 2008), who compared alternative confidence interval estimation methods, including bootstrap and jackknife techniques, for the parameters of a parallel two-component system model with dependent failure and time-varying covariates, showing that the jackknife method outperforms bootstrap techniques for censored data. (Arasan & Lunn 2009), extended a parallel system survival model based on the bivariate exponential to include a time-varying covariate, evaluating parameter estimates at various censoring levels, comparing fixed vs. time-varying covariate models, and studying Wald, likelihood ratio, and jackknife methods for constructing confidence intervals, with applications to diabetic retinopathy data.

Following that, (Manoharan, Arasan, Midi & Adam 2015) compared the performance of Wald, likelihood ratio, and jackknife confidence intervals for the parameters of the log-normal distribution in the presence of left-truncated and right-censored survival data, finding that the jackknife method outperformed the others, particularly for small sample sizes with left-truncated data and low censoring. (Kiani, Arasan, Midi et al. 2012) examined the Gompertz model with time-dependent covariates and right-censored data, comparing its performance at different censoring levels and sample sizes, and evaluating Wald and jackknife methods for confidence intervals.

Survival models with interval-censored data have been explored by authors such as (Kiani & Arasan 2013), who extended the Gompertz model with time-dependent covariates for interval-censored data, comparing the performance of Wald and likelihood ratio methods for confidence interval estimation. The study highlighted the effectiveness of these methods in handling interval-censored data. (Fang, Arasan, Midi & Bakar 2015) compared jackknife and bootstrap confidence interval estimates for the parameters of a log-logistic model with censored data and covariates, evaluating their performance through coverage probability studies at various error probability levels and censoring proportions.

(Alharbi, Jayanthi, Haizum & Ling 2022) extended the generalized exponential model to include covariates for interval-censored data, evaluating the maximum likelihood estimator and Wald confidence intervals, with better performance observed at larger sample sizes and lower censoring proportions. Then, (Al-Hakeem, Arasan, Mustafa & Peng 2023) extended the generalized exponential distribution to incorporate time-dependent covariates for

interval-censored data, comparing maximum likelihood estimations and finding better performance with larger sample sizes and lower attendance probabilities. (Manoharan, Arasan, Midi & Adam 2020) assessed the performance of local influential diagnostics for the extended log-normal model with time-dependent covariates, left-truncation, and case-k interval censoring, comparing it with global diagnostics through a simulation study.

More recently, several models and inference methods have been developed for interval-censored survival data. For instance, (Zhou, Sun & Ibrahim 2021, Zhou & Sun 2021) explored transformation models and estimation techniques. (García Meixide, Lema & Vilar 2024) proposed a sparse neural network AFT model for interval-censored outcomes, demonstrating improved prediction performance over classical methods using real-world biomedical data. (Lou, Li & Sun 2024) developed a two-step semiparametric transformation approach to handle missing covariate issues, supported by simulations and an Alzheimer's disease dataset. (Zhang, Li & Weng 2023) introduced a valid inference procedure post-variable selection for the Cox model with interval-censored data, using lasso and asymptotic techniques. Lastly, (Pal, Peng & Aselisewine 2023) discussed a support vector–based semiparametric cure model that accommodates interval-censored survival times.

Other research related to survival models with covariates include (Arasan & Ehsani 2011), who applied a repairable system model for interval failure data with a time-dependent covariate, evaluating several NHPP-based models on ball bearing failure data and using bootstrapping for variance estimation. They found that the proposed model was effective and easy to implement. (Manoharan, Arasan, Midi & Adam 2017) extended the three-parameter log-normal survival model to incorporate left-truncated and right-censored data with covariates. They applied bootstrap inferential procedures to estimate the parameters and assessed the model's performance through a simulation study.

The Cox-Snell residuals (r_{C_i}) are commonly used for checking the fit of a model in survival analysis. When the data is positively skewed because of censoring, the Cox-Snell residuals tend to be smaller or less informative because they are based on the assumption that all observations are fully observed, which is not the case with censored data, as pointed out by (Cox & Snell 1968). To correct this, the Cox-Snell residuals can be modified by adding a positive surplus to make it more reliable. Two conventional modifications of r_{C_i} take the surplus as the mean ($r^{'}_{C_i}$) and median ($r^{''}_{C_i}$) of the standard exponential distribution, see (Cox & Snell 1968). The use of the median of the standard exponential distribution for the surplus was proposed by (Crowley & Hu 1977) as they found the mean tends to inflate the residual far too much. Normally, the arithmetic mean works well when the data is simple and does not have extreme values or outliers, as discussed by (Huber 1981). If the data contains extreme values, the arithmetic mean may not be ideal, as it can overly increase the residuals.

For survival data, which can often follow an exponential or skewed pattern, the geometric mean is a better option because it handles this kind of data more effectively. However, when the data contains extreme values or outliers, the harmonic mean is preferred because it is less affected by these extremes. (Naslina, Jayanthi, Syahida & Bakri 2020) and (Lai & Arasan 2020) deduced that the modified Cox-Snell residuals for the Gompertz model based on the empirical harmonic mean perform better than both standard and other modified Cox-Snell residuals. (Arasan & Midi 2021) concluded that harmonic mean and jackknife har-

monic mean residuals perform significantly better, especially when censoring proportions are high.

In the case of interval-censored data, where the exact timing of an event is unknown but falls within a specified range, traditional methods may not yield accurate results. When data is interval-censored, the Cox-Snell residuals themselves are also interval-censored. (Farrington 2000) recommends replacing the interval residuals with expected values under $\exp(1)$. However, this approach may be impractical for more complex models or when the data exhibits mixed-case censoring. To address this, this study proposes a change to the Cox-Snell residuals by using the jackknife bias-corrected harmonic mean and random imputation, which is better at dealing with heavy censoring. This adjustment is expected to give more reliable results, as shown in (Arasan & Midi 2023), especially when the data is censored in different ways and contains outliers, which can improve model assessment accuracy.

(Arasan & Midi 2023) introduced a method using the bias-corrected bootstrap harmonic mean and random imputation to adjust residuals for the extreme minimum value regression with right- and interval-censored data. The extreme minimum value regression model only accommodates monotonic hazards with a simpler data structure. Their study demonstrated that these adjusted residuals were effective for assessing model adequacy and identifying influential observations. In contrast, the current study focuses on modifying the Cox-Snell residuals using the jackknife bias-corrected harmonic mean and random imputation for a two-parameter distribution with a bathtub-shaped hazard, which has a more complex data structure. While both the study by (Arasan & Midi 2023) and the present work aim to improve residuals in the presence of censoring, our approach emphasizes the jackknife technique, particularly in cases of mixed censoring with a non-monotonic hazard rate. Although these two studies explore similar goals, they propose different models and computational techniques for addressing residual issues in survival analysis.

2. Methodology

2.1. The model

Let T be a non-negative random variable representing the survival time of an event. The density and survivor functions for the TBP model by (Chen 2000) are given by Eqs. (1) and (2).

$$f(t,\lambda,\gamma) = \lambda\gamma t^{\gamma-1}\exp\left(t^{\gamma}+\lambda(1-e^{t^{\gamma}})\right), \quad (1)$$

$$S(t,\lambda,\gamma) = \exp\left(\lambda\left(1-e^{t^{\gamma}}\right)\right), \quad t>0. \quad (2)$$

The effect of the covariates can be incorporated into the model by allowing the parameter λ to be a function of the covariates. If the vector of covariate values is $x' = (x_0, x_1, \cdots, x_{p-1})$, and the vector of regression coefficients is $\beta' = (\beta_0, \beta_1, \cdots, \beta_{p-1})$, then $\lambda = e^{-\beta' x}$, where $\gamma > 0$ represents an unknown parameter. The density and survivor functions for the TBPR

model are given by Eqs. (3) and (4).

$$f(t,\beta,\gamma) = \gamma t^{\gamma-1}\exp\left(-\beta'\mathbf{x}+t^{\gamma}+e^{-\beta'\mathbf{x}}(1-e^{t^{\gamma}})\right), \tag{3}$$

$$S(t,\beta,\gamma) = \exp\left(e^{-\beta'\mathbf{x}}(1-e^{t^{\gamma}})\right), \quad t>0. \tag{4}$$

The distribution has a monotonically increasing hazard function when $\gamma \geq 1$ and may have a bathtub-shaped hazard function when $\gamma < 1$. Consider the case where there are lifetimes for $i = 1,2,\ldots,n$ observations. Let the left and right endpoints for the i^{th} subject be t_{L_i} and t_{R_i}, respectively. To distinguish between censoring types for each observation, we define an indicator variable δ_i as follows:

$$\delta_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ observation is interval-censored,} \\ 0 & \text{if the } i^{\text{th}} \text{ observation is right-censored.} \end{cases} \tag{5}$$

The likelihood function for the full sample with interval and right-censored data is shown by Eq. (6).

$$l(\beta,\gamma) = \prod_{i=1}^{n}\left[S(t_{L_i})-S(t_{R_i})\right]^{\delta_i}\left[S(t_{L_i})\right]^{(1-\delta_i)}. \tag{6}$$

So, for the TBPR model the likelihood and log-likelihood functions for the full sample are shown by Eqs. (7) and (8).

$$l(\beta,\gamma) = \prod_{i=1}^{n}\left[e^{-\beta'x_i(1-exp(t_{L_i}^{\gamma}))}-e^{-\beta'x_i(1-exp(t_{R_i}^{\gamma}))}\right]^{\delta_i}\left[e^{-\beta'x_i(1-e^{t_{L_i}^{\gamma}})}\right]^{(1-\delta_i)} \tag{7}$$

$$\begin{aligned} L(\beta,\gamma) = {} & \sum_{i=1}^{n}\delta_i\left\{\log\left[e^{-\beta'x_i(1-exp(t_{L_i}^{\gamma}))}-e^{-\beta'x_i(1-exp(t_{R_i}^{\gamma}))}\right]\right\} \\ & +(1-\delta_i)\left\{e^{-\beta'x_i(1-e^{t_{L_i}^{\gamma}})}\right\}. \end{aligned} \tag{8}$$

The estimates for β and γ are obtained by solving the likelihood equations using any iterative technique suited for nonlinear equations. The inverse of the observed information matrix, denoted as $i(\widehat{\beta},\widehat{\gamma})$, can be computed from the second partial derivatives of the log-likelihood function, evaluated at $\widehat{\beta}$ and $\widehat{\gamma}$, providing estimates for the variance and covariance, as shown in Eq. (9).

$$\widehat{Var}(\widehat{\beta},\widehat{\gamma}) = [i(\widehat{\beta},\widehat{\gamma})]^{-1}. \tag{9}$$

2.2. The residuals

The Cox-Snell residual for the i^{th} subject is given by $r_{C_i} = \hat{H}(t_i) = -\log(\hat{S}(t_i))$, where $\hat{H}(t_i)$ and $\hat{S}(t_i)$ are the estimated cumulative hazard and survivor functions, respectively. As discussed by (Cox & Snell 1968), a challenge arises when dealing with censored data, particularly right-censored observations, as these residuals tend to underestimate the true values. To address this, we propose modified Cox-Snell residuals using bias-corrected harmonic means via the jackknife method and random imputation, depending on the type of censoring.

Right-Censored Data

To adjust Cox-Snell residuals under right-censoring, the jackknife bias-corrected harmonic mean is applied. The i^{th} jackknife sample is constructed by removing the i^{th} observation from the original dataset of n observations, as described by (Efron & Tibshirani 1994) and defined in Eq. (10).

$$t_{(i)} = (t_1, t_2, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n) \tag{10}$$

Let $\hat{\theta}_{h(i)}$ be the harmonic mean from the i^{th} jackknife sample. The average of these harmonic means is:

$$\hat{\theta}_{h(.)} = \sum_{i=1}^{n} \frac{\hat{\theta}_{h(i)}}{n}$$

The jackknife estimate of bias is given by $(n-1)(\hat{\theta}_{h(.)} - \hat{\theta}_h)$, where $\hat{\theta}_h$ is the harmonic mean of the full dataset. The jackknife bias-corrected estimate can then be obtained as shown in Eq. (11).

$$\hat{\theta}_{h_{jack}} = n\hat{\theta}_h - (n-1)(\hat{\theta}_{h(.)}). \tag{11}$$

The altered Cox-Snell residual when data is right-censored using the jackknife bias-corrected estimate is given by Eq. (12).

$$r_{C_i}^{\text{jack}} = r_{C_i} + \hat{\theta}_{h_{jack}} \quad \text{for the } i^{\text{th}} \text{ subject} \tag{12}$$

Interval-Censored Data

Residual analysis under interval-censoring is more complex (Farrington 2000). A practical method introduced by (Arasan & Midi 2023) uses random imputation. Let $S(\cdot)$ denote the model-based survivor function. For the i^{th} subject, generate R values from the uniform distribution $U(S(t_{R_i}), S(t_{L_i}))$, then transform these to obtain pseudo-lifetimes $t_i^{\,r}$, for $r = 1, 2, \ldots, R$. The imputed lifetime is estimated as:

$$t_i^{'} = \sum_{r=1}^{R} \frac{t_i^{\,r}}{R}$$

The adjusted Cox-Snell residual for interval-censored data is then:

$$r_{C_i}^{\text{int}} = \hat{H}(t_i^{'}) = -\log(\hat{S}(t_i^{'})) \tag{13}$$

General Formulation

Combining both scenarios, the modified Cox-Snell residual is defined as:

$$r_{C_i}^{*} = \begin{cases} r_{C_i}^{\text{jack}} & \text{if data is right-censored} \\ r_{C_i}^{\text{int}} & \text{if data is interval-censored} \end{cases} \tag{14}$$

Following that, the adjusted martingale and deviance residuals using the jackknife bias-corrected estimate are given by Eqs. (15) and (16).

$$r_{M_i}^{*} = \delta_i - r_{C_i}^{*}. \tag{15}$$

$$r_{D_i}^{*} = \text{Sgn}(r_{M_i}^{*})[-2(r_{M_i}^{*} + \delta_i \ln(\delta_i - r_{M_i}^{*}))]^{1/2}. \tag{16}$$

The score or Schoenfeld residual (r_{S_i}) was proposed by (Schoenfeld 1982), and is derived from the first derivatives of the log-likelihood function with respect to its parameters. Consequently, these residuals exhibit varying values for each parameter in the model. Since the data is both interval- and right-censored, the adjusted score residuals can be obtained using the imputed lifetimes discussed in Section 2.2. Let,

$$\tilde{t} = \begin{cases} t_i^{'} & \text{for } t_i \text{ interval-censored,} \\ t_{L_i} & \text{for } t_i \text{ right-censored.} \end{cases} \tag{17}$$

The log-likelihood for the full sample is given by Eq. (18).

$$\begin{aligned} \ell(\beta,\gamma) &= \sum_{i=1}^{n} \delta_i \log f(\tilde{t}_i,\beta,\gamma) + (1-\delta_i)\log S(\tilde{t}_i,\beta,\gamma) \\ &= \sum_{i=1}^{n} \delta_i \left[\log\gamma + (\gamma-1)\log\tilde{t}_i - \beta'\mathbf{x}_i + \tilde{t}_i^{\gamma} + e^{-\beta'\mathbf{x}_i}(1-e^{\tilde{t}_i^{\gamma}})\right] \\ &\quad + (1-\delta_i)\left[e^{-\beta'\mathbf{x}_i}(1-e^{\tilde{t}_i^{\gamma}})\right] \end{aligned} \tag{18}$$

The adjusted score residuals ($r_{S_i}^{*}$) can now be calculated from the components of the first derivatives of the log-likelihood function with respect to its parameters, β and γ, evaluated at their respective MLEs, see Eqs. (19) and (20).

$$\frac{\partial L(\beta,\gamma)}{\partial \beta_j} = \sum_{i=1}^{n} -x_{ij}\left[\delta_i + e^{-\beta'\mathbf{x}_i}(1-e^{\tilde{t}_i^{\gamma}})\right], j = 0, 1\cdots, p-1, \tag{19}$$

$$\frac{\partial L(\beta,\gamma)}{\partial \gamma} = \sum_{i=1}^{n}\left[\delta_i\left(\frac{1}{\gamma}+\ln\tilde{t}_i\right)+\tilde{t}_i^{\gamma}\ln\tilde{t}_i\left(\delta_i-e^{-\beta'\mathbf{x}_i}e^{\tilde{t}_i^{\gamma}}\right)\right]. \quad (20)$$

The plot of $r^*_{S_i}$ versus the observation number should be randomly distributed around zero for a good fit. Index plots of the score residuals for each covariate in the fitted model are useful at indicating extreme observations and outliers.

3. Simulation study

Two simulation studies were designed to assess different aspects of the proposed residual diagnostics as follows.

- **Simulation Study I (Sim I)** was designed to identify the most suitable modified residuals by evaluating their ability to assess model adequacy across different levels of censoring and sample sizes.
- **Simulation Study II (Sim II)** builds on the findings of Sim I and investigates the effectiveness of the best-performing residuals from Sim I in detecting extreme or influential observations, which is vital for model diagnostics in clinical survival data.

Sim I was conducted using 1000 replications, at $n = 50$, 80 and $n = 120$, with approximate right censoring proportions (cp) of 0.30, 0.40, 0.50, 0.55 and 0.60 for the TBPR model. The objective is to compare the effectiveness of altered Cox-Snell residuals, utilizing bias-corrected jackknife harmonic mean and multiple imputation, $r^*_{C_i}$ against r'_{C_i} and r''_{C_i}, using mid-point imputation. Mid-point imputation estimates lifetimes by using the average of the t_{L_i} and t_{R_i}. It assumes the true value is near the middle of the range. The simulation study only examines the effectiveness of the altered Cox-Snell residuals, as the values of the martingale and deviance residuals are based on these altered Cox-Snell residuals.

The values of β_0, β_1, and γ were set to 3.3, 0.95, and 0.42, respectively, to mimic the lifetime of cancer data, measured in months. Survival times were derived using the inverse transformation method. For the i^{th} observation, the censoring time c_i follows an exponential distribution with parameter μ, where the value of μ is adjusted to achieve the desired approximate right censoring proportion in our dataset. The covariate was simulated as a categorical variable with proportions set to $P = 0.5$ to mimic the distribution of treatment types among patients. The parameter estimates can be obtained by solving the likelihood equations using an iterative procedure designed for nonlinear equations. In this study, the maximum likelihood estimators for all parameters were computed employing the Newton-Raphson iterative method.

To generate interval-censored data, we utilized a sequence of 24 check-up times, $\tau_1, \tau_2, \cdots, \tau_\kappa$, spaced at two-month intervals, assuming all subjects attended these check-ups. Subsequently, we determined whether the uncensored lifetimes, t_i, fell within any of these intervals. If t_i fell within the interval (τ_m, τ_{m+1}) where $m \leq \kappa$, then the corresponding left and right bounds, t_{L_i} and t_{R_i}, for the i^{th} observation, were set to τ_m and τ_{m+1}, respectively. Otherwise, if $t_i > \tau_\kappa$, t_i would be right-censored at τ_κ.

To assess the efficacy of various modifications of the Cox-Snell residuals, it is necessary to derive the estimated Kaplan-Meier survivor function based on the values of these altered residuals. Let $\hat{S}(r^*_{C_i})$ represent the estimated Kaplan-Meier survivor function derived from the adjusted Cox-Snell residuals. The plot of $\log[-\log(\hat{S}(r^*_{C_i}))]$ against $\log(r^*_{C_i})$ should ideally manifest as a linear function with unit slope and intercept zero, as expected when the residuals follow an unit exponential distribution under a correctly specified model. Consequently, by applying the same methodology to the other residuals, their performances can be compared based on the mean absolute deviation (MAD) of three key metrics: the intercept, slope, and correlation coefficient R, from their ideal values of 0, 1, and 1, respectively, indicating a well-fitting model.

Sim II, with 1000 replications, was also carried out using sample sizes of $50, 80$, 200 and 360, along with approximate right censoring proportions (cp) set at 0.10 and 0.30 for the TBPR model. The covariate was simulated as categorical variable with a proportion set to $P = 0.5$ to mimic the distribution of two different treatment types among patients. The purpose of this simulation study is to assess and compare the effectiveness of the best adjusted residuals in detecting extreme observations and outliers. Two data points were randomly chosen and perturbed to yield extreme observations compared to the others. This was achieved by altering the m^{th} lifetime, t_m, by an amount $\omega = 3.5$ scaled by the standard deviation of the lifetimes, s_t, and the largest censored observation, t_{max}, resulting in $t^{'}_m = t_m + \omega s_t + t_{max}$.

The detection percentage was determined based on whether the randomly selected outliers produced the two largest absolute values of the adjusted residuals. For the score residuals, the residual corresponding to the covariate parameter was used to detect outliers. The detection rate was further categorized into two cases: the percentage of datasets where both outliers were detected and the percentage of datasets where only one outlier was detected. In some cases, the methods did not detect any outliers. The overall detection rate was calculated by considering both full and partial detections. Specifically, full weight was given to cases where both outliers were detected, while half weight was assigned to cases where only one outlier was detected.

3.1. Simulation results

The results of Sim I are given in Figures 1-3. The plots demonstrate that the newly proposed adjusted residual, $r^*_{C_i}$ consistently exhibits significantly lower MAD values for intercept, slope, and correlation coefficient (R) across all levels of censoring proportions and sample sizes. This indicates superior performance by $r^*_{C_i}$ in assessing model adequacy. Although performance of $r^{''}_{C_i}$ is marginally superior to that of $r^{'}_{C_i}$, $r^*_{C_i}$ notably surpasses both in indicating a well-fitted model. As n increases, the gap narrows, but $r^*_{C_i}$ still maintains a clear advantage, supporting its robustness across different data conditions. These results confirm the effectiveness of the jackknife bias correction and random imputation method in improving residual-based diagnostics for right- and interval-censored data.

Sim II results, shown in Table 1, evaluate the ability of four adjusted residuals, $r^*_{C_i}$, $r^*_{M_i}$, $r^*_{D_i}$, and $r^*_{S_i}$, to detect outliers under two censoring scenarios: $cp = 0.10$ and $cp = 0.30$. The values outside parentheses correspond to $cp = 0.10$, while those in parentheses refer to

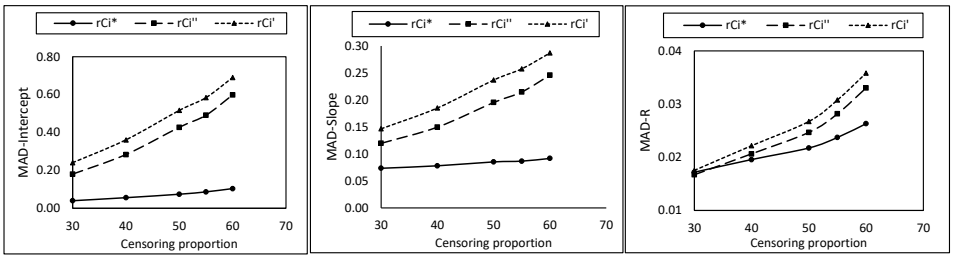

Figure 1. MAD for TBPR model at $n = 50$

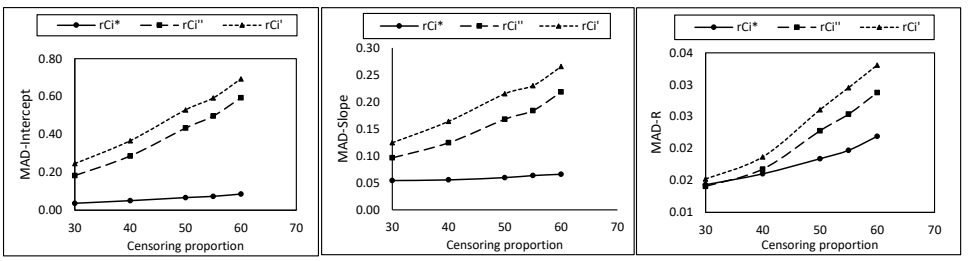

Figure 2. MAD for TBPR model at $n = 80$

$cp = 0.30$. The goal is to measure how often the two deliberately perturbed observations are correctly identified as the most extreme. The results indicate that the newly proposed adjusted Cox-Snell residual, $r^*_{C_i}$, performs best in detecting both outliers in over 99% of simulations even at small sample sizes, and maintaining perfect detection rates at $n = 200$ and above. It is followed by $r^*_{M_i}$, then $r^*_{D_i}$ and $r^*_{S_i}$, in terms of detection accuracy.

The performance of all methods improve as n increases and when $n = 360$, where all residuals except the adjusted score residuals achieved 100% detection for both outliers. When $cp = 0.3$, $r^*_{C_i}$ and $r^*_{M_i}$ remain robust, maintaining overall detection above 95%, even with a sample size as low as $n = 50$. However, the performance of $r^*_{D_i}$ declines rapidly, achieving only 65.9% overall detection compared to $r^*_{S_i}$, which achieves 86.3%. Once again, all performances improve as n increases, particularly $r^*_{D_i}$, which begins to outperform $r^*_{S_i}$ when $n \geq 200$. However, only $r^*_{C_i}$ and $r^*_{M_i}$ achieve 100% detection for both outliers at all censoring levels when $n = 360$.

These findings demonstrate the effectiveness of $r^*_{C_i}$ for assessing model adequacy and detecting outliers, especially in complex censoring settings. Comparing the performance of different residuals across sample sizes and censoring levels also offers practical guidance for researchers and clinicians in selecting suitable diagnostics for survival analysis. Together, the results from both simulation studies confirm the reliability and robustness of the proposed residual adjustments.

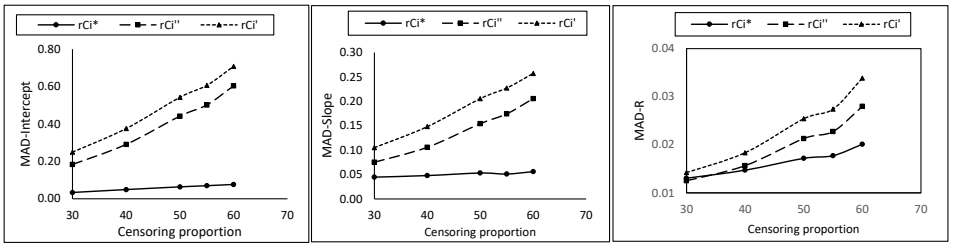

Figure 3. MAD for TBPR model at $n = 120$

Table 1. Percentage detection for different sample sizes and censoring proportions

n	Type	2 obs	1 obs	Overall
50	$r^*_{C_i}$	99.0 (97.5)	1.0 (2.5)	99.5 (98.8)
	$r^*_{M_i}$	97.3 (90.2)	2.7 (9.8)	98.7 (95.1)
	$r^*_{D_i}$	71.7 (39.6)	26.9 (52.5)	85.2 (65.9)
	$r^*_{S_i}$	76.1 (76.3)	20.8 (20.0)	86.5 (86.3)
80	$r^*_{C_i}$	100.0 (99.1)	0.0 (0.9)	100.0 (99.6)
	$r^*_{M_i}$	99.7 (97.1)	0.3 (2.9)	99.9 (98.6)
	$r^*_{D_i}$	86.5 (55.3)	13.2 (40.3)	93.1 (75.5)
	$r^*_{S_i}$	84.1 (80.3)	14.5 (16.9)	91.4 (88.8)
200	$r^*_{C_i}$	100.0 (100.0)	0.0 (0.0)	100.0 (100.0)
	$r^*_{M_i}$	100.0 (100.0)	0.0 (0.0)	100.0 (100.0)
	$r^*_{D_i}$	99.0 (89.1)	1.0 (10.9)	99.5 (94.6)
	$r^*_{S_i}$	94.3 (88.5)	5.7 (10.7)	97.2 (93.9)
360	$r^*_{C_i}$	100.0 (100.0)	0.0 (0.0)	100.0 (100.0)
	$r^*_{M_i}$	100.0 (100.0)	0.0 (0.0)	100.0 (100.0)
	$r^*_{D_i}$	100.0 (98.5)	0.0 (1.5)	100.0 (99.3)
	$r^*_{S_i}$	97.1 (93.8)	2.9(6.1)	98.6 (96.9)

Values outside parentheses correspond to $cp = 0.10$, while values in parentheses correspond to $cp = 0.30$.

4. Real example on Hodgkin's Disease

In this section, we apply the proposed methods to a real dataset to demonstrate the practical applicability of the modified residuals. The dataset comprises the survival times (in months) of 35 patients diagnosed with Hodgkin's Disease and treated with nitrogen mustards, as originally analyzed by (Bartolucci & Dickey 1977). The survival time represents the duration from treatment initiation to either death or censoring. Patients were classified into two groups: Group 1 received minimal or no prior therapy, while Group 2 underwent heavy prior therapy. Among these patients, 9 were right-censored, resulting in a censoring proportion of $cp = 0.257$, which falls within the range considered in our simulation studies.

This dataset was selected to evaluate the diagnostic performance of the adjusted residuals in detecting model fit and influential observations in a real-world clinical scenario. We focus on checking the fit of the TBPR model and testing the modified residuals with both right- and interval-censored data. By comparing the results from our simulations with the

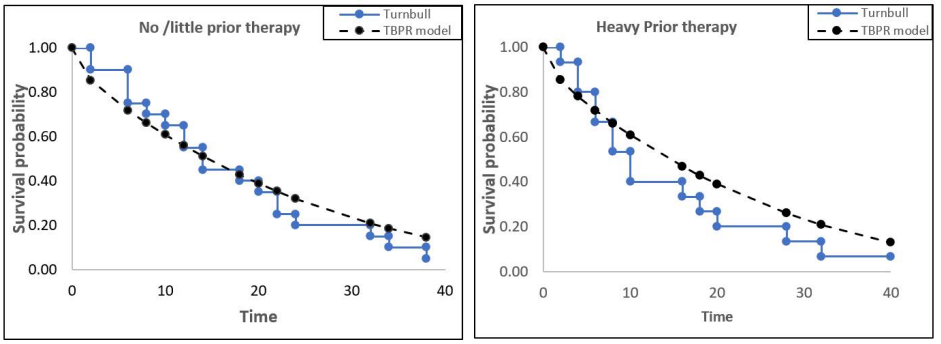

Figure 4. Turnbull and TBPR Survivor Function Estimates for Hodgkin's Disease data by Group

real data, this section will help confirm the findings from the simulation study and shows how useful the proposed methods are for model diagnostics in a clinical context.

The TBPR model was fitted using the treatment group as a categorical covariate. To align with the objectives of this study, the data was modified to create interval-censored data with a 2-month width. To assess the model fit, we obtain the Turnbull estimate of the survivor function (TB) and compare it with the parametric survivor function obtained using the TBPR model for each patient group. Figure 4 presents the plots, indicating that employing the TBPR model would be rather appropriate for the dataset. The survival functions for the two groups indicate that the patients who received little or no prior therapy have slightly better chances of survival than the patients who received heavy prior therapy.

Table 2 displays the parameter estimates obtained from fitting the TBPR model to the Hodgkin's Disease dataset, with group as the covariate. The p-value associated with β_1 indicates a lack of statistically significant difference between patients who received minimal or no prior treatment and those who underwent heavy prior treatments. According to the estimated parameters, the median survival time for patients in the first and second groups is 14.5 and 14.2 months, respectively, indicating a relatively small difference. Figures 5 and 6 show the index plots for $r^*_{C_i}$, $r^*_{M_i}$,$r^*_{D_i}$, and $r^*_{S_i}$ for the Hodgkin's Disease data.

All plots except the $r^*_{D_i}$ plot indicate that observation 30 exceeds the two standard deviations from the mean limit, respectively. Thus, it is important that we investigate this observation thoroughly. Patient 30 had the longest censored lifetime of approximately 40 months among those who underwent extensive prior treatment. All other observations, whether they experienced failure (uncensored) or were censored, had survival times shorter than observation 18. The $r^*_{S_i}$ plot was the only one that singled out observation 18 in addition to observation 30. This was the second largest censored lifetime of approximately 30 months among those who underwent extensive prior treatment. All other patients, whether they experienced an event (failure) or were censored, had survival times shorter than that of observation 18. However, since observation 18 was not flagged as extreme in either the $r^*_{C_i}$ and $r^*_{M_i}$ plots, both of which exhibited superior performance in the simulation study, it is unlikely to be a true outlier.

Table 2. Estimates and 95% Wald interval for the parameters of TBPR Model

Parameter	Estimates	Std.Err	Z	P Val	lower	upper
β_0	2.794	0.693	4.030	0.000	1.435	4.153
β_1	-0.0199	0.404	-0.049	0.961	-0.811	0.772
γ	0.341	0.035	9.830	0.000	0.273	0.410

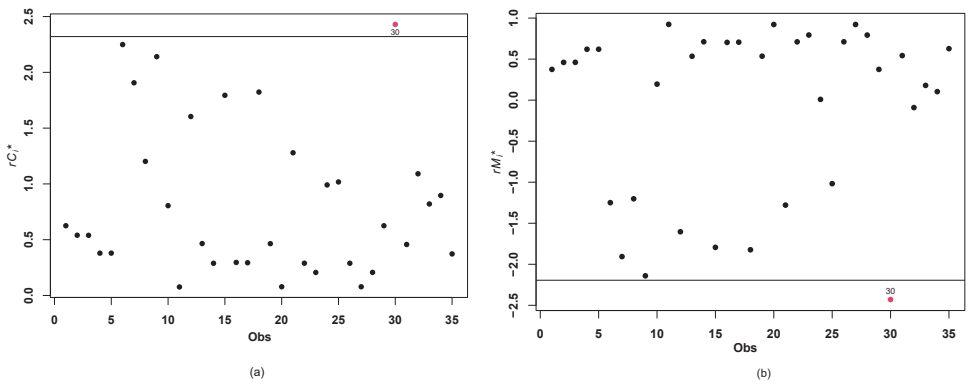

Figure 5. Index plot of adjusted Cox-Snell (a) and adjusted martingale (b) residuals for Hodgkin's Disease data.

The analysis of the Hodgkin's Disease dataset confirmed the findings from the simulation studies. The adjusted residuals, particularly $r^*_{C_i}$ and $r^*_{M_i}$, effectively identified observation 30 as an influential outlier, demonstrating their reliability across both simulated and practical scenarios. This is clearly illustrated in Figure 5, where observation 30 appears as a distinct outlier with substantially higher residual values in both the $r^*_{C_i}$ and $r^*_{M_i}$ plots, reinforcing the simulation findings. These results support the practical effectiveness of the proposed methods in real-world survival analysis, even under moderate censoring. The TBPR model showed a good fit to the data, and patients with minimal or no prior therapy exhibited slightly better survival outcomes.

These real data findings align with those observed in Simulation Study II. The adjusted residuals $r^*_{C_i}$ and $r^*_{M_i}$ consistently identified the most extreme observations, confirming their robustness and diagnostic value. Figure 5 highlights observation 30 as a clear outlier, further validating the ability of these residuals to detect influential cases. In addition, the comparison of model-based and Turnbull survivor curves supports the adequacy of the TBPR model. Together, the results demonstrate that the proposed residual adjustments serve as reliable tools for evaluating model fit and identifying outliers, making them valuable for practical use in survival analysis with interval- or right-censored data.

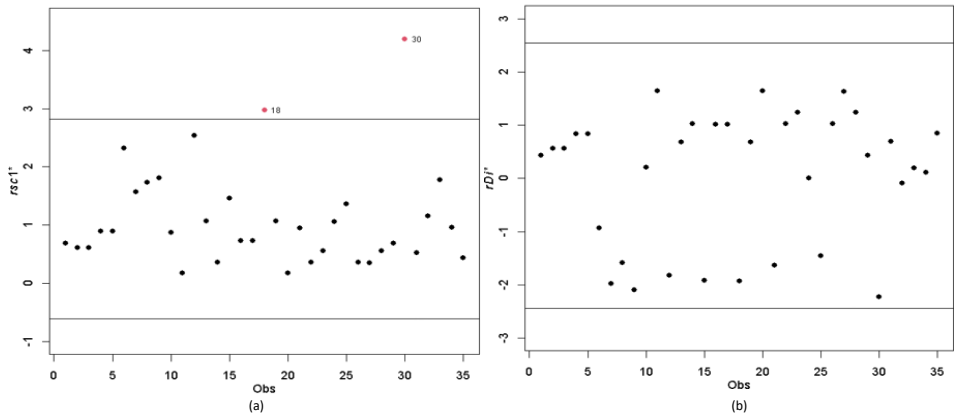

Figure 6. Index plot of adjusted score (a) and adjusted deviance (b) residuals for Hodgkin's Disease data.

5. Conclusion

In this study, we aimed to develop and evaluate modified residuals for model diagnostics in survival analysis, particularly for the TBPR model with right- and interval-censored data. Specifically, we explored the performance of various residuals modified using bias-corrected jackknife harmonic means with random imputation. The results of the first simulation study showed that the newly proposed adjusted residual, $r^*_{C_i}$, consistently outperformed other variations of the Cox-Snell residuals in detecting model fit, exhibiting significantly lower mean absolute deviation (MAD) values across all levels of censoring proportions and sample sizes. While r''_{C_i} performed slightly better than r'_{C_i}, $r^*_{C_i}$ still proved superior in identifying well-fitted models.

The second simulation study showed that $r^*_{C_i}$ and $r^*_{M_i}$ residuals were particularly effective in detecting influential observations, while $r^*_{S_i}$ and $r^*_{D_i}$ residuals performed poorly, although their performance improved as sample sizes increased. These findings highlight the importance of selecting the appropriate residuals for specific types of censoring and sample sizes. Thus, the newly proposed $r^*_{C_i}$ residual significantly outperformed other variations of the Cox-Snell residuals in terms of model diagnostics and detecting extreme observations. Our objective was to improve upon traditional residuals and test their effectiveness through both simulation and real data.

We also applied the proposed methods to a modified real dataset on Hodgkin's Disease patients. The goal was to demonstrate how the proposed residual adjustments perform in a real-world survival analysis. We focused on assessing the fit of the TBPR model and testing the modified residuals with both right- and interval-censored data, which are common in survival analysis. By comparing the results from our simulations with the real data, we were able to confirm the findings from the simulation study and show how useful the proposed methods are for model diagnostics in a clinical context. The results showed that the modified residuals were effective in detecting extreme and influential observations, for

the TBPR model, which aligns with the objectives of this study. For instance, the $r^*_{C_i}$ and $r^*_{M_i}$ residuals successfully identified outliers in the Hodgkin's Disease data, confirming their utility in practical, real-world survival analysis. Therefore, the study's objectives were successfully achieved: the proposed modifications to the Cox-Snell residuals improved model diagnostics and provided meaningful results in both simulated and real-world data contexts.

The methods presented in this research, being computationally intensive and empirically driven, can easily be extended to other models, such as bivariate or parallel-system models, and can handle different types of data, including truncated, left-censored, and mixed-case censored data. Further exploration could involve using double bootstrap techniques to refine these diagnostics. Finally, the analysis of the Hodgkin's Disease dataset illustrates that the TBPR model is suitable for the data, with patients who received minimal or no prior therapy having slightly better survival outcomes compared to those who received heavy prior therapy.

References

Al-Hakeem, H. A., Arasan, J., Mustafa, M. S. B. and Peng, L. F., (2023). Generalized exponential distribution with interval-censored data and time dependent covariate. *Communications in Statistics-Simulation and Computation,* 52(12), pp. 6149–6159.

Alharbi, N., Jayanthi, A., Haizum, A. and Ling, W., (2022). Assessing performance of the generalized exponential model in the presence of the interval censored data with covariate. *Austrian Journal of Statistics,* 51(1), pp. 52–69.

Arasan, J., Ehsani, S., (2011). Modeling repairable system failures with interval failure data and time dependent covariate. *Journal of Modern Applied Statistical Methods,* 10, pp. 618–624.

Arasan, J., Lunn, M., (2008). Alternative interval estimation for parameters of bivariate exponential model with time varying covariate. *Computational Statistics,* 23, pp. 605–622.

Arasan, J., Lunn, M., (2009). Survival model of a parallel system with dependent failures and time varying covariates. *Journal of Statistical Planning and Inference,* 139(3), pp 944–951.

Arasan, J., Midi, H., (2021). Jackknife and bootstrap estimates for modified residuals of the log-logistic model. *in AIP Conference Proceedings*, Vol. 2423(1), AIP Publishing LLC, p. 070009.

Arasan, J., Midi, H., (2023). Bootstrap based diagnostics for survival regression model with interval and right-censored data. *Austrian Journal of Statistics,* 52(2), pp. 66–85.

Bartolucci, A. A., Dickey, J. M., (1977). Comparative bayesian and traditional inference for gamma-modeled survival data. *Biometrics* pp. 343–354.

Chen, Z., (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics & Probability Letters,* 49(2), pp. 155–161.

Cox, D. R., Snell, E. J., (1968). 'A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological),* 30(2), pp. 248–265.

Crowley, J., Hu, M., (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association,* 72(357), pp. 27–36.

Efron, B., Tibshirani, R. J., (1994). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York.

Fang, L. Y., Arasan, J., Midi, H. and Bakar, M. R. A., (2015). Jackknife and bootstrap inferential procedures for censored survival data. *in* AIP Conference Proceedings, Vol. 1682(1), AIP Publishing.

Farrington, C. P., (2000). Residuals for proportional hazards models with interval-censored survival data. *Biometrics,* 56(2), pp. 473–482.

García Meixide, A., Lema, M. and Vilar, J. M., (2024). A bayesian semiparametric mixture cure model for doubly interval-censored data. *Computational Statistics & Data Analysis,* 190, p. 107925.

Huber, P. J., (1981). *Robust Statistics*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Ismail, I., Arasan, J., Safie, M. and Mohd Safari, M. A., (2022). Bathtub hazard model with covariate and right censored data. *Journal of Quality Measurement and Analysis JQMA,* 18(3), pp. 1–15.

Kiani, K., Arasan, J., (2013). Gompertz model with time-dependent covariate in the presence of interval-, right-and left-censored data. *Journal of Statistical Computation and Simulation,* 83(8), pp. 1472–1490.

Kiani, K., Arasan, J., Midi, H. et al., (2012). Interval estimations for parameters of gompertz model with time-dependent covariate and right censored data. *Sains Malaysiana,* 41(4), pp. 471–480.

Lai, M. C., Arasan, J., (2020). Single covariate log-logistic model adequacy with right and interval censored data. *Journal of Quality Measurement and Analysis,* 16(2), pp. 131–140.

Lawless, J. F., (1982). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

Lou, Y., Li, G. and Sun, J., (2024). Interval-censored quantile regression based on fractional counting process. *Statistical Analysis and Modeling,* XX(XX), pp. 1–20.

Manoharan, T., Arasan, J., Midi, H. and Adam, M. B., (2015). A coverage probability on the parameters of the log-normal distribution in the presence of left-truncated and right-censored survival data. *Malaysian Journal of Mathematical Sciences,* 9(1).

Manoharan, T., Arasan, J., Midi, H. and Adam, M. B., (2017). Bootstrap intervals in the presence of left-truncation, censoring and covariates with a parametric distribution. *Sains Malaysiana,* 46(12), pp. 2529–2539.

Manoharan, T., Arasan, J., Midi, H. and Adam, M. B., (2020). Influential measures on log-normal model for left-truncated and case-k interval censored data with time-dependent covariate. *Communications in Statistics-Simulation and Computation,* 49(6), pp. 1445–1466.

Naslina, A. M. N. N., Jayanthi, A., Syahida, Z. H. and Bakri, A. M., (2020). Assessing the goodness of fit of the gompertz model in the presence of right and interval censored data with covariate. *Austrian Journal of Statistics,* 49(3), pp. 57–71.

Pal, N., Peng, Y. and Aselisewine, S., (2023). Bayesian cure rate modeling of interval censored data based on negative binomial distribution. *Statistics & Probability Letters,* 204, p. 109984.

Schoenfeld, D., (1982). Partial residuals for the proportional hazards regression model. *Biometrika,* 69(1), pp. 239–241.

Sun, J., (2006). *The statistical analysis of interval-censored failure time data*, Vol. 3, No. 1, Springer.

Zhang, J., Li, G. and Weng, C., (2023). A semiparametric transformation model for multivariate interval-censored failure time data. *Lifetime Data Analysis,* 30, pp. 41–65.

Zhou, H., Sun, J., (2021). Semiparametric transformation model for interval-censored survival data with covariate measurement error. *Biometrics,* 77(2), pp. 523–535.

Zhou, H., Sun, J. and Ibrahim, J. G., (2021). A review on interval-censored survival data: models, inference methods, and applications. *Statistical Methods in Medical Research,* 30(5), pp. 1312–1336.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 155–168, https://doi.org/10.59139/stattrans-2026-009
Received – 09.05.2025; accepted – 06.11.2025

The application of BERTopic models to the analysis of Polish research publications in the field of economics and management

Paweł Lula[1]

Abstract

The main objective of the article is to analyze topics from the field of economics and management discussed in the Polish publications from 2000 to 2024. The research process allowed the identification of the main topics and the evaluation of their importance in subsequent years covered by the analysis. The BERTopic model was chosen as the main research method. The paper presents both the theoretical basis of the employed research method and the results of its application to the analysis of the Polish publication achievements registered in the Scopus database. The paper presents a description of topics identified, a specification of the relationship between them and changes in the importance of each topic between 2000 and 2024. All calculations were performed using computer programs prepared in Python language.

Key words: publication achievements, topic modelling, BERTopic method.

1. Introduction

The analysis of issues discussed in Polish publications related to the field of economics and management in the period 2000–2024 was the main goal of the research. Topic modelling, and BERTopic model in particular, was chosen as the main research tool.

Topic modelling belongs to main subareas of the natural language processing. It allows for identification of main issues raised in large collections of documents and for the evaluation of the significance of identified topics. The development of methods of topic modelling and analysis can be observed since the 1990s. A brief overview of the approaches used in this field can be found in Section 2 of the paper. Section 3 presents BERTopic models, while Section 4 discusses methods for assessing topic model's quality. Section 5 presents the results of the analysis of Polish publication achievements in the area of economics and management in the period 2000–2024.

[1] Krakow University of Economics, Krakow, Poland. E-mail: pawel.lula@uek.krakow.pl.
ORCID: https://orcid.org/0000-0003-2057-7299.

2. Topic modelling

Topic modelling allows for the identification and description of main issues discussed in a collection of documents. Having analyzed works on the use of statistical methods for natural language processing, several different approaches to the problem of topic modelling in documents can be identified:

1. Algebraic methods – among which Latent Semantic Analysis (LSA) (Deerwester *et al.*, 1990) is the best known solution. This method is based on frequency matrix representation and allows for the presentation of documents and words in a common base in which dimensions correspond to latent semantic components that can be interpreted as the main issues represented in the corpus. From the computational side, LSA is based on the SVD decomposition of the frequency matrix. Also, non-negative matrix factorization can be used for topics identification (Lee and Seung, 1999).

2. Probabilistic methods – in this approach every topic is described by specifying the distribution over words and every document is represented by the distribution over topics. Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003) method is the best known representative of this group of models. LDA can be considered as a generalization of the probabilistic latent semantic analysis (Hofmann, 1999).

3. Transformer-based models. Transformers can be defined as linguistic models that are able to process sequences of tokes (Vaswani *et al.*, 2017). They take into account the semantic aspects of words by using an embedding-based representation. They allow of describing the relationships between words through the use of the attention mechanism. Complex neural networks are used in their construction, where the learning process is carried out based on large corpora of documents. BERTopic technique is one of the most popular approaches belonging to this group of models (Grootendorst, 2022).

3. BERTopic model

BERTopic technique allows for the identification, description and analysis of topics discussed in the collection of documents. This method consists of the following steps:

1. Calculation of sentence embeddings.
2. Dimensionality reduction of embeddings.
3. Identification of topics by clustering of reduced embeddings.
4. Building topic's description.

3.1. Calculation of sentence embeddings

An embedding is a vector representing a given object in the semantic space. The more similar the objects are to each other, the smaller the distance between their embeddings. In natural language processing, embeddings can represent words, sentences, paragraphs or whole documents. Embeddings should present linguistic objects embedded in their context. One of the first researchers to draw attention to the crucial role of context in understanding words was John Rupert Firth (Firth, 1962).

For comprehensive presentation of the process of calculating embeddings of sentences, the architecture of the BERT model should first be presented (Devlin et al., 2019). Taking the transformer architecture as a starting point, it can be concluded that the BERT model performs the functions of the encoder, which determines the numerical representation for tokens comprising the input sequence. BERT is a neural network model which:

- takes as input a sequence of tokens forming two sentences,
- is trained to solve two types of tasks: predicting the missing word in a sentence based on the remaining words, and checking whether two input sentences form a logical sequence,
- uses the attention mechanism to describe the relationships between words forming input sentences,
- is used for the calculation of embeddings - output values of the neural network calculated for a given input word form its embedding.

SBERT (Reimers and Gurevych, 2019) is a version of the BERT model optimized to calculate sentence embeddings. For SBERT model, it is assumed that one sentence is provided as an input and that values of embedding vector for the whole sentence are produced as an output.

3.2. Dimensionality reduction of embeddings

Sentence embeddings calculated with the use of the SBERT model are vectors with several hundred elements. During the current step of the analysis, embeddings are reduced to vectors of a few or a dozen elements. Most often this operation is performed using the UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction) algorithm, which was introduced in (McInnes, Healy and Melville, 2020).

The UMAP algorithm uses weighted graphs to describe the structure of objects in high-dimensional space and the structure of objects' projections in low-dimensional space. The main objective of the method is to determine such a configuration of objects in a low-dimensional space for which the dissimilarity measure between the graphs describing the distribution of objects in each space will be the smallest. In graphs

describing the distribution of objects, edges are created between each node and its n nearest neighbours. The weight assigned to the edge between i-th and j-th vertex defines the probability that a relationship between these vertices exists. All probabilities of link existence between the i-th vertex and the vertices not belonging to its neighborhood are assumed to be zero. The matrices $\mathbf{W}_{n\times n}$ and $\mathbf{V}_{n\times n}$ are the weight matrices of the graphs describing the arrangement of objects in high-dimensional and low-dimensional space. Cross entropy is taken as a measure of the dissimilarity of graphs:

$$H(\mathbf{W},\mathbf{V}) = -\sum_{i,j} w_{ij}\log v_{ij} + \big(1-w_{ij}\big)\log\big(1-v_{ij}\big) \tag{1}$$

Using the UMAP method, the optimization algorithm searches for such a distribution of object projection in low-dimensional space for which $H(\mathbf{W},\mathbf{V})$ takes the smallest value.

To summarize the current section, it may be stated that the UMAP-step transforms sentence embeddings into vectors with several elements, in a way that minimizes the loss of semantic information of sentences.

3.3. Cluster analysis of reduced embeddings

In this step of the analysis sentence embeddings are grouped into clusters with the use of the HDBSCAN method (Hierarchical Density-Based Spatial Clustering of Applications with Noise) (Campello Ricardo J. G. B. and Moulavi, 2013).

In the first step, the HDBSCAN method estimates the probability density function for the analyzed set of points. Next, potential clusters are extracted by finding regions of data space corresponding to every peak of probability density function. The main problem which should be solved is related to the distinction of peaks representing clusters from peaks corresponding to a group of objects forming a part of a larger cluster. This decision is based on the comparison of probability masses of descendant clusters with the probability mass of ancestor cluster reduced by the sum of children masses. If probability masses of descendant clusters are dominant then a current cluster should be split into two new clusters. HDBSCAN splits objects into clusters in a way that maximizes the sum of probability masses of recognized clusters.

The idea presented above is implemented by performing the following steps:

1. Calculation of mutual reachability distances between every pair of embeddings.

   Let us assume that $d(x,y)$ is a distance between x and y points and r_x^n and r_y^n are radii of the smallest circles with centers respectively at point x and y containing n points belonging to the neighborhood of each of these points. Then the mutual reachability distance between x and y point may be defined as:

$$d_{MRD}(x,y) = max\big(d(x,y), r_x^n, r_y^n\big) \tag{2}$$

If points being compared are densely distributed in the space, then $d_{MRD}(x, y)$ is equal to $d(x, y)$. In the case of sparsely distributed points, the $d_{MRD}(x, y)$ is greater than $d(x, y)$.

2. Building a minimum spanning tree (MST).

   Every pair of objects x and y, for witch $d_{MRD}(x, y) > 0$ is linked by an edge to create an undirected graph with $d_{MRD}(x, y)$ as weights. Next, a minimum spanning tree is found with the use of Prim's algorithm (Prim, 1957). This operation is equivalent to dendrogram building with the use of single linkage method and mutual reachability distance.

3. Performing a pruning process.

   Leaves of the MST are combined to form groups containing the required number of objects.

4. Extraction of clusters.

   The main objective of this step is to answer the question whether the probability mass of the descendant clusters is high enough to separate them into separate clusters. Estimation of the probability mass corresponding to a given cluster is performed by analyzing the weights assigned to the edges leading from the node forming a given cluster to the nodes where potential descendant clusters are created.

During this step of analysis sentences are grouped into clusters. Clusters in which the number of elements exceeds the declared threshold value are treated as topics. The remaining sentences are treated as noise.

3.4. Building topic's description

For every extracted topic, its description is built. It has a form of a sequence of words which are crucial to a given topic. A class-based version of the TFIDF schema is used to create topic description (Sparck Jones, 1972). The algorithm is composed of several steps:

1. All sentences assigned to every cluster are merged into separate document,
2. For a set of documents obtained as a result of step 1, a frequency matrix $\mathbf{TF}_{[W\times C]} = [f_{ij}]$, where $i = 1, \ldots, W$ indicates a word, and the $j = 1, \ldots, C$ represents a cluster, symbol f_{ij} denotes the number of occurrences of the i-th word in the j-th cluster.
3. Weights are calculated with the use of the formula:

$$w_{ij} = f_{ij} \times \log\left(1 + \frac{A}{f_i}\right) \tag{3}$$

where A is an average number of words per class and f_i denotes frequency of the i-th word across all classes.

4. Labels for j-th cluster are created by merging words with highest values of w_{ij}.

4. BERTopic model quality

One of the main methods used to evaluate topic models is coherence measure C_V, which can be defined as the average of consistency coefficients calculated for the n most important words for each topic. The below presentation of C_V is based on (Rijcken, 2023).

The concept of pointwise mutual information (PMI) is a starting point for defining the consistency of words. PMI is a measure of association between two events x and y and can be defined as:

$$\mathrm{pmi}(x;y) = \log\frac{P(x;y)}{P(x)P(y)} \tag{4}$$

PMI compares the probability of the simultaneous occurrence of two events with the probability of their simultaneous occurrence when they are independent. The PMI value can be normalized using the formula:

$$\mathrm{npmi}(x;y) = \frac{\mathrm{pmi}(x;y)}{-\log(P(x;y))} \tag{5}$$

where npmi is a normalized (to $[-1;1]$ range) pointwise information and $-\log(p(x;y))$ is a self-information (Shannon information) related to the message about simultaneous occurrence of x and y.

Assuming that:

- $\boldsymbol{D} = \{d_1, d_2, \dots, d_{|\boldsymbol{D}|}\}$ is a set of documents,
- $\boldsymbol{d_i} = [w_{i,1}^d, w_{i,2}^d, \dots, w_{i,|d_i|}^d]$ defines the i-th document as a list of words,
- $S(d_i, j, \sigma)$ is a sliding window defined for the i-th document, starting at the j-th position and including σ words,
- $T = \{t_1, t_2, \dots, t_{|T|}\}$ is a set of identified topics,
- $V_k = [v_{k,1}, v_{k,2}, \dots, v_{k,N}]$ defines a list of N words defining the k-th topic.

Next, the matrix of association coefficients between words defining every topic should be created. For topic k the matrix $\mathbf{Q}_k$ has a form:

$$\mathbf{Q}_k = \begin{bmatrix} q_{1,1}^k & \dots & q_{1,N}^k \\ \dots & \dots & \dots \\ q_{N,1}^k & \dots & q_{N,N}^k \end{bmatrix} \tag{6}$$

where:

$$q_{x,y}^k = nmpi(v_{k,x}, v_{k,y}) \tag{7}$$

For topic k, normalized PMI values are calculated using probabilities of occurrence of words $v_{k,x}$ and $v_{k,y}$ inside the sliding window $S(d_i, j, \sigma)$ moving through all documents in $\boldsymbol{D}$. It may be expressed as:

$$nmpi\left(v_{k,x}, v_{k,y}\right) = \frac{\log\frac{P\left(v_{k,x}, v_{k,y}\right)+\varepsilon}{P\left(v_{k,x}\right)P\left(v_{k,y}\right)}}{-\log\left(P\left(v_{k,x}, v_{k,y}\right)+\varepsilon\right)} \tag{8}$$

where $P\left(v_{k,x}, v_{k,y}\right)$ is defined as:

$$P\left(v_{k,x}, v_{k,y}\right) = \frac{\sum_{a=1}^{|D|}\sum_{b=1}^{|d_a|-\sigma+1} g\left(a,b,\sigma,v_{k,x},v_{k,y}\right)}{\sum_{a=1}^{|D|}(|d_a|-\sigma+1)} \tag{9}$$

where:

$$g\left(a, b, \sigma, v_{k,x}, v_{k,y}\right) = \begin{cases} 1 & if\ v_{k,x}, v_{k,y} \in S(d_a, b, \sigma) \\ 0 & otherwise \end{cases} \tag{10}$$

and $P\left(v_{k,x}\right)$ is calculated using the formula:

$$P\left(v_{k,x}\right) = \frac{\sum_{a=1}^{|D|}\sum_{b=1}^{|d_a|-\sigma+1} h\left(a,b,\sigma,v_{k,x}\right)}{\sum_{a=1}^{|D|}(|d_a|-\sigma+1)} \tag{11}$$

where:

$$h\left(a, b, \sigma, v_{k,x}\right) = \begin{cases} 1 & if\ v_{k,x} \in S(d_a, b, \sigma) \\ 0 & otherwise \end{cases} \tag{12}$$

Next, for every topic, vector $\boldsymbol{M}_k$ is calculated:

$$\boldsymbol{M}_k = \left[m_{k,1}, m_{k,2}, \ldots, m_{k,N}\right] = \left[\sum_{j=1}^{N} q_{j,1}^{k}, \sum_{j=1}^{N} q_{j,2}^{k}, \ldots, \sum_{j=1}^{N} q_{j,N}^{k}\right] \tag{13}$$

Elements of $\boldsymbol{M}_k$ are calculated as sums of values located in subsequent columns of $\mathbf{Q}_k$. $\boldsymbol{M}_k$ may be treated as a k-th topic representation.

To calculate the coherence measure for a given set of topic, the $\boldsymbol{C}$ matrix is first calculated.

$$\boldsymbol{C}_{[|T|\times N]} = \begin{bmatrix} sim(\boldsymbol{M}_1, \boldsymbol{q}_1^1) & \ldots & sim(\boldsymbol{M}_1, \boldsymbol{q}_N^1) \\ \ldots & \ldots & \ldots \\ sim\left(\boldsymbol{M}_{|T|}, \boldsymbol{q}_1^{|T|}\right) & \ldots & sim\left(\boldsymbol{M}_{|T|}, \boldsymbol{q}_N^{|T|}\right) \end{bmatrix} \tag{14}$$

where symbols $\boldsymbol{q}_j^k$ represent the j-th row of the $\mathbf{Q}_k$ matrix and $sim(.)$ is a cosine similarity between vectors.

Finally, the coherence measure C_V is calculated as an arithmetic average of elements of the $\boldsymbol{C}$ matrix.

Values of the C_V coefficient belong to the range [0; 1]. Higher values indicate higher consistency of topics. When deciding on the number of topics, it is advisable to maximize this indicator.

5. The analysis of Polish research publications in the fields of economics and management

5.1. The scope of the analysis

The dataset included titles and abstracts of research publications published in the period 2000–2024, with at least one Polish author, registered by the Scopus database and assigned to *BUSI* (business), *ECON* (economics) or *DECI* (decision science) areas. The total number of publications which met the above conditions was 36445, but the analysis covered 35626 publications that had a title and an abstract in English.

Basic quantitative indicators describing the whole set of Polish publication achievements (36445 works) are presented in Figure 1.

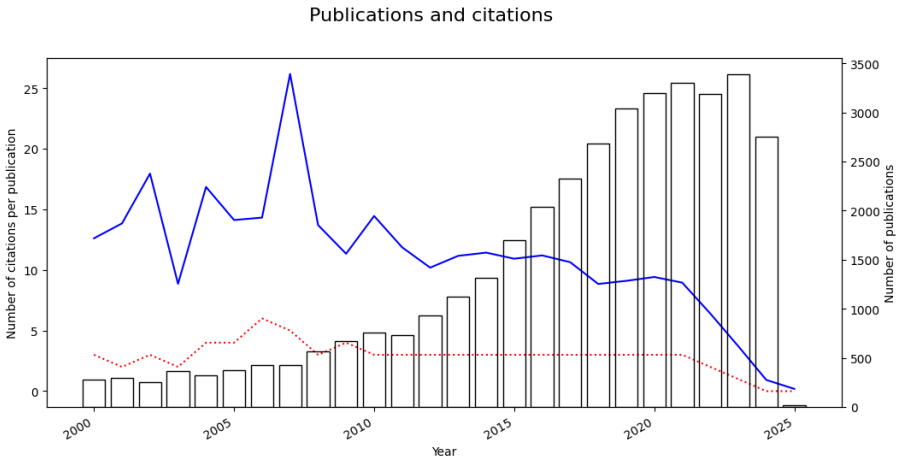

Figure 1. Number of publications (bar plot, right axis), number of citations per publication (average value – blue solid line, left axis; median value – red dotted line, left axis)

Source: own work based on Scopus database.

A rapidly increasing number of publications can be seen by 2021. The number of published papers seems to stabilize in the following years. In contrast, the number of citations per published paper has been decreasing over the past 15 years.

5.2. BERTopic model building and interpretation

The BERTopic model was used for the analysis of titles and abstracts of Polish publications. First, documents were split into tokens with a form of sentences. All tokens with 28 or less letters were removed (these tokens most often contained names of publishing houses or names of affiliated institutions). Finally, in the analysis 283576 sentences were used.

Next, embeddings for sentences were calculated with the use of the SBERT model.

Several BERTopic models were tested and finally the model with 9 topics was chosen. This decision was taken on the basis of the C_V coherence, which, depending on the number of topics, took values shown in Table 1.

Table 1. Values of the C_V coherence for models with different number of topics

Number of topics	C_V
6	0.4228
7	0.4338
8	0.4438
9	0.4823
10	0.4643

Source: own work.

Table 2 shows the number of tokens (sentences) assigned to every topic.

Table 2. Number of sentences assigned to every topic

Topic ID	*Topic -1*	*Topic 0*	*Topic 1*	*Topic 2*	*Topic 3*
Count	133358	108453	18851	13650	3351
Topic ID	***Topic 4***	***Topic 5***	***Topic 6***	***Topic 7***	
Count	2432	1689	1656	136	

Source: own work.

Topic -1 represents all sentences which have been identified as noise and are not related to any of the recognized topics.

In order to interpret each topic, lists of the words most closely related to each topic have been created (Figure 2).

Figure 2. The most important words for identified topics

Source: own work.

Topic 0 covers issues related to economic development and management methods and their implementation in Poland. The key issue addressed under *Topic 1* is decision support methods. *Topic 2* represents issues specific to commodity science. Issues specific to natural environment and regional development are discussed within *Topic 3*. Subjects related to image processing are discussed under *Topic 4*. Health care issues are related to *Topic 5*. Biology and genetics issues are related to *Topic 6*. *Topics 7* is related to mathematics, in particular to geometry.

Issues specific to identified topics are in many cases related to each other. A visualization of the similarity matrix is shown in Figure 3.

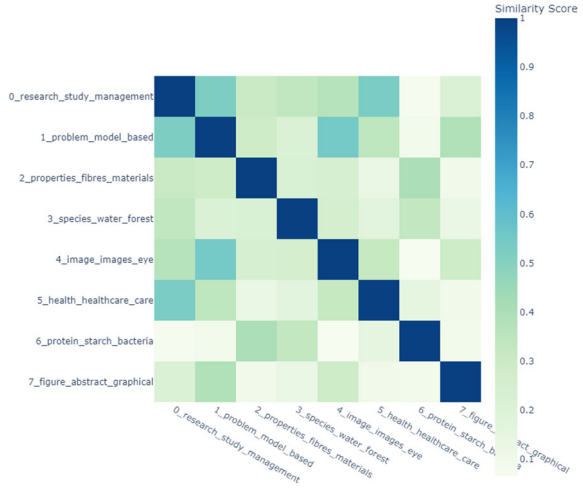

Figure 3. Visualization of the similarity matrix between topics

Source: own work.

A useful tool for analyzing relationships between topics can also be an intertopic distance map presented in Figure 4.

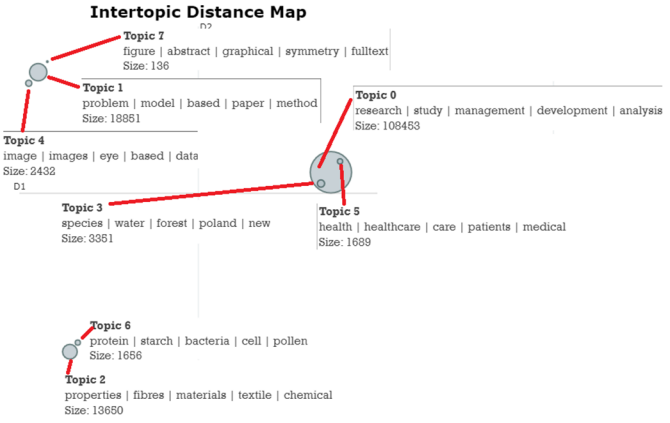

Figure 4. Intertopic distance map for identified topics

Source: own work.

An analysis of Figure 4 indicates that three groups of themes can be identified:

- Group 1: Topic 1, Topic 4, Topic 7.
- Group 2: Topic 0, Topic 3, Topic 5.
- Group 3: Topic 2, Topic 6.

In the next step of the research, a sentence-topic matrix was estimated to determine the importance of every topic in every single sentence. Next, information about topic contribution to every sentence was aggregated at the level of every document. The aggregation was done by calculating geometric average for values relating to sentences that formed a given document. Then, using the same approach, an aggregation of the importance of each topic was carried out for each year included in the scope of analysis. The results are presented in Figure 5.

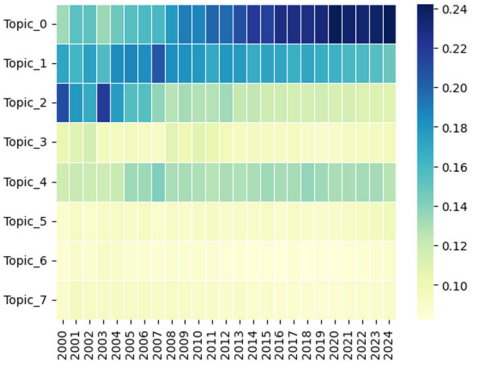

Figure 5. Changes in topics' importance in Polish research works over years

Source: own work.

Analyzing the data presented in Figure 5, it is worth noting that the values presenting the importance of each topic in consecutive years are relative (they add up to unity for each year). At the beginning of the current century, the greatest publication achievements were related to decision support systems and commodity science. Since the second decade of the 21st century, economic development and management issues have played a key role in the publication output. In contrast, the importance of commodity science and decision support systems has been declining. Interest in image processing methods, which fall under the umbrella of multivariate analysis, is also noticeable. The importance of the other topics was rather low.

6. Conclusions

The research carried out allows for formulation of the following conclusions:

1. In quantitative terms, the Polish publication achievements in the field of economics and management has increased significantly since the beginning of this century, although the number of publications stabilized in the last few years. The growth potential seems to be exhausted.
2. The quality of the analyzed publication achievements, measured by the number of citations, has not shown any positive change for the last 15 years.
3. Main topics discussed in Polish publications included: economic development, management methods, decision support solutions, commodity science issues, natural environment and regional development, health care system, biology and genetics and mathematics.
4. Topics related to economic development and management issues gained the most importance in the last two decades.
5. Decision support systems and commodity science issues have lost their relevance.
6. The importance of the quantitative approach remains noticeable and unchanged.
7. The remaining topics have relatively small significance.
8. The use of the BERTopic model has made it possible to analyze large text datasets and aggregate the results.
9. Further research on BERTopic and other topic modelling methods should be considered as necessary.

References

Blei, D., Ng, A. and Jordan, M., (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (3), pp. 993–1022.

Campello Ricardo J. G. B. and Moulavi, D. and S. J., (2013). Density-Based Clustering Based on Hierarchical Density Estimates, in V.S. and C.L. and M.H. and X.G. Pei Jian and Tseng (ed.) *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 160–172.

Deerwester, S. *et al.*, (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6), pp. 391–407.

Devlin, J. *et al.*, (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://arxiv.org/abs/1810.04805.

Firth, J. R., (1962). A synopsis of linguistic theory, 1930–1955, in *Studies in Linguistic Analysis*. Oxford: Blackwell.

Grootendorst, M., (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* [Preprint].

Hofmann, T., (1999). *Probabilistic Latent Semantic Indexing*. New York: ACM.

Lee, D. D., Seung, H. S., (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401, pp. 788–791.

McInnes, L., Healy, J. and Melville, J., (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Available at: https://arxiv.org/abs/1802.03426.

Prim, R. C., (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6), pp. 1389–1401. Available at: https://doi.org/10.1002/j.1538-7305.1957.tb01515.x.

Reimers, N., Gurevych, I., (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Available at: https://arxiv.org/abs/1908.10084.

Rijcken, E., (2023). *CV Topic Coherence Explained. Understanding the metric that correlates the highest with humans.*

Sparck Jones, K., (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1), pp. 11–21.

Vaswani, A. *et al.*, (2017). Attention is all you need, in *Advances in Neural Information Processing Systems*.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 169–186, https://doi.org/10.59139/stattrans-2026-010
Received – 09.05.2025; accepted – 17.12.2025

The role of education and gender in shaping career paths of Polish millennials: a shared frailty survival model analysis

Wioletta Grzenda[1], **Agnieszka Marszałek**[2]

Abstract

Our study aims to examine the influence of gender and the level of education on job mobility among young employees, using the Polish labor market as an example. When analyzing job changes, we go beyond previous studies by considering the duration of individual job episodes and the time-varying nature of some characteristics in young people, such as the level of education or the marital status. Our analysis was based on survival analysis methods, including frailty models. Using data from the Generation and Gender Survey, we found that the impact of the examined factors on job mobility varied by gender. We observed that the influence of having a child on job mobility was significant only for women. Mothers had a lower risk of job changes than childless women. The stabilization of men's careers takes place over time and is associated with leaving the family home and marriage. Moreover, having higher education has a greater impact on the risk of job changes for men than for women.

Keywords: education, gender, job mobility, survival analysis.

1. Introduction

Millennials, known as the Y generation, consider work less significant in their lives, prioritize leisure to a greater extent, and exhibit a weaker work ethic compared to individuals from the Baby Boomers and Generation X (Twenge, 2010). Moreover, Millennials are perceived as people who are motivated by higher pay, quickly become dissatisfied and leave their jobs (AbouAssi et al., 2021). Simultaneously, young employees expect job stability as much or even more than their counterparts from the Baby Boomer and Generation X generations at the same age (Twenge, 2010). Larasati and Aryanto (2020) point to Generation Y as a generation that, despite many advantages,

[1] Collegium of Economic Analysis, Institute of Statistics and Demography, SGH Warsaw School of Economics, Warsaw, Poland. E-mail: wgrzend@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-2226-4563.

[2] Independent researcher, Poland. E-mail: marszalek.agnieszka@outlook.com. ORCID: https://orcid.org/0000-0003-4906-6484.

such as self-confidence, independence and social activity, has a poor reputation as job-hoppers. Job-hopping refers to an employee's frequent and voluntary inter-organizational transitions, not necessarily related to a change in the nature of work itself (Steenackers and Guerry, 2016; Lake et al., 2018). This phenomenon was initially referred to as the 'hobo syndrome' in the 1970s and was explained as the tendency for an employee to migrate between organizations, driven not necessarily by rational motives but rather by a sudden urge for change (Ghiselli, 1974). The purpose of such behavior is to find the best job to meet some subjective criteria. One of them may be a willingness to increase earnings that can be obtained by offering one's work experience to another employer. The Redmond and McGuinness (2019) study results confirm that previous status employment influences future employee wage increases. However, the impact of job mobility on wage growth depends on gender and education (Pearlman, 2018).

Analyses of young people's behavior in the labor market indicate a negative correlation between a person's age and their propensity to switch employers. At the beginning of their careers, young people are more likely to switch jobs than their older colleagues (Steenackers and Guerry, 2016). Moreover, young women tend to job-hop significantly more than young men. Also, Larasati and Aryanto (2020), based on a literature review, conclude that young women change jobs more often than young men. The objective of our study is to examine the influence of gender and education on job mobility among people from Generation Y. Furthermore, as we analyze the factors influencing gender-related job changes, we investigate the causes of excessive job mobility and discuss the consequences of job-hopping.

In our study, we focus on Poland as an example of a country where unemployment among young people is particularly low compared to other European countries (Eurostat, 2023). However, traditional gender-based social roles in this country are still considered important (Kasprzak, 2023). Moreover, research shows that professional and family careers are interdependent (Landmesser, 2013; Grzenda, 2019). We used data from the first and second waves of the Polish Generations and Gender Survey (GGS), which were conducted in 2010–2011 and 2014–2015, respectively. While the realm of Millennials' behavior in the job market has been thoroughly explored, there are still some gaps that lead us to the following research questions: (Q1) What are the differences in the impact of factors determining job mobility based on gender? (Q2) What impact does education have on the risk of job changes?

Our contribution to the literature is twofold. First, we aim to go beyond previous studies by identifying differences by gender in the impact of factors such as education, age, and having a child on the risk of job mobility. Second, we make full use of the

longitudinal approach, taking into account in the analysis not only job changes but also the duration of each job episode as well as changes over time in the values of other characteristics, such as education or marital status. Thus, the results of our study contribute to research on the importance of the role of gender and education in the employment decisions of young people and on the factors that predispose individuals to follow a specific career path.

2. Review of the literature

2.1. Labor turnover and job mobility

The primary driver for job changes among Millennials is the pursuit of job satisfaction (Campione 2015; Hassan et al., 2020). However, as highlighted by Campione (2015), the factors that push them away tend to carry more weight than the positive factors that draw them in. One of the factors influencing the retention of Millennials in a company is pay. Redmond and McGuinness (2019) show that individuals who have worked for another employer before taking up their current position are more likely to receive pay raises than people who were previously unemployed. However, too frequent job changes do not necessarily yield positive outcomes. Yankow (2022) found that individuals who exhibit moderate job changes within the first 2 years of entering the labor market but subsequently reduce their mobility actually achieve higher wages compared to both those who remain in the same job and those who consistently change jobs. Generation Y's inclination for frequent job changes in pursuit of fulfilling work challenges employers in retaining skilled labor and coping with high turnover within this generation (Hassan et al., 2020). According to the human capital theory, the departure of an experienced or skilled worker may result in a decrease in future productivity (Becker, 1964). It is claimed that apart from the loss of tacit knowledge and experience, employee turnover is also associated with excessive costs related to HR administration and recruitment and training costs of new workers (Huang and Zhang, 2016). The consequences of losing an employee, especially management staff, are so great that companies, after losing executives to other companies, significantly increase their incumbent executives' pay (Gao et al., 2015). Furthermore, the departure of one employee might have a negative impact on the job satisfaction and productivity of other employees who stay within the organization (Steenackers and Guerry, 2016). Companies wishing to retain Millennials in the organization should focus primarily on work-life balance issues, flexible time, and paid leave, and avoid extreme hours and irregular hours schedules worked (Twenge, 2010; Campione, 2015).

2.2. Job mobility and gender

The research results on the behavior of young people in the labor market do not indicate clear conclusions regarding the tendency toward job mobility by gender. There has been a long-standing debate in the literature about the gender differences in employment and wage (Wootton, 1997; Pedulla 2016; Blau and Kahn, 2017; Reichelt et al., 2021; Zamarro and Prados, 2021) as well as the relationships between paid work and motherhood (Boeckmann et al., 2015; Zhou, 2017; Cabello-Hutt, 2020; Cukrowska-Torzewska and Matysiak 2020; Schmitt, 2021). Based on the above literature, it can be concluded that gender inequality in the labor market is a consequence of various factors and does not necessarily reflect the biological roles fulfilled by women. According to Boeckmann et al. (2015), maternal employment is shaped by institutional and cultural contexts, which make men less involved in caring for small children than women. Looze (2017) found that preschool-age children largely immobilize white American women, as they discourage these women from making types of voluntary job changes. On the other hand, in the initial stages of their careers, women are more likely than men to change employers (Steenackers and Guerry, 2016). This is related to the search for a rewarding and stable job that will allow for childcare after starting a family. Similar conclusions are provided by the results of earlier research by Matysiak (2009) on fertility and female employment in Poland. It was found that young Polish women, before starting a family, are highly active in their search for a stable position in the labor market that would enable them to pursue their professional lives and have children (Matysiak, 2009). Also, Kaufman and White (2015), when examining gender differences among Swedish workers, showed that having secure employment is more important for women than for men. The lower willingness of women to quit their jobs has also been confirmed by Moynihan and Landuyt (2008) when examining state government jobs. Disparities in career path patterns and tendencies to change employers based on gender and parental status influence one's professional career trajectory, and consequently, disparities in current and future wages. Reshid (2019) found that although men and women change jobs and occupations simultaneously, women receive a significantly lower wage return on mobility than men. Moreover, differences in women's professional mobility, particularly concerning their maternal status, lead to disparities in their earnings and negatively impact their future professional careers (Looze, 2017).

2.3. Job mobility and education

The individual labor market behavior is significantly influenced by the acquired human capital, and one of its main indicators is education. The Millennial Generation is reporting higher levels of educational attainment than earlier generations (Ng and

Johnson, 2015). However, the research results on the relationship between education level and job changes are not fully consistent. Grosemans et al. (2020) primarily focused their research on the transition from higher education to the workforce and concluded that increased occupational mobility is observed during this period. However, they emphasize the importance of distinguishing between deliberate exploration and floundering. Based on the research by Ignaczak et al. (2022), it can be concluded that higher education affects professional careers in two different ways. Well-educated employees are in demand by companies, which makes it easier for such people to find a job that meets their expectations and, at the same time, lowers the risk of future dismissal. On the other hand, higher demand in the labor market makes such people more confident when deciding to change employers, because the action is less risky. Thus, workers with a college degree have a higher tendency to job-hopping than individuals with a relatively low education level (Ignaczak et al., 2022). Also, Ng and Johnson (2015) reported that the increased level of education, notably in the field of graduate management education, among Millennials, is instrumental in enhancing their capacity for career mobility, with a particular emphasis on transitioning between sectors. In contrast, Steenackers and Guerry (2016) in their study of the Belgian labor market state that the level of education has no impact on the job-hopping behavior of an employee and having more job alternatives is not always connected to an increased tendency of job switching.

3. Data

To model the employment trajectories of young individuals in the Polish labor market, we used data from Wave 1 and Wave 2 of the Generations and Gender Survey Poland (GGS-PL). The GGS-PL survey is part of an international research the Generations and Gender Programme (GGP) designed to obtain information on demographic processes with consideration of the economic, social, and cultural context. In our analysis, we included respondents who, at the time of the second survey, were aged between 18 and 29 and had previously undertaken at least one job in the private or public sector (excluding self-employment). The upper age limit of the respondents was based on Arnett's (1998; 2006) findings, in which he states that young people reach full social maturity around the age of 30. Given the assumptions adopted in our study, the total number of participants was 543, with 49.63% being women (270) and 50.37% men (273). Employment history was reconstructed based on the first and second waves of the GGS survey. The statistics on respondents' number of jobs are presented in Table 1. It was found that the maximum number of job changes by respondents was 7. Furthermore, more women than men had only one job. Biemann et al. (2012) indicated that a career path is significantly influenced not only by gender, but

also by age, marital status, having a child, education, and employment sector. In our study, we also consider these characteristics to see to what extent they are related to the job changes of Polish millennials.

Table 1. Number of respondent's jobs undertaken until Wave 2 of the GGS-PL by gender

Number of jobs	Number of respondents			Per cent		
	Women	Man	Total	Women	Man	Total
1	156	145	301	57.78	53.11	55.43
2	61	70	131	22.59	25.64	24.13
3	36	34	70	13.33	12.45	12.89
4	10	11	21	3.70	4.03	3.87
5	3	6	9	1.11	2.20	1.66
6	2	7	9	0.74	2.57	1.66
7	2	0	2	0.74	0.00	0.36

Source: own calculations; data from Generations and Gender Survey Poland.

The dependent variable, which in survival analysis is the time to failure, for each respondent and each his/her job was defined as the number of months from the start of this job to its termination in the case of employment termination. In the case of people who had not terminated their employment relationship with their last employer at the time of the second wave of the study, the time was counted until Wave 2 of the GGS. In addition, a censoring variable was created and assigned a value of 1 if the event occurred, that is, if the respondent terminated the employment relationship, and 0 otherwise.

In the next stage of the research, we verified which of the considered demographic, socio-economic, and work-related characteristics changed over time. It was obtained that only attributes such as the respondent's sex, place of residence in childhood, and father's and mother's education level are constant in time. The remaining characteristics, such as the respondent's age group, education level, marital status, sector of employment, and information on whether the respondent has at least one child and whether he or she has ever lived without parents, are time-varying. In addition, the new variable describing the age group was created by categorizing the respondent based on his/her date of birth. The first category (18-24 years) is the age range into which adolescents are classified, while the second group (25-29 years) is those in the so-called emerging adulthood stage (Arnett, 2006). The set of potential independent variables selected for modelling is presented in Tables 2 and 3.

Table 2. Sample characteristics by gender – variables constant in time

Variable	Categories	Per cent		
		Women	Men	Total
Place of residence in childhood	A city of 100,000 or more residents	24.07	27.01	25.55
	A city of under 100,000 residents	37.78	35.40	36.58
	Rural areas	38.15	37.59	37.87
Father's education level	Basic vocational, lower secondary, primary, incomplete primary	70.00	67.15	68.57
	Higher, postsecondary and vocational secondary, general secondary	30.00	32.85	31.43
Mother's education level	Basic vocational, lower secondary, primary, incomplete primary	53.70	54.38	54.04
	Higher, postsecondary and vocational secondary, general secondary	46.30	45.62	45.96

Source: own calculations; data from Generations and Gender Survey Poland.

Table 3. Sample characteristics by gender – time-varying variables

Variable	Categories	Per cent					
		At the beginning			At the end		
		Women	Men	Total	Women	Men	Total
Age group	18-24 years	99.26	99.27	99.26	48.89	41.97	45.40
	25-29 years	0.74	0.73	0.74	51.11	58.03	54.60
Education level	Lower secondary, primary, incomplete primary	17.78	23.08	20.44	4.44	9.12	6.80
	Basic vocational	6.67	21.61	14.18	8.89	25.18	17.10
	General secondary	28.52	23.81	26.15	14.81	14.23	14.52
	The first stage of tertiary, postsecondary and vocational secondary	34.07	25.64	29.83	44.44	37.59	40.99
	The second stage of tertiary and higher	12.96	5.86	9.39	27.41	13.87	20.59
Is married	No	100.00	99.63	99.82	70.37	81.02	75.74
	Yes	0.00	0.37	0.18	29.63	18.98	24.26
Has at least one child	No	100.00	99.63	99.82	71.11	77.74	74.45
	Yes	0.00	0.37	0.18	28.89	22.26	25.55
Has ever lived without parents	No	24.81	49.08	37.02	24.81	49.27	37.13
	Yes	75.19	50.92	62.98	75.19	50.73	62.87
Sector of employment	Public	24.44	15.75	20.07	19.26	16.79	18.01
	Private	75.56	84.25	79.93	80.74	83.21	81.99

Source: own calculations; data from Generations and Gender Survey Poland.

4. Methods

In our study, we used generalization of the Cox proportional hazard model. For each individual i, let $\mathbf{x} = [\, x_1, \ldots, x_k]^T$ denote the vector of independent variables, and $\boldsymbol{\beta} = [\, \beta_1, \ldots, \beta_k]$ denote the vector of regression coefficients, then the hazard function for the model of proportional hazards takes the form:

$$h_i(t) = h_0(t) \exp(\mathbf{x_i}\boldsymbol{\beta}), \tag{1}$$

where $h_0(t)$ denotes baseline hazard. Given the non-parametric form of the baseline hazard, the partial likelihood method is used to estimate model parameters (Cox, 1972; Cox, 1975, Cox and Oakes, 1984). This form of the Cox model is dedicated to estimating the occurrence of a single event. In the case of multiple events, it is required to adjust the formula. In 1982 Andersen and Gill proposed a generalized version of the Cox proportional hazards model dedicated to recurring event data called the Andersen-Gill model or intensity model (Andersen and Gill, 1982). This model relates the event recurrence intensity function to the covariates in a multiplicative manner. The method uses a counting process approach, treating each individual as a process of counting multiple events with essentially independent increments. The model assumption is that the risk of an event occurrence in time t does not change regardless of whether past events have occurred or not, which implies the independence of recurring events.

Let $h_{ik}(t)$ represent the hazard function of the k-th event for i-th individual at time t and $h_0(t)$ represent the common baseline hazard for all events/individuals. The hazard function for the Andersen-Gill model then takes the form:

$$h_{ik}(t) = h_0(t) \exp(\mathbf{x_i}\boldsymbol{\beta}). \tag{2}$$

When the assumption of recurring events independence is met, it is possible to estimate the risk of all events using the event times of each observed event. The Andersen-Gill model aims to estimate the same quantity as the Cox proportional hazard model, but the estimation is based on more information because the person who experienced the event remains at risk of subsequent events. Consequently, the corresponding partial probability is based on a larger number of events and a modified set of risks. If this assumption is not met, the Andersen-Gill model is still applicable, but it requires some modification known as the proportional means model (Lin et al., 2000).

The other option is to extend the Andersen-Gill model into the frailty model by adding random effects to it, which would allow us to consider the unobservable heterogeneity of individuals. The model created in such a way is called the shared frailty model. It is assumed that for each individual, there is more than one observation within each cluster, and all the observations within the cluster share the same level of frailty. Using the frailty term makes it possible to correct some or all of the errors in the coefficients caused by unobserved heterogeneity. The model is estimated with the use

of the penalized partial likelihood method (Ripatti, Palmgren, 2000); thus, the parameter estimates for the fixed effects obtained from the shared frailty model differ from the proportional means model (Allison, 2010). The hazard function of the k-th event for i-th individual (in the i-th cluster) takes the form:

$$h_{ik}(t) = h_0(t)\exp(\mathbf{x_i}\boldsymbol{\beta}) + \gamma_i, \tag{3}$$

where γ_i is a random effect for the i-th individual. The random components are assumed to be independent and distributed identically.

The generalized version of the frailty model is a model that includes both time-independent and time-dependent covariates:

$$h_{ik}(t) = h_0(t)\exp(\mathbf{x_i}\boldsymbol{\beta} + \mathbf{z_i}(t)\boldsymbol{\delta}) + \gamma_i, \tag{4}$$

where $\mathbf{x_i}$ is a vector of time-independent covariates, $\boldsymbol{\beta}$ is a coefficient vector for time-independent covariates, $\mathbf{z_i}$ denotes a vector of time-dependent covariates, and $\boldsymbol{\delta}$ is a coefficient vector for time-dependent covariates.

Given that the shared frailty model is estimated using the penalized partial likelihood method, it is recommended to use a suitably modified test when fitting the model. One of the recommended methods is to use the Wald test with generalized degrees of freedom (Gray, 1992; Therneau and Grambsch, 2000), which was also used in this work.

5. Results

Given that the results of previous research indicate the existence of some differences in labor market behavior patterns between men and women (Moynihan and Landuyt, 2008; Matysiak, 2009; Kaufman and White, 2015; Steenackers and Guerry, 2016), we built three models: a general model for all respondents and two additional models by gender.

Based on the assumption that there may be notable differences between individuals in their behavior in the labor market, in the first stage of the research, we assessed the statistical significance of the random effect to verify if the shared frailty model is justified for our study. It was obtained that the random effect is statistically significant for each of the three models.

Based on the preliminary data analysis, it was determined that most of the respondents' characteristics vary over time. Therefore, to construct the model, the waiting time for the occurrence of an event for each respondent was divided into subintervals, ensuring that all examined features remain constant within these designated subperiods.

Following the analysis of the available data and the Wald test, the set of covariates for the main model was established. The study aimed to investigate gender differences in labor market behavior; therefore, an identical specification of explanatory variables was applied in the models for women and men. The results obtained from the shared frailty are presented in Table 4 (all respondents) and Table 5 (women and men separately).

Table 4. Estimated parameters with standard error, p-value and hazard ratio – general model

Covariate	Parameter estimate	Standard error	*p*-value	Hazard ratio
Age (*ref. 18–24 years*)				
25-29 years	-0.578	0.363	0.112	0.561
Education level (*ref. Lower secondary, primary, incomplete primary*)				
Basic vocational	-0.104	0.212	0.624	0.901
General secondary	0.006	0.194	0.976	1.006
The first stage of tertiary, postsecondary and vocational secondary	0.321	0.189	0.090	1.379
The second stage of tertiary and higher	1.205	0.233	<.001	3.337
Is married (*ref. No*)				
Yes	-0.470	0.222	0.034	0.625
Has at least one child (*ref. No*)				
Yes	-0.380	0.256	0.137	0.684
Has ever lived without parents (*ref. No*)				
Yes	-0.155	0.122	0.204	0.856
Sector of employment (*ref. Public*)				
Private	-0.271	0.124	0.029	0.763
Sex *(ref. Man)*				
Woman	0.059	0.119	0.623	1.060

Source: own calculations; data from Generations and Gender Survey Poland.

The first shared frailty model included all respondents. We found that older respondents have a lower risk of job mobility. People aged 25 to 29 had a 43.9% lower hazard of job mobility than people aged 18 to 25. However, based on the obtained p-value, it cannot be concluded that this characteristic is statistically significant. Analyzing the variable describing the respondents' education, we found that this factor influences the risk of job mobility. Respondents with a first stage of tertiary, postsecondary, and vocational secondary level of education had a 37.9% greater hazard of job mobility compared to lower secondary, primary, and incomplete primary education. In contrast, respondents with at least a second stage of tertiary education had more than 3 times higher hazard of job mobility than the least educated respondents. Furthermore, married people had a 37.5% lower hazard of job mobility than other respondents. In addition, the results show that the sector of employment is a statistically significant factor too. Young people working in the private sector had

a 23.7% lower hazard of job mobility in comparison to people working in the public sector. In the case of the first shared frailty model, characteristics such as having at least one child or ever living without parents proved to be statistically insignificant. Presented interpretations remain valid under the ceteris paribus assumption.

In the subsequent research stage, two models were constructed - one including only women and the other including only men, to better understand gender differences in the behavior of young individuals in the labor market. Comparing the results of these two models, it can be concluded that age was an important factor only in the case of men. Male respondents in the older age group (25-29 years) had a 65.4% lower hazard of job mobility compared to the younger group.

Table 5. Estimated parameters with standard error, p-value and hazard ratio – model for women and model for men

Covariate	Women				Men			
	Parameter estimate	**Standard error**	***p*-value**	**Hazard ratio**	**Parameter estimate**	**Standard error**	***p*-value**	**Hazard ratio**
Age (*ref. 18–24 years*)								
25-29 years	-0.120	0.465	0.797	0.887	-1.060	0.609	0.082	0.346
Education level (*ref. Lower secondary, primary, incomplete primary*)								
Basic vocational	0.122	0.366	0.739	1.129	-0.047	0.286	0.869	0.954
General secondary	-0.206	0.282	0.464	0.814	0.105	0.293	0.719	1.111
The first stage of tertiary, postsecondary and vocational secondary	0.050	0.287	0.863	1.051	0.572	0.273	0.037	1.771
The second stage of tertiary and higher	0.756	0.327	0.021	2.130	1.953	0.388	<.001	7.047
Is married (*ref. No*)								
Yes	-0.336	0.273	0.219	0.714	-0.632	0.400	0.115	0.532
Has at least one child (*ref. No*)								
Yes	-0.672	0.406	0.098	0.511	-0.123	0.352	0.726	0.884
Has ever lived without parents (*ref. No*)								
Yes	-0.141	0.190	0.458	0.869	-0.230	0.178	0.197	0.794
Sector of employment (*ref. Public*)								
Private	-0.365	0.167	0.029	0.694	-0.143	0.200	0.474	0.866

Source: own calculations; data from Generations and Gender Survey Poland.

The results also indicate that the level of education was a factor that more strongly differentiated the employment of men than that of women. Both the first stage of tertiary, postsecondary and vocational secondary and the second stage of tertiary and higher education levels had a positive effect on the risk of job mobility. Male respondents with the first stage of tertiary, postsecondary and vocational secondary education had a 77.1% higher hazard of terminating their jobs compared to the least educated group, and respondents with the highest education level had more than 7 times higher hazard of job mobility compared to the least educated group of male respondents. In the case of female respondents, it was revealed that only women with at least a second stage of tertiary or higher education had statistically significant, more than 2 times higher hazard of job mobility compared to the least educated female respondents.

Other statistically significant factors influencing the risk of job mobility in the case of female respondents were having a child, as well as the employment sector. Women who had at least one child had a 48.9% lower hazard of job mobility compared to women without children. Females working in the private sector had job mobility hazard lower by 30.6% compared to those working in the public sector. These factors were statistically insignificant in the case of male respondents.

Moreover, based on the results of the Wald test with generalized degrees of freedom, it can be concluded that in the case of men, the marital status and the history of living accommodation also influenced the risk of job mobility. Men who were married had a 46.8% lower hazard of job mobility compared to those unmarried. If the male respondent had ever lived without parents, his hazard of job mobility was lower by 20.6% compared to those who had lived with their parents all their lives. All interpretations remain valid under the ceteris paribus assumption.

6. Discussion and Conclusions

Our study focuses on Generation Y, which has a significantly different approach to employment than the earlier Generation X (Twenge, 2010). People from Generation Y often live in a hurry and focus on their development, which makes them less loyal to their employers (Robak, 2017). Millennials exhibit a higher propensity for changing jobs and employers more frequently than their predecessors, and they also display a greater readiness to embrace career shifts that may not necessarily involve upward mobility (Lyons et al., 2012). Our study aimed to examine the influence of gender (Q1) and education (Q2) on job mobility among young individuals. We revealed that, among Polish Millennials, gender did not influence the risk of job change, whereas it did play a significant role in determining the impact of other factors on job mobility, including education level.

Gender disparities in the labor market often stem from traditionally held social roles for women, with motherhood being a key aspect (Kaufman and White, 2015; Steenackers and Guerry, 2016; Cukrowska-Torzewska and Matysiak, 2020). This is confirmed by the results of our research, which indicate that the impact of having a child on job mobility was significant only for women. Furthermore, women with at least one child had a lower risk of job changes compared to childless women. The recognition of incongruity between women's careers and their duties as mothers, as well as the consequent adjustments they make, influences women's gender role perceptions as they transition into motherhood. This is reflected in women's different attitudes towards job mobility both before and after the birth of a child (Zhou, 2017). Furthermore, Bass (2015) demonstrates that gendered expectations related to parenthood may play a significant role in perpetuating patterns of labor market inequality, even before the practical constraints of parenthood come into play. Based on our findings, it can be concluded that, in the case of men, having a child does not directly impact professional mobility. However, the stabilization of men's careers occurs with age and is associated with leaving the family home as well as marriage. In the case of women, these factors had no impact on professional mobility. In conclusion, we agree with Boeckmann et al. (2015) that women's employment patterns are determined more by motherhood than gender. Moreover, we show that this finding also applies to career mobility.

While women are, on average, better educated than men (Cukrowska-Torzewska and Lovasz, 2016), their employment situation is not necessarily more favorable. We found that the level of education mattered for job mobility for Polish Millennials, but this factor shaped the labor market behavior of men more strongly than that of women. For women, only having at least a second stage of tertiary education statistically significantly reduced the risk of job changes compared to women with the lowest level of education, while for men this risk was also reduced by having a first stage of tertiary, postsecondary or vocational secondary education, likewise versus the least educated men. Nevertheless, our findings are in line with previous research, indicating that individuals with higher education exhibit a stronger inclination toward professional mobility compared to those with relatively lower levels of education (Ignaczak et al., 2022; Ng and Johnson, 2015).

Considering the employment sector, AbouAssi et al. (2021) note that American youth tend to change jobs frequently, but only within a given sector, not across sectors. Moreover, the strongest predictor of public sector employees changing jobs within the sector is job dissatisfaction. In the case of Polish Millennials, the employment sector was significant only for women, with the risk of changing jobs being lower for women employed in the private sector. It can, therefore, be concluded that the public sector, previously associated with employment stability, is no longer attractive to young

people. Taking into account the results of previous research indicating the impact of salary on the choice of career path (Redmond and McGuinness, 2019; Pearlman, 2018) in the case of Poland, this may be related to lower salaries offered in the public sector compared to the private sector.

Millennials, when changing employers, seek a satisfying job to meet their subjective criteria (Campione 2015; Hassan et al., 2020). Polish Millennials had average cross-organizational mobility, with approximately half of them still in their first work at the time of the second wave of the GGS. Our evidence suggests that there are differences in the patterns of job mobility of young women and men in the Polish job market. Job changes early in one's professional career have both advantages and disadvantages. On the one hand, more frequent transitions can provide diverse professional experience and facilitate the discovery of a satisfying job, which may serve as a stepping stone to a successful future career. On the other hand, the lack of professional stability can hinder leaving the parental home and starting a family.

Simultaneously, frequent job changes among young individuals pose a significant challenge for employers. Considering the costs of turnover, employers have to make every effort to attract and retain valuable employees, particularly Millennials (Campione, 2015). The findings we have obtained can provide decision-makers with valuable insights for shaping strategies aimed at reducing employee turnover among Generation Y.

7. Limitations

This study has some limitations. First, the GGS-PL survey does not provide information on whether stopping work resulted from voluntary reasons or was determined by other factors. Moreover, we lacked detailed job-specific information, apart from the sector in which each individual was employed. Such details could have provided additional insights into gender-based disparities in the professional mobility of young individuals.

References

AbouAssi, K., McGinnis Johnson, J. and Holt, S. B., (2021). Job mobility among millennials: Do they stay or do they go?, *Review of Public Personnel Administration*, Vol. 41(2), pp. 219–249.

Allison, P. D., (2010). *Survival analysis using SAS: a practical guide*. Cary, North Carolina: Sas Institute.

Andersen, P. K., Gill, R. D., (1982). Cox's regression model for counting processes: a large sample study, *The annals of statistics*, pp. 1100–1120.

Arnett, J. J., (1998). Learning to stand alone: The contemporary American transition to adulthood in cultural and historical context. *Human development*, *41*(5-6), 295-315.

Arnett, J. J., (2006). Emerging Adulthood: Understanding the New Way of Coming of Age. In J. J. Arnett & J. L. Tanner (Eds.), *Emerging adults in America: Coming of age in the 21st century* (pp. 3–19). American Psychological Association.Bass, B. C., (2015). Preparing for parenthood? Gender, aspirations, and the reproduction of labor market inequality, *Gender & Society*, Vol. 29(3), pp. 362–385.

Becker, G., (1964), *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. New York: National Bureau of Economic Research.

Biemann, T., Zacher, H. and Feldman, D. C., (2012). Career patterns: A twenty-year panel study, *Journal of Vocational Behavior*, Vol. 81(2), pp. 159–170.

Boeckmann, I., Misra, J. and Budig, M. J., (2015). Cultural and institutional factors shaping mothers' employment and working hours in postindustrial countries, *Social Forces*, Vol. 93(4), pp. 1301–1333.

Blau, F. D., Kahn, L. M., (2017). The gender wage gap: Extent, trends, and explanations, *Journal of economic literature*, Vol. 55(3), pp. 789–865.

Cabello-Hutt, T., (2020). Changes in work and care trajectories during the transition to motherhood, *Social Science Research*, Vol. 90, 102439.

Campione, W. A., (2015). Corporate offerings: Why aren't millennials staying?, *Journal of Applied Business & Economics*, Vol. 17(4).

Cox, D. R., (1972). Regression models and life-tables, *Journal of the Royal Statistical Society: Series B* (Methodological), Vol. 34(2), pp. 187–202.

Cox, D. R., (1975). Partial likelihood, *Biometrika*, Vol. 62(2), pp. 269–276.

Cox, D. R., Oakes, D., (1984), *Analysis of Survival Data*. London: Chapman and Hall.

Cukrowska-Torzewska, E., Matysiak, A., (2020). The motherhood wage penalty: A meta-analysis, *Social science research*, Vol. 88, 102416.

Cukrowska-Torzewska, E., Lovasz, A., (2016). Are children driving the gender wage gap? Comparative evidence from Poland and Hungary, *Economics of Transition*, Vol. 24(2), pp. 259–297.

Eurostat, (2023). *Unemployment rate by age*. Retrieved October 31, 2023, from https://ec.europa.eu/eurostat/databrowser/view/tepsr_wc170/default/table?lang=en.

Gao, H., Luo, J., Tang, T., (2015). Effects of managerial labor market on executive compensation: Evidence from job-hopping, *Journal of Accounting and Economics*, Vol. 59(2-3), pp. 203–220.

Ghiselli, E. E., (1974). Some perspectives for industrial psychology, *American Psychologist,* Vol. 29(2), pp. 80–87.

Gray, R. J., (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, *87*(420), 942-951.

Grosemans, I., Hannes, K., Neyens, J. and Kyndt, E., (2020). Emerging adults embarking on their careers: Job and identity explorations in the transition to work, *Youth & Society*, Vol. 52(5), pp. 795–819.

Grzenda, W., (2019). *Modelowanie karier zawodowej i rodzinnej z wykorzystaniem podejścia bayesowskiego*. Warszawa: PWN.

Hassan, M. M., Jambulingam, M., Alagas, E. N., Uzir, M. U. H. and Halbusi, H. A., (2020). Necessities and ways of combating dissatisfactions at workplaces against the Job-Hopping Generation Y employees, *Global Business Review*, 0972150920926966.

Huang, P., Zhang, Z., (2016). Participation in Open Knowledge Communities and Job-Hopping, *MIS Quarterly*, Vol. 40(3), pp. 785–806.

Ignaczak, L., Raffestin, L. and Voia, M., (2022). Do the determinants of employment duration vary across employment spells?, *Applied Economics*, Vol. 54(9), pp. 1011–1029.

Kasprzak, E. K., (2023). Career patterns and career satisfaction of women and men in Poland in years 1990-2010, *Journal of Gender Studies*, pp. 1–14.

Kaufman, G., White, D., (2015). What makes a "good job"? Gender role attitudes and job preferences in Sweden, *Gender Issues*, Vol. 32, pp. 279–294.

Lake, C. J., Highhouse, S. and Shrift, A. G., (2018). Validation of the job-hopping motives scale, *Journal of Career Assessment*, Vol. 26(3), pp. 531–548.

Landmesser, J. M., (2013). Wykorzystanie metod analizy czasu trwania do badania aktywności ekonomicznej ludności w Polsce. *Rozprawy Naukowe i Monografie*. Warszawa: Szkoła Główna Gospodarstwa Wiejskiego w Warszawie.

Larasati, A., Aryanto, D. B., (2020). Job-Hopping and the determinant factors. *In 5th ASEAN Conference on Psychology, Counselling, and Humanities* (ACPCH 2019), pp. 54–56.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z., (2000). Semiparametric regression for the mean and rate functions of recurrent events, *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), Vol. 62(4), pp. 711–730.

Looze, J., (2017). Why Do (n't) they Leave?: Motherhood and women's Job Mobility, *Social Science Research*, Vol. 65, pp. 47–59.

Lyons, S. T., Schweitzer, L., Ng, E. S. and Kuron, L. K., (2012). Comparing apples to apples: A qualitative investigation of career mobility patterns across four generations, *Career Development International*, Vol. 17(4), pp. 333–357.

Matysiak, A., (2009). Employment first, then childbearing: Women's strategy in post-socialist Poland, *Population studies*, Vol. 63(3), pp. 253–276.

Moynihan, D. P., Landuyt, N., (2008). Explaining turnover intention in state government: Examining the roles of gender, life cycle, and loyalty, *Review of Public Personnel Administration*, Vol. 28(2), pp. 120–143.

Ng, E. S., Johnson, J. M., (2015). Millennials: Who are they, how are they different, and why should we care, *The multigenerational workforce: Challenges and opportunities for organisations*, pp. 121–137. https://api.semanticscholar.org/ CorpusID: 169202435.

Pearlman, J., (2018). Gender differences in the impact of job mobility on earnings: The role of occupational segregation, *Social Science Research*, Vol. 74, pp. 30–44.

Pedulla, D. S., (2016). Penalized or protected? Gender and the consequences of nonstandard and mismatched employment histories, *American sociological review*, Vol. 81(2), pp. 262–289.

Reshid, A. A., (2019). The gender gap in early career wage growth: The role of children, job mobility, and occupational mobility, *Labour*, Vol. 33(3), pp. 278–305.

Reichelt, M., Makovi, K., and Sargsyan, A., (2021). The impact of COVID-19 on gender inequality in the labor market and gender-role attitudes, *European Societies*, Vol. 23(sup1), pp. S228–S245.

Redmond, P., McGuinness, S., (2019). The gender wage gap in Europe: Job preferences, gender convergence and distributional effects, *Oxford Bulletin of Economics and Statistics*, Vol. 81(3), pp. 564–587.

Ripatti, S., Palmgren, J., (2000). Estimation of multivariate frailty models using penalized partial likelihood, *Biometrics*, Vol. 56(4), pp. 1016–1022.

Robak, E., (2017). Expectations of generation Y connected with shaping the work-life balance. The case of Poland, *Oeconomia Copernicana*, Vol. 8(4), pp. 569–584.

Schmitt, C., (2021). The impact of economic uncertainty, precarious employment, and risk attitudes on the transition to parenthood, *Advances in Life Course Research*, Vol. 47, 100402.

Steenackers, K., Guerry, M. A., (2016). Determinants of job-hopping: an empirical study in Belgium, *International Journal of Manpower*, Vol. 37(3), pp. 494–510.

Therneau, T. M., Grambsch, P. M., (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag

Twenge, J. M., (2010) A Review of the Empirical Evidence on Generational Differences in Work Attitudes, *Journal of Business and Psychology*, Vol. 25, pp. 201–210.

Wootton, B. H., (1997). *Gender differences in occupational employment*, *Monthly Labor Review*, Vol. 120, pp. 15–24.

Zamarro, G., Prados, M. J., (2021). Gender differences in couples' division of childcare, work and mental health during COVID-19, *Review of Economics of the Household*, Vol. 19(1), pp. 11–40.

Zhou, M., (2017). Motherhood, employment, and the dynamics of women's gender attitudes, *Gender & Society*, Vol. 31(6), pp. 751–776.

Yankow, J. J., (2022). The effect of cumulative job mobility on early-career wage development: Does job mobility actually pay?, *Social Science Quarterly*, Vol. 103(3), pp. 709–723.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 187–204, https://doi.org/10.59139/stattrans-2026-011
Received – 15.12.2024; accepted – 17.11.2025

Is the GPT model suitable for sentiment analysis? Testing for geographical, political and gender bias

Agnieszka Choczyńska[1]

Abstract

The new generation of Large Language Models, based on Generative Pre-trained Transformers (GPT) can be useful for automatic text annotation and sentiment analysis. However, they tend to learn the bias from training data, which can lead to distorted results. In this paper, the GPT-4o-mini model by OpenAI is tested for the presence of geographical, political and gender bias in the case of Polish economic news headlines. It has been found that the model consistently differs in sentiment scores for the same sentence, depending on the country mentioned. A remedy to this problem is proposed, which masks the references to countries and nationalities using the GPT model. Some differences in sentiment scores resulting from explicit references to gender or political parties are also identified, although these types of bias are considerably weaker than geographical bias.

Key words: large language models, geographical bias, gender bias, political bias, sentiment analysis.

1. Introduction

Large Language Models (LLMs) based on Generative Pre-trained Transformers (GPT) are increasingly used in scientific research. Text annotation is one of the applications that can benefit from the GPT models (Kheiri and , 2023). On the one hand, they are faster and more affordable than human annotators. On the other, they are more versatile than the language models trained for a narrow purpose and do not require a training dataset (Kocoń *et al.*, 2023). Given that most of the information created by the human population is in natural language, having a universal, ready-to-use text-mining tool would be beneficial for in social science.

There are, however, some obstacles to overcome. LMMs tend to absorb the biases found in the texts on which they are trained (Rozado, 2020). A growing amount of research finds gender (Radaideh, Kwon and Radaideh, 2025; Lee *et al.*, 2024; Zhu, Wang and Liu, 2024), nationality (Manvi *et al.*, 2024; Aslan 2024), political (Retzlaff, 2024; Rozado, 2023), and other types of bias (Huang *et al.*, 2020) in the text generated by the GPT models. Other studies focus on word embeddings (vector representations of words created during model training) and find that models pick up stereotypical associations from the natural language (Garg *et al.*, 2018; Rozado, 2020).

[1] AGH University of Krakow, Krakow, Poland. E-mail: aghachocz@agh.edu.pl.
ORCID: https://orcid.org/0000-0001-7134-567X.

One could suspect that the bias learned by the generative AI also impacts their abilities in text annotation. For example, they would assign lower sentiment to the sentences mentioning a country or demographic group the training datasets were biased against. In their analysis of sentences related to the energy industry, Radaideh, Kwon and Radaideh (2025) found that the GPT-2 model assigned a lower score to the sentences mentioning nuclear energy, male gender, old age or conservative political ideology.

However, the topic is still understudied. Firstly, most of the existing literature focuses on the English language, while one of the benefits of the GPT models is their multilingualism. The studies uncovering algorithmic bias in GPT models typically analyze the word embeddings or the impact of bias on text generation. Despite sentiment analysis being a popular application for this kind of models, it is still not well known how the bias can distort its outcomes.

This paper analyzes the bias in GPT-based text annotations, focusing on the Polish language and the texts broadly related to economics. The issue is approach from a new angle, using a set of fictional economic news headlines with positive, negative, or neutral implications for the mentioned country. Headlines are generated with different country names each time and prompt the GPT-4o-mini model to assign the sentiment to the sentence. The results show that countries significantly impact the sentiment score (p-value < 0.001), even though the headline does not change.

The same framework is used to test if the GPT models exhibit political bias in sentiment analysis. A set of headlines mentioning political orientation (left-wing or right-wing party), power dynamics (ruling party or the opposition), and the names of the main political parties in Poland are generated. As a control, there are also provided headlines where no political party, position or orientation is mentioned.

The results show that the sentences without any mention of a party tend to have the highest sentiment score, though the differences are generally small. A positive sentence gets, on average, a higher sentiment score if it mentions the ruling party, but negative sentences about the ruling party get lower scores than if they mention the opposition. No bias was found with regard to political orientation or particular party names.

Similarly, the paper assesses the presence of gender bias in sentiment analysis. The generated sentences include either a) direct mention of gender (e.g. men, women, male, female), b) mention of a fictional male or female name, or c) gendered grammatical forms. In Polish, verbs, nouns and adjectives have gendered forms, so it is impossible to change the gender of the subjects by just replacing names and pronouns.

The analysis shows very small effect of gender bias. Among the positive sentences, the ones mentioning female names received higher sentiment scores than the ones mentioning male names. Among negative sentences, the model assigned slightly lower scores to the ones that mentioned the female gender. The successes of particular women may be perceived more positively, as they are typically framed as a bigger breakthrough. On the other hand, if the problem is related to women (negative sentences with direct mention of gender), it is perceived as a bigger problem - and assigned a lower sentiment - than if it was related to men. However, no other configurations yielded significant differences. In particular, there is no evidence of gender grammatical forms impacting the sentiment score.

Finally, there is proposed a method of dealing with inherent geographical bias by censoring country names from the text. A dataset of economic news headlines from the public TV portal is used for this purpose. The GPT-4o-mini model is prompted to assign a sentiment score from -5 (strongly negative) to 5 (strongly positive) towards each country mentioned in the text. Next, there are create anonymized sentences by replacing all references to countries with codes and run sentiment analysis for that modified dataset.

The model consistently overstates the sentiment for Poland. For Germany and Russia, it tends to produce more neutral scores (e.g. less positive for positive news and less negative for negative news), while the references to the US receive more extreme values. However, the model's performance in country recognition and anonymization is unsatisfactory, leaving a room for improvement.

Although the researchers have been long aware of the problem of algorithmic bias, this paper expand the current knowledge by showing how it can impact the outcomes of sentiment analysis in a non-English language. Tests for geographical, gender and political bias reveal that they are all present, with the first one being by far the strongest. This study may be helpful for those looking to apply the GPT model to sentiment analysis, especially for the case of news analysis.

2. Literature review

2.1. Applications of the GPT models in sentiment analysis

With their natural language processing abilities, the GPT models could be used for sentiment analysis and other text-mining tasks. They are many times faster and more affordable than human annotators. Unlike specialized machine learning models, they can perform a wide variety of tasks on different forms of text. Some of the existing solutions are available through API, meaning that the researcher does not need to have the computational power or storage needed to train a large model.

These models can outperform untrained text annotators (Gilardi, Alizadeh and Kubli, 2023) and, in some cases, the state-of-the-art solutions (Kheiri and Karimi (2023); Fatouros *et al.*, 2023). However, their performance varies by task and dataset, and they should not be treated as a universally good solution (Curry, Baker and Brookes, 2024). Most research finds the GPT models to underperform, compared to the high-tuned, specialized language models (Kocoń *et al.*, 2023; Liyanage, Gokoni and Mago, 2024; Krugmann and Hartmann, 2024; Kristensen-McLachlan *et al.*, 2023).

However, there are a few caveats to this research. First, in the case of OpenAI, we do not know the full list of language corpora these models were trained on. If researchers perform the tests using publicly available datasets, the model may have already seen them, meaning that its performance on new data may be overestimated (Kocoń *et al.*, 2023; Ahuja *et al.*, 2023). Secondly, comparing a specialized model trained for a specific task with a GPT model in a zero-shot approach may underestimate the latter's abilities with additional training. It is still difficult to determine the extent of additional training this model would need to match the performance of a specialized solution, and if it would be indeed substantially smaller than preparing a model from scratch.

Finally, with the fast pace of AI development, it is difficult to assess their performance. Most of the aforementioned studies are not yet published, and the models they test will likely be obsolete before they do. Overall, researchers call for caution and additional validation before using GPT models for text annotation (Pangakis, Wolken and Fasching, 2023; Kristensen-McLachlan *et al.*, 2023; Ollion *et al.*, 2023; Curry, Baker and Brookes, 2024).

Most of the research on LLMs in text analysis is focused on the English language. A Common Crawl corpus, widely used in model training, has 45% of English texts, so one could expect models will be the most proficient in this language (Dac Lai *et al.* 2023). In a comprehensive evaluation of several LLMs (including GPT-3.5 and 4) in 70 languages, Ahuja *et al.*, (2023) noticed a worse performance for prompts written in non-English language. For the same reason, Etxaniz *et al.*, (2023) proposed translating the problem to English before analysis. Similar results were found by Dac Lai *et al.* (2023). However, there are some studies that counter these findings, not finding the benefits of English prompts (e.g. Debess, Simonsen and Einarsson, 2024). Both model and task-specific aspects may interfere with the results, although most of the research seems to find the benefit of English prompts.

This study focuses on bias in sentiment analysis and not on the accuracy of the model. Based on the research above, the authors decided to write the prompts in English and set the temperature parameter to 0.25, which gave the most consistent results in the previous experiments with the GPT model.

2.2. Bias in sentiment analysis

The GPT models are based on embeddings, which are vector representations learned from a large corpus of natural language. Closely related words in natural language should end up relatively close in the vector space (Garg *et al.*, 2018). If the corpus contains stereotypical associations between words, the model will likely incorporate that information and express human-like bias (Caliskan, Bryson and Narayanan, 2017). Moreover, the bias will not necessarily be diminished by more training, if the additional training dataset contains bias as well (Radaideh, Kwon and Radaideh, 2025).

Rozado (2020) found that most of the research on bias in word embedding models considered gender bias (93% of analyzed papers) and racial bias (54%). They performed a wider analysis of associations between positive words and terms related to gender, age, race, religiosity, affluence, and political orientation in several natural language corpora used in LLMs training. They found that positive words were more associated with women and femininity, youth, beauty, affluence and liberal political orientation. Typical African-American names and religiosity held negative associations, but the results for direct mentions of race or sexual orientations were mixed, with different directions, depending on the corpora. Garg *et al.*, 2018 found that associations between gender/race and certain occupations were correlated with the factual proportions of employees. The additional bias was largely explained by stereotypes held by the population.

The research on bias in the GPT models mainly focused on the stereotypical or toxic elements in the generated text. Huang *et al.*, (2020) asked the model to finish sentences with notions of different genders and occupations, finding that it can produce more negative

outputs in certain contexts. A similar framework was used by Lee *et al.,* 2024 in the South Korean context and Zhu, Wang and Liu (2024) in Chinese, finding bias related to gender and nationality.

Manvi *et al.*, (2024) performed a study of geographical bias, prompting the model to rank the countries, according to objective facts (e.g. population density), objective facts uncorrelated with geographic position (e.g. solar flux), and subjective opinion (e.g. attractiveness of the citizens). They found that the model systematically underestimates or overestimates the ranks of objective facts, despite being able to provide precise numbers when prompted. There is also a bias against the regions of lower socioeconomic conditions in rankings of subjective opinions. The authors tested 5 models and the GPT-4 exhibited the lowest bias.

Another form of bias is of a political nature. When asked questions from the political compass test, the GPT model leaned towards liberalism (Retzlaff, 2024). These findings are supported in the analysis of word embeddings by Rozado, 2020, but the political bias is not as well studied as that related to gender, race or nationality.

To what extent the bias built in the word embeddings or present in the generated text would impact the sentiment analysis? This topic is not yet well studied. Radaideh, Kwon and Radaideh (2025) studied the impact of bias on the sentiment scores assigned by five LLMs, including the GPT-2 model, in the case of sentences related to the energy industry. They generated sentences in which they switched terms related to energy source, politics, gender, age and ethnicity, and prompted the models to assign sentiment scores to each configuration. They found that the mentions of nuclear energy, conservative ideology, male gender, old age and white race usually lowered the sentiment score, however, with some variations between models. A similar approach is applied in this analysis.

2.3. Bias mitigation

Strategies to mitigate bias include a) creating more fair and balanced datasets, b) fairness-aware model training, and c) algorithmic debiasing (Srinivasan *et al.*, 2024, Liu, 2025). The first strategy may be done by balancing the dataset to obtain equal number of observations for majority and minority group (Han, Baldwin and Cohn, 2022). In an unbalanced dataset, minority groups may be classified with a higher error, due to their limited representation in the dataset. Another strategy is to embed fairness into the training process, designing loss-function so that it takes the bias into account.

Specifically in LLMs, it is possible to mitigate biases by manipulating word embeddings. Zhao *et al.*, (2018) used this approach to neutralize the gender connotations of (by definition) gender-neutral occupations. For example, a word "nurse" may refer to any gender, but its vector representation appears closer to "female" in the embedding space, as historically most nurses were women. Zhao *et al.*, (2018) captured the distances between occupations and genders and used them as weights in training of a gender-neutral model. Ravfogel *et al.*, (2020) presented an Iterative Null-Space Projection, a method of removing certain properties from neural representations. Liang *et al.*, (2021) tested their approach on the embeddings of the GPT-2 model.

The first problem is that these methods require *a priori* knowledge of all underprivileged groups and biases. Utama, Moosavi and Gurevych (2020) proposed a framework in which the first "shallow" model is trained on a limited dataset to pick up existing stereotypes and biases, which are further used to down-weight biased observations, lowering their impact on the final model. This is based on the assumption that biases represent the most superficial knowledge, that would be learned first by the model presented with limited data. A similar approach was tested by Orgad and Belinkov (2023).

The second problem is that these methods require either access to the data or repeating the training process. In the case of large, pre-trained models this may not be feasible. An alternative approach was proposed by Liu (2021), who attempted to mitigate political bias. They obtained the hidden states from the GPT-2 model and transformed them so that a gender neutral embedding was of equal distance to the two options, in this case liberal and conservative. However, they noticed a trade of between fairness and fluency and accuracy (see also: Nadeem, Bethke and Reddy, 2021, Liang *et al.*, 2021).

2.4. Hypotheses development

In this analysis, three types of sentiment bias are considered: geographical, political and gender bias. As public TV covers international news, one should expect the bias against particular countries could make a big difference in sentiment. The first hypothesis is based on the results obtained by Manvi *et al.*, (2024) and Rozado (2020):

H1: The GPT model is biased against countries of low socio-economic status.

Economic and business news may often reference political parties as well when they report government economic policy, investments, state-owned companies or corruption affairs. If the training data corpora contain positive associations with liberal and progressive political ideology (Rozado, 2020), one could expect the second hypothesis to be true:

H2: The GPT model is biased against right-wing parties.

Finally, the study considers the aspect of gender bias. In Polish, most parts of the speech have gendered forms. Every time a news headline mentions a person or a group of people, their gender is revealed through grammar. If the GPT model associates female names or grammatical forms with more positive sentiment, it could distort the analysis. Hence the third hypothesis:

H3: The GPT model is biased against men.

Although researchers note other dimensions of bias, they are not likely to impact the sentiment analysis in this case. Poland is a rather racially homogenous country and mentions of race are not common in economic news articles. Due to the economic focus, headlines do not generally mention physical appearance, sexual orientation or disabilities of the subjects, so these aspects were ommited as well.

3. Testing the GPT models for text annotation

3.1. Data

The data used in this analysis are news headlines from the business section of the Polish public TV internet portal. The dataset spans from 2012-06-13 to 2024-09-13 and consists

of 17,554 pieces. Each news piece is composed of a headline and a one- or two-sentence description, that introduces a longer video material.

The first part of this study uses generated headlines similar to these articles but constructed in a way suitable for bias testing. For geographical bias sentences have to mention exactly one country. They had to include one party or party members for the political test. In the case of the gender bias test, each sentence had to either directly mention gender or a fictional person of a specified gender. In the second part of the study, the original headlines are used to test how geographical bias impacted sentiment analysis in a real-life scenario.

Similarly to human annotators, the GPT model will not always return the same output for the same prompt. First, there is a check of replicability of the GPT text-annotation task results. A random sample of 1000 headlines is selected and the model is prompted to perform two text-mining tasks. The first is to assign a sentiment score from -5 (strongly negative) to 5 (strongly positive). The second one is to extract all countries mentioned in the text. Each analysis is performed 10 times in different sessions to assess the consistency of the results.

The results are fairly consistent. In 66.4% of cases, the model returned the same sentiment in each round, and only four times (0.4%) the difference was 3 points or more. For the country recognition task, the Jaccard similarity index was applied. For each pair of outputs, it counts the number of countries provided in both outputs (intersection of sets), divided by the overall number of countries that appeared in them (union of sets). The average score is 0.955.

All tasks are carried out with the GPT-4o-mini model by OpenAI using the Batches interface. Batches enable scheduling of a larger portion of API requests for asynchronous processing. As each text is to be analyzed independently, it is a suitable option for text-mining tasks.

3.2. Geographical bias test

The test uses 30 sentences (10 positive, 10 negative, and 10 neutral in sentiment). The sentences are similar to the news headlines in the TVP dataset, but constructed in a way that only one country is mentioned in a headline, and the country name is interchangeable. English translations of example headlines are provided below:

Positive: Prices no longer on the rise. Inflation in XXX falls quicker than expected.

Neutral: XXX is struggling with drought. The government is implementing special support programs for farmers.

Negative: In XXX, the problem of unemployment is getting worse. Every fifth adult is looking for a job.

The full set of headlines can be found in the repository (https://github.com/agachocz/SiT_GPT_bias_appendix.git). After generating the sentences, the XXX placeholder is replaced with one of the 194 country names in a correct grammatical form. Then the GPT-4o-mini model is prompted to assign the sentiment score to each sentence.

If the model's sentiment analysis abilities are not impaired by geographical bias, each sentence should be assigned the same score across countries. The sentiment analysis is repeated 10 times to test if potential differences in the outputs occur consistently.

Table 1 presents the test results: minimum, maximum and average score obtained for each group of sentences, along with the Kruskal-Wallis test statistic. Kruskal-Wallis test is a non-parametric alternative for ANOVA, more suitable for analyzing differences in rankings distributions. Under H_0 hypothesis, there are no statistically significant differences withing groups of sentences, therefore the model assigns the same sentiment score consistently, regardless of a country mentioned.

Table 1. The results of geographical bias test. Significance codes: * < 0.05, ** < 0.01, *** < 0.001

Group	Min score	Max score	Average score	Kruskal-Wallis test	Correlation with GDP PC
Positive	2	5	4.15	855.98***	-0.003
Neutral	-3	-3	0.534	602.36***	-0.055
Negative	-5	-1	-2.9	399.93***	-0.057
All	-5	5	0.593	77.209	

As presented in Table 1, the differences were significant in all sentiment categories, with the highest Kruskal-Wallis statistic for positive sentences. However, there is no significant effect for all categories taken together. One reason may be that the bias is not linear - some countries may score more positively in positive sentences and more negatively in negative.

This effect can be seen in Figure 1, presenting the average sentiment scores for all countries within categories. Note that the maps have independent colour scales so the differences are better visible. In positive sentences, Poland scores the highest among all countries, because the good news may seem the best for the Polish language users if it is related to their own country. However, among negative sentences, the sentiment for Poland is at the bottom of the scale, since the bad news may seem worse when they hit close to home.

Similarly, Russia scores the lowest in positive sentences and is about in the middle of the scale for negative ones. Since it is responsible for aggression on Ukraine near the Polish border, positive news about the Russian economy may be perceived more negatively, and negative ones - more positively.

To directly test Hypothesis 1, the correlation is computed between the average sentiment and the GDP per capita. The data on GDP expressed in purchasing power parity is sourced from the Worldometer database (Worldometer, 2024), excluding 14 countries for which there was no data. The correlation coefficients are presented in the last column of Table 1.

In all sentiment categories, correlations between sentiment scores and GDP per capita are small and insignificant. Contradictory to the hypothesis, the sentiment does not seem to depend on the economic development of the countries. Looking at Figure 1, it is clear that the sentiment scores are the lowest for the countries that may be perceived as a threat by Polish citizens. The most notable being Russia, but also North Korea, Afghanistan and China.

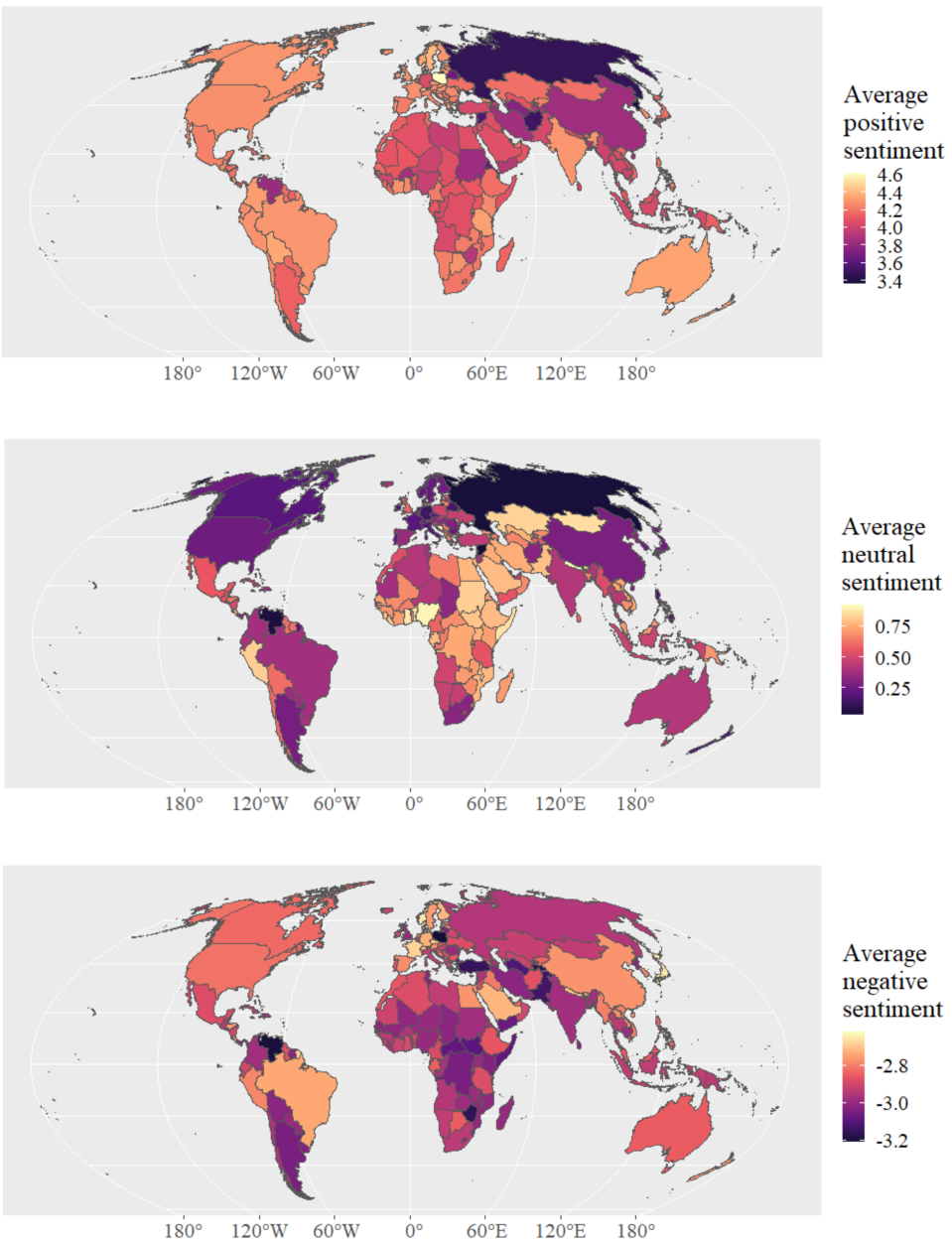

Figure 1. Average sentiment of positive, neutral, and negative sentences. Note that each map has an independent colour scale.

3.3. Political bias test

The same framework is used in the political bias test. Instead of country names, there are two terms for the party's position (ruling party, opposition), two terms for its political orientation (left-wing, right-wing), and eight names of the main political parties in Poland. The example sentences are as follows:

Positive: Renowned economist praises XXX's proposition. "This action is long overdue."

Neutral: XXX has published its new programme. What does it plan for seniors?

Negative: Millions in grants, zero results. The foundation linked to XXX MPs will come under scrutiny.

Additionally, there are generated the same sentences without any political terms, or with anonymous ones, such as "one of the parties" or "this party". This will provide the benchmark for the political terms. Again, the model is used to assign sentiment scores on a scale of -5 to 5 and repeat this task 10 times.

Table 2 provides the average scores for each term and the Kruskall-Wallis test statistics within groups. In general, any mention of politics lowers the sentiment of the sentence. Headlines without a term related to a party have the highest scores among positive sentences.

Table 2. The results of the political bias test. Significance codes: * < 0.05, ** < 0.01, *** < 0.001

Political term	Positive		Neutral		Negative	
	Average	Test	Average	Test	Average	Test
No term	3.470	-	0.340	-	-2.840	-
Position (ruling/opposition) + no term						
Ruling	3.32	11.889	0.31	0.067	-3.06	18.924
Opposition	2.97	**	0.38		-2.71	***
Orientation (left-wing/right-wing) + no term						
Left-wing	3.06	12.005	0.2	2.358	-2.78	0.944
Right-wing	2.98	**	0.26		-2.85	
Party names + no term						
KO	3.20	11.149	0.44	4.779	-2.780	6.901
Konfederacja	3.15		0.39		-2.770	
Nowa Lewica	3.20		0.40		-2.740	
PiS	3.26		0.21		-2.920	
PO	3.19		0.45		-2.810	
Polska2050	3.28		0.46		-2.770	
PSL	3.09		0.39		-2.780	

The Kruskal-Wallis test statistic is significant in the political orientation group for positive sentences. However, the pairwise Wilcox Rank Sum test reveals that only a difference between left/right orientation and no political term is significant. There is no difference between scores for left-wing and right-wing sentences, but they both score lower than sentences with no adjective related to political orientation. As there are no significant effects of bias toward specific party names, the Hypothesis 2 is rejected.

As for the position of the party in the political system, the ruling party tends to get higher scores in positive sentences, and lower for the negative ones, compared to the opposition. This may be a sign of more polarized opinions toward ruling parties, or assigning them a higher responsibility for the positive or negative outcomes.

3.4. Gender bias

Finally, the gender bias is tested. There are constructed sentences with mentions of gender, either explicitly (man/woman or male/female), through a fictional male or female name (Jan Nowak and Anna Kowalska), or just as a grammatical form (because nouns, verbs and adjectives have gendered alterations in Polish). The examples of headlines are provided below:

Positive: Infrastructure Minister Jan Nowak at the opening of the new power plant. "A milestone towards green energy."

Neutral: Polish men are innovative, but few of their discoveries live to see patents. Conclusions of a new CSO report.

Negative: Company founded by Jan Nowak in huge financial trouble. It is likely to declare bankruptcy.

The positive sentence above has a feminized version: «Infrastructure Minister **Anna Kowalska** at the opening of the new power plant. "A milestone towards green energy."» Then the name is removed to create an indirectly gendered sentence. Although in English this sentence would not imply the gender of the Minister, in Polish the word "Minister" would have two forms.

The example of a neutral sentence does not contain a name but a direct reference to the male gender, which can be switched to "women" for the female version. The indirect sentences are "Poles are innovative, but few of their discoveries live to see patents. Conclusions of a new CSO report.", where "Poles" has gendered form ("Polki" or "Polacy").

The results of the analysis are provided in the Table 3. In general, gendered grammatical forms do not differentiate sentiment scores. Among positive sentences, the ones including a female name received significantly higher scores than the ones including a male name. Possibly, the individual success of a woman is perceived as a bigger breakthrough (and more impressive by that) than the same success of a man. However, negative sentences mentioning the female gender received lower scores than the same sentences related to the male gender. This may be because problems seem more grave if they are related to women.

Overall, the results support Hypothesis 3, that the sentiment analysis with the GPT-4 model is biased against men. However, the effects are miniscule, compared to the scale of geographical bias, and the bias is related only to two specific cases.

Table 3. The results of gender bias test. Significance codes: * < 0.05, ** < 0.01, *** < 0.001

Type	Women avg	Men avg	Kruskall-Wallis test
	Positive		
Explicit gender	4	4	-
Gendered name	3.81	4.09	4.067*
Grammatical gender	4.05	3.90	2.620
	Neutral		
Explicit gender	0.1	-0.15	0.639
Gendered name	1.45	1.35	0.514
Grammatical gender	0.64	0.42	1.123
	Negative		
Explicit gender	-2.93	-2.73	4.248*
Gendered name	-3.16	-3.2	0.435
Grammatical gender	-3.15	-3.17	0.148

4. Anonymization

In the previous section, it was found that the GPT-4o-mini model is vulnerable to strong geographical bias. A method to remedy this problem is proposed by masking all references to country or nationality in the source text. This section is dedicated to a sentiment analysis of economic media headlines with and without masking and compare the results. The prompts used for particular tasks are provided in the repository (https://github.com/agachocz/SiT_GPT_bias_appendix.git).

First, the GPT model is prompted to modify the headlines by masking all countries or nationalities with codes: AAA, BBB, CCC, and so on. An example of this transformation could be:

Input: Hundreds of Estonian companies still trade with Russia. Estonian exports to Russia fell drastically after the latter invasion of Ukraine, but over 300 companies registered in Estonia kept trading with this country.

Output: Hundreds of AAA companies still trade with BBB. AAA exports to BBB fell drastically after the latter invasion of CCC, but over 300 companies registered in AAA kept trading with this country.

In the same prompt, the model is also asked to return the dictionary in the form "Country:code", in this case: "Estonia:AAA;Russia:BBB;Ukraine:CCC", to easily retrieve country names behind the codes after sentiment analysis.

Next, the model is asked to separately assign a sentiment score to the original and masked headlines. Finally, there is a comparison of the results to see to what extent the bias related to country names impacted the sentiment analysis. For this purpose, there were selected four countries with the biggest coverage. Poland has the highest number of mentions (5,733 news pieces), followed by Russia (1,399), Germany (1,235) and the US (1,017).

Table 4. Differences between sentiment assigned by the GPT model for sentences with and without anonymization. A positive average means that anonymized mentions receive higher scores than the ones revealing country names. N refers to the number of mentions

	Poland	Germany	Russia	US
n	3,403	865	1178	824
Correlation	0.854	0.768	0.709	0.773
		Positive		
n	2,363	348	375	454
average diff	-0.066	0.856	1.570	-1.040
std deviation	1.510	2.080	2.800	2.370
		Neutral		
n	348	164	171	139
average diff	-0.891	-0.445	0.474	0.583
std deviation	2.050	1.700	1.880	2.080
		Negative		
n	692	353	632	231
average diff	-1.240	-1.050	-0.231	0.342
std deviation	2.250	1.440	1.440	1.680

The correlation coefficients between sentiment scores for original and anonymized sentences are presented in Table 4. The scores are quite similar, as the correlation varies from 0.709 for Russia to 0.854 for Poland. The impact of bias related to country names is not very high.

The articles are split into sentiment categories based on the sentiment score from anonymized sentences: positive for scores higher than 2, negative for lower than -2, and neutral for scores between. Most mentions of Poland and the US were positive, while mentions of Germany and Russia were largely negative. Table 4 presents the difference between scores from anonymized and original sentences and present the averages and standard deviations.

In all categories, the averages for Poland are negative, meaning that mentions of Poland resulted in higher scores than thise that would have been obtained from the anonymized sentence. In the case of Russia and Germany, the average difference for positive sentences is positive, so mentions of these countries make the sentence seem less positive, compared to an anonymized sentence. However, for the negative sentences, there is the same pattern as for Poland, where the sentiment with country mention tends to be less negative than the sentiment without it. The reverse is true for the mentions of the United States. Here, the positive sentences typically receive higher scores, and negative sentences get even lower scores than the ones with masked country names.

Overall, the model seems to consistently overstate the sentiment, when Poland is mentioned. For Germany and Russia, it tends to produce more neutral scores (e.g. less positive for positive news and less negative for negative news), while the mentions of the US receive more extreme values.

However, the model does not handle these tasks perfectly. There are cases, where the outputs do not match, either because the model does not recognize all countries mentioned in the original sentence, makes mistakes in anonymizing, or fails to provide sentiment for all codes from masked sentences. The cases with multiple countries, nationality adjectives, or mentions of geographical regions, institutions, companies and other entities tend to produce mismatched outputs. These problems can be somewhat remedied by providing more precise prompts or cleaning the data afterwards, but the method still leaves room for improvement.

5. Conclusions

One of the drawbacks of automatic text annotations with Large Language Models is algorithmic bias. The models tend to learn the stereotypical associations present in human-created data used to train them and it may distort the results.

In the case of Polish economic news, this study tests for the presence of geographical, political and gender bias in sentiment scores assigned by the GPT-4o-mini model by OpenAI. The method uses generated sentences with interchangeable terms related to country names, political parties or gender, and prompt the model to assign sentiment scores on the scale from -5 (strongly negative) to 5 (strongly positive). If the model was unbiased, there should be no difference in sentiment scores.

The results show a significant effect of geographical bias. The GDP model tends to judge the same sentences as more or less positive, depending on the country mentioned. Contrary to other studies, sentiment bias was not correlated with the GDP of countries. For

the positive sentences, the lowest scores were obtained for countries violating international law and human rights, such as Russia or North Korea.

Political and gender bias were not that strong. Mentions of Polish parties did not differentiate sentiment and references to both left-wing and right-wing political orientations resulted in lower sentiment than no mention at all. Sentiment related to the ruling party was more polarized, compared to the opposition. Positive sentences mentioning the former were more positive, and the negative ones - were more negative. Positive sentences were also judged slightly more positively if they mentioned a female name, but negative sentences received more negative scores when the female gender was mentioned.

A remedy to the problem of geographical bias is proposed by masking all countries mentioned in the text. For this purpose, a dataset of economic news headlines from the public TV portal is used. The GPT model is prompted to assign a sentiment score from -5 (strongly negative) to 5 (strongly positive) towards each country mentioned in the text. Next, the anonymized sentences are created by replacing all references to countries with codes and run sentiment analysis for that modified dataset.

The study finds that the impact of bias depends on the country. For Poland, the model consistently provided higher scores for original sentences, compared to the anonymized headlines. For Germany and Russia, it tended to give less positive scores for positive news and less negative for negative news. On the contrary, the references to the US received more extreme values than the same sentences with masked country mentions. However, the model's performance in country recognition and anonymization has a room for improvement.

From the practical point of view for the GPT model users, testing for social bias is a necessity. The model can exhibit stereotypical tendencies when assigning sentiment, and the types of bias may depend on the specific data and use case. Recognizing and mitigating such biases should be a standard procedure in any sentiment analysis using large language models. This study shows that anonymization of the inputs may be a simple solution to deal with geographical bias, although in a larger scale, it requires a more reliable model to remove geographical markers from the text.

These findings could be useful for social science researchers interested in using Large Language Models for text analysis, especially if the source text is in Polish. The problem of algorithmic bias, especially related to countries and nationality, significantly affects the outcomes of sentiment analysis. However, the analysis is limited to Polish and short news articles about the economy. There may be cases where gender or political bias plays a greater role, or where sentiment analysis is distorted in other kinds of bias that have not been tested for here.

Additional data

Tables with generated sentences and all prompts used in the analysis are available in the repository: https://github.com/agachocz/SiT_GPT_bias_appendix.git.

References

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K. and Sitaram, S., (2023). MEGA: Multilingual Evaluation of Generative AI, in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, pp. 4232–4267. https://aka.ms/MEGA.

Aslan, F., (2024). *Bias assessment in Large Language Models*, PhD thesis, Tilburg University, Tilburg

Caliskan, A., Bryson, J. J. and Narayanan, A., (2017). Semantics derived automatically from language corpora contain human-like biases Science, 356, pp. 183–186. https://doi.org/10.1126/science.aal4230.

Curry, N., Baker, P. and Brookes, G., (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT, *Applied Corpus Linguistics*, 4(1). https://doi.org/10.1016/j.acorp.2023.100082.

Dac Lai, V., Trung Ngo, N., Pouran Ben Veyseh, A., Man, H., Dernoncourt, F., Bui, T. and Huu Nguyen, T., (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning, in 'Findings of the Association for Computational Linguistics: EMNLP 2023', *Association for Computational Linguistics, Singapore*, pp. 13171–13189 .

Debess, I. N., Simonsen, A. and Einarsson, H., (2024). Good or Bad News? Exploring GPT-4 for Sentiment Analysis for Faroese on a Public News Corpora, Technical report, *ELRA Language Resource Association*. https://huggingface.co/datasets/hafsteinn/.

Etxaniz, J., Azkune, G., Soroa, A., Lopez De Lacalle, O. and Artetxe, M., (2023). Do Multilingual Language Models Think Better in English?, in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, vol. 2, Association for Computational Linguistics, Mexico City, pp. 550–564

Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G. and Kyriazis, D., (2023). Transforming sentiment analysis in the financial domain with ChatGPT, *Machine Learning-with Applications*, 14, 100508.

Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J., (2018), Word embeddings quantify 100 years of gender and ethnic stereotypes, *Proceedings of the National Academy of Sciences of the United States of America*, 115(16), pp. 3635–3644.

Gilardi, F., Alizadeh, M. and Kubli, M., (2023). ChatGPT outperforms crowd workers for text-annotation tasks, *Proceedings of the National Academy of Sciences of the United States of America*, 120(30) .

Han, X., Baldwin, T., and Cohn, T., (2022). Balancing out Bias: Achieving Fairness Through Balanced Training. In Y. Goldberg, Z. Kozareva, and Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2022, pp. 11335–11350. https://doi.org/10.18653/v1/2022.emnlp-main.779.

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D. and Kohli, P., (2020). Reducing Sentiment Bias in Language Models via Counterfactual Evaluation, *arXiv*. http://arxiv.org/abs/1911.03064.

Kheiri, K., Karimi, H., (2023). SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning, *arXiv*. http://arxiv.org/abs/2307.10234.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, , Wojtasik, K., Woźniak, S. and Kazienko, P., (2023). ChatGPT: Jack of all trades, master of none, *Information Fusion*, 99 101861. https://doi.org/10.1016/j.inffus.2023.101861.

Kristensen-McLachlan, R. D., Canavan, M., Kardos, M., Jacobsen, M. and Aarøe, L., (2023). Chatbots Are Not Reliable Text Annotators, *arXiv* . http://arxiv.org/abs/2311.05769.

Krugmann, J. O. and Hartmann, J., (2024). 'Sentiment Analysis in the Age of Generative AI', *Customer Needs and Solutions*, 11(1).

Lee, S., Kim, D., Jung, D., Park, C. and Lim, H., (2024). Exploring Inherent Biases in LLMs within Korean Social Context: A Comparative Analysis of ChatGPT and GPT-4, in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 4, pp. 93–104.

Liang, P. P., Wu, C., Morency, L.-P. and Salakhutdinov, R., (2021). Towards Understanding and Mitigating Social Biases in Language Models, *ICML*. https://arxiv.org/abs/2106.13219.

Liu, R., Jia, C., Wei, J., Xu, G., Wang, L. and Vosoughi, S., (2021). Mitigating Political Bias in Language Models Through Reinforced Calibration, in *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*. https://doi.org/10.48550/arXiv.2104.14795.

Liu, Z., (2025). Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies, *Journal of Transcultural Communication*, 3(2), pp. 224–244. https://www.degruyterbrill.com/document/doi/10.1515/jtc-2023-0019/html.

Liyanage, C. R., Gokani, R. and Mago, V., (2024). GPT-4 as an X data annotator: Unraveling its performance on a stance classification task, *PLoS ONE*, 19 .

Manvi, R., Khanna, S., Burke, M., Lobell, D. and Ermon, S., (2024), Large Language Models are Geographically Biased, in *Proceedings of the 41st International Conference on Machine Learning*, 1409, pp. 34654 - 34669.

Nadeem, M., Bethke, A. and Reddy, S., (2021). StereoSet: Measuring stereotypical bias in pretrained language models, in 'Proceedings ofthe 59th Annual Meeting ofthe Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing', *Association for Computational Lingusitics*, pp. 5356–5371.

Ollion, É, Shen, R., Macanovic, A. and Chatelain, A., (2023). ChatGPT for Text Annotation? Mind the Hype! https://doi.org/10.31235/osf.io/x58kn.

Orgad, H. and Belinkov, Y., (2023). BLIND: Bias Removal With No Demographics, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, pp. 8801–8821. https://doi.org/10.18653/v1/2023.acl-long.490.

Pangakis, N., Wolken, S. and Fasching, N., (2023). Automated Annotation with Generative AI Requires Validation, *arXiv*. http://arxiv.org/abs/2306.00176.

Radaideh, M. I., Kwon, H. and Radaideh, M. I., (2025). Fairness and Social Bias Quantification in Large Language Models for Sentiment Analysis, Knowledge-based Systems 319, 113569. https://doi.org/10.1016/j.knosys.2025.113569.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M. and Goldberg, Y., (2020). Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256.

Retzlaff, N., (2024). Political Biases of ChatGPT in Different Languages, Preprints.org, URL: www.preprints.org.

Rozado, D., (2020). Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types, *PLoS ONE*, 15(4).

Rozado, D., (2023). The Political Biases of ChatGPT, textitSocial Sciences, 12(3), 148. https://doi.org/10.3390/socsci12030148.

Srinivasan, N., Perumalsamy, K., Sridhar, K., Rajendran, G. and Kumar, A. A., (2024). Comprehensive Study on Bias In Large Language Models, *International Refereed Journal of Engineering and Science*, 13(2), pp. 77–82 .

Utama, P. A., Moosavi, N. S. and Gurevych, I., (2020). Towards Debiasing NLU Models from Unknown Biases, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7597–7610

Worldometer, (2024). Worldometer GDP per capita dataset. https://www.worldometers.info/gdp/gdp-per-capita/, accessed: 23.09.2024.

Zhao, J., Zhou, Y., Li, Z., Wang, W. and Chang, K.-W., (2018). Learning Gender-Neutral Word Embeddings, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853. http://arxiv.org/abs/1809.01496.

Zhu, S., Wang, W. and Liu, Y., (2024). Quite Good, but Not Enough: Nationality Bias in Large Language Models – A Case Study of ChatGPT, *arXiv*. http://arxiv.org/abs/2405.06996.

STATISTICS IN TRANSITION new series, March 2026
Vol. 27, No. 1, pp. 205–220, https://doi.org/10.59139/stattrans-2026-012
Received – 03.07.2024; accepted – 19.05.2025

Survey sampling in wartime: addressing challenges for cross-national and longitudinal studies

Olena Kaminska[1]

Abstract

Conflict situations pose unique challenges before survey sample design, three of which are examined in this paper in the context of the ongoing full-scale war in Ukraine. Firstly, defining the population becomes a complex task as a result of substantial population changes. This includes internal displacement, a significant proportion of the population emigrating (but with a potential future return), and a considerable number of individuals actively engaged in combat who are currently unreachable for interviews. Secondly, defining households, particularly for longitudinal studies, is complicated by the temporary separation of many families. Thirdly, accurate accounting for vacant and demolished dwellings is essential, as national statistics often lack precise and reliable data in these areas. This study outlines the design of two samples: one for a cross-national European Social Survey in Ukraine (ESS), and the other for the longitudinal household panel study, UKRAINS. By carefully addressing these complex sampling challenges, it is possible to develop high-quality probability samples that account for population mobility and unpredictability in a wartime context.

Keywords: sample design, population definition, eligibility, conflict, war.

1. Introduction

Conflict situations bring significant changes to populations, and these need to be reflected in sample design to ensure accurate representation. Unlike the usual population dynamics in peaceful times, which include births, deaths, and relatively stable immigration and emigration patterns, conflict introduces large-scale and unpredictable changes. The full-scale invasion of Ukraine by Russia in 2022 exemplifies these challenges, presenting unique obstacles for survey sampling.

Firstly, the conflict has led to the non-coverage of territories temporarily occupied by Russian forces, areas of active fighting, and adjacent unsafe regions. Additionally,

[1] Institute for Social and Economic Research, University of Essex, Essex, United Kingdom.
E-mail: olena@essex.ac.uk. ORCID: https://orcid.org/0000-0002-2138-1311.

there has been a substantial internal displacement, with people moving from Eastern and Southern Ukraine to safer areas in Central and Western Ukraine (UNHCR, Regional Bureau for Europe, 2024). The war has also resulted in significant emigration, primarily of women and children, as men are largely prohibited from leaving the country during the war (UNHCR, Regional Bureau for Europe, 2024). Furthermore, individuals currently serving in the military are not accessible for surveys.

Household structures have also been profoundly affected, many of which are temporarily separated. The two main types include: a household where the wife (and children, if any) has left the country for safety while the husband remains in Ukraine, and another where one member is serving in the military while the other stays home. Despite the physical separation, these households often continue to function as cohesive units, maintaining financial and emotional support and intending to reunite as soon as circumstances allow (e.g. the war ends).

The conflict has also resulted in a large proportion of vacant dwellings. Families may have moved abroad, joined the military, or relocated within the country to safer areas. The proportion of vacant houses varies by region, urban versus rural settings, and proximity to the frontline, reflecting the perceived safety of these areas. Reliable information on vacant housing is difficult to obtain prior to sampling, with many estimates lacking precision.

These complexities create a new context for planning survey sample design. This paper addresses these challenges through two separate designs: a cross-national survey and a longitudinal household survey. After a look at the Ukrainian context and literature review of the survey sampling in conflict, we follow with a brief description of the two studies, their objectives, and their designs. We then discuss the impact of wartime on the definition of a population, the associated challenges, and our solutions for a sample design. Next, we address the difficulties in defining households and our approaches to these issues. Vacant houses problem and how they are accounted for in the sample design is explored next. Finally, we describe sampling-related challenges during fieldwork. The discussion section summarizes the main challenges in detail and concludes with suggestions for future studies in similar conflict situations.

2. Ukrainian context

The Russian invasion of Ukraine began in 2014 with the annexation of Crimea, followed by the occupation of large parts of Eastern Ukraine, including the regional centers of Luhansk and Donetsk. This led to a significant internal displacement, with many internally displaced persons (IDPs) relocating to areas under Ukrainian control. It is estimated that the initial stage of the war resulted in approximately 1.4 million IDPs moving to safer locations within Ukraine (Mikheieva & Kuznetsova, 2023). This is

around 3% of the total population of Ukraine of 45.4 million estimated as of 2014 (Kulu et al., 2022). The Minsk I and II agreements (OSCE, 2014; OSCE, 2015) led to a cessation of active conflict until 2022, allowing IDPs to settle and survey organizations to interview them in their new homes. This period provided invaluable experience in surveying IDPs, residents near frontlines, those in occupied territories, and decommissioned military personnel (e.g. Cafiero et al., 2021; ILO Decent Work Technical Support Team and Country Office for Central and Eastern Europe, 2016). During this time, active fighting was confined to the frontline and occupied territories, leaving the rest of Ukraine relatively safe.

The full-scale invasion launched on February 24, 2022, marked a dramatic escalation, with Russia attempting to occupy most, if not all, of Ukraine. Military engagements occurred in Central, Eastern, and Southern Ukraine, and Russian missiles, along with military drones, targeted the entire country, leaving no place safe for civilians. This resulted in around 8.2 million individuals leaving Ukraine, at least temporarily, and around 8 million being internally displaced in the first months of the war (Rogoza, 2023). This is a substantial proportion of 37.3 million population living in Ukraine-controlled territories as of 2019 (Rogoza, 2023). Following the invasion, martial law was imposed, prohibiting males of military age (18-60) from leaving Ukraine, except under specific conditions.

Some people have returned home, but as of February 2024 the number of new IDPs moving to safer areas within Ukraine who remain displaced is estimated at 3.7 million (IOM UN migration, 2024). An estimated 6.4 million people remain abroad as of the end of 2023 (UNHCR, Regional Bureau for Europe, 2024). Of those who are abroad 6.0 million are hosted in countries across Europe with 88 percent of them being women and children (UNICEF, 2024). Of those who stayed in Europe the most common countries for Ukrainians to find refuge as of 2023 was Germany (1.03 million), Poland (994,000), and Czech Republic (448,000). Those outside of Europe, just over 400,000, are mainly in Canada (220,000) and the USA (over 100,000) (UNHCR, Regional Bureau for Europe, 2024). Separately, 2.8 million Ukrainians have registered residence in Russia since 2022, including 19,500 forcibly deported children (Rogoza, 2023).

Since the beginning of the full-scale invasion Ukraine has experienced profound additional demographic changes: mass casualties among soldiers and civilians, including children; a significant decline in birth rates due to excess mortality and deteriorating living conditions; large-scale relocations from Eastern to Western Ukraine; substantial emigration; demographic shifts in major cities such as Odesa, Dnipro, Lviv, and Kyiv due to both evacuation and relocation; the movement of entrepreneurs and manufacturing to safer areas; and extensive destruction of civilian infrastructure, particularly housing (Libanova and Pozniak, 2022).

Despite the ongoing conflict, some Ukrainians, estimated at 900,000 (UNHCR, Regional Bureau for Europe, 2024) as of the end of 2023 have returned from abroad to their homes, and additionally 298,000 returned from abroad to a different location from their original home, while many of homes still remain under occupation or near the frontline. Additionally, some IDPs continue to return to their homes within Ukraine, particularly in de-occupied areas. Many, however, are still awaiting the opportunity to return, with estimates of those remaining abroad about 6.4 million (UNHCR, Regional Bureau for Europe, 2024).

3. Literature on sampling in conflict situations

In any context, the primary goal of a sampling statistician is to obtain a representative sample of a population in the most cost-efficient manner. Representativeness is ensured by giving everyone in the population a chance to be selected, with this chance being known as a specific probability (Kish, 1965). Sampling in conflict situations shares the same fundamental challenges as in other contexts but faces two additional significant issues: large population movements, such as an influx of refugees to neighboring countries or internal displacement of people (Anguilera et al., 2020; Mneimneh et al., 2014; Lubbad, 2024), and often a lack of up-to-date sampling frame information, which quickly becomes unrepresentative of the new reality of population distribution (Anguilera et al., 2020; Mneimneh et al., 2014; Lubbad, 2024; Box and Thomas, 1944). Similar challenges are encountered in surveys conducted in post-conflict situations, partly due to the population movement back home after temporary shelter in other parts of the country or abroad (Lynn, 2004).

Sampling theory has made significant advances in the field of selection where no sampling frame is available (e.g. West, 2016), for hard-to-reach populations (e.g. Raifman et al., 2022), or even where the population is highly mobile (Raifman et al., 2022; Reichel and Morales, 2017). By combining these methods with classic selection techniques driven by the context of the available sampling frame information, statisticians can develop probability samples in conflict situations. Common solutions to the lack of up-to-date sampling frame and population distribution information include using satellite imagery to define first-stage clusters (Anguilera et al., 2020; Mneimneh et al., 2014; Lubbad, 2024), satellite photographs with random point estimates and circles around them as clusters (Shannon, 2012), multiple frames, respondent-driven samples (Mneimneh et al., 2014; Khoury, 2019), and random walk samples (Shannon, 2012; Spagat, 2012).

Simultaneously, several considerations must be addressed: due to safety concerns, some insecure areas may be excluded (Mneimneh et al., 2014), or access to affected communities may be limited (Spagat, 2012; Mneimneh et al., 2014). When situations

rapidly change, flexibility in sample design is required to reflect new security concerns and population flows (Mneimneh et al., 2014). Importantly, noncontact and nonresponse (Mneimneh et al., 2014; Lubbad, 2024) and mistrust (Lubbad, 2024) may also be higher. Additionally, new sampling frames may become available during conflicts, such as ration card lists in WWII UK (Box and Thomas, 1944), UNICEF immunization lists in Afghanistan, and food distribution lists in Kosovo (Mneimneh et al., 2014).

Ukraine has successfully conducted social surveys during the ongoing armed conflict since 2014, primarily via face-to-face mode, and since COVID-19 also via phone mode (largely mobile phones) (Paniotto, 2022). Among survey errors, the war has impacted coverage error, unit nonresponse, and measurement error (for sensitive questions) but has not affected sampling error, item nonresponse, interviewer error, or processing error (Paniotto, 2022). In the Ukrainian context, coverage error results from the inability to survey occupied territories by Russia, and to a lesser extent, areas near the frontline. However, this noncoverage in terms of population percentage does not equate to the proportion that lived in these areas before the war. A significant portion of that population have moved to safer parts of Ukraine, where they are reachable and can be interviewed by a survey organization. Similarly, areas close to the frontline have negligible populations remaining, while most residents have moved to other parts of Ukraine or abroad (Paniotto, 2022).

A separate consideration involves newcomers to occupied territories, where Russia has intentionally illegally settled Russians (Myroshnychenko, 2023). Although they reside in the geographically recognized territory of Ukraine, they never lived there before and did not have Ukrainian citizenship. Including or excluding them from the population is a theoretical consideration that should reflect research aims. In practice, while the war is ongoing and Russia controls these territories, it is likely that there will not be access for an objective social survey of Russian settlers in Ukrainian territories. Such proportions are not negligible, as it is estimated that 500,000 – 800,000 illegal new Russian settlers have been living in Crimea since 2014 (Ostiller and The Kyiv Independent news desk, 2023), while the Crimean population in the 2001 Census was just over 2 million (Wikipedia contributors, 2024); and 40,000 Russian citizens settled in Mariupol as of 1 July 2023 with Russians planning for this number of go over 300,000 with pre-invasion population of Mariupol of around 420,000 (Myroshnychenko, 2023).

Estimating population totals and density in Ukraine presents an additional challenge. With significant population movements, an outdated Census from 2001, and no reliable administrative data depicting the full population picture, innovative methods are required. Sarioglo and Ogay (2022) propose an innovative method of population estimation by modelling population size, density, and location using mobile network and mobile phone operators' data combined with administrative registers and a social survey on mobile device usage. This approach provides timely estimates of population

size across regions and localities and potentially their changes. Yet, the approach is limited to Ukraine-controlled territories where Ukrainian mobile service is available. In areas not controlled by Ukrainian government Russia blocks the mobile signal to Ukrainian networks, and data usage of mobile phones becomes unavailable, thus preventing population estimates in occupied territories.

Finally, the Ukrainian context also poses challenges for longitudinal studies. Establishing a longitudinal panel before the start of a war has substantial advantages, as continuing through the war can provide invaluable information on opinion and life changes. However, this also brings methodological challenges. A two-wave opinion study by PONARS collected in December 2019 via face-to-face mode and in October 2022 via telephone in Ukraine faced significant challenges due to large population movements, especially from Eastern parts of Ukraine (Rickard et al., 2023). This resulted in substantial attrition with important differences across regions, with higher attrition rates in areas with higher civilian fatalities, often concentrated on the frontline and in occupied territories. This is critical for opinion surveys in Ukraine because Eastern Ukraine had different political attitudes, including those towards Russia and NATO, compared to Western Ukraine prior to the full-scale war. The task of a methodologist is to disentangle true changes in opinion from changes due to attrition caused by the war. Here, attrition combines two factors: typical nonresponse and eligibility definition, where a significant portion of the population may become ineligible. Rickard et al. note that they excluded Ukrainians who moved abroad (to Western countries or to Russia) between the first and second waves of data collection, which may include substantial proportions of Eastern Ukrainians. Such exclusion may lead to observed changes in opinions, which need to be accounted for when analyzing longitudinal data for pre- and post-war analysis.

4. Sampling theory and conflict

The first step in planning a sample design is to define the population we aim to represent. This population consists of units of analysis and should be entirely driven by the theoretical needs of the research, without consideration of practical challenges, sampling frame limitations, or other constraints. For example, in the Ukrainian context, we might be interested in residents or households currently living in Ukraine (encompassing the entire 1991 recognized territory).

The first challenge faced by survey data collection in conflict zones is the safety of interviewers, particularly in face-to-face surveys. Data cannot be collected in occupied territories or areas near the frontline, resulting in non-coverage. Addressing this issue involves acknowledging these limitations in the analysis and interpretation of results, as there is little scope for a practical solution.

The next step in sample design is to ensure the probability nature of the sample, meaning every eligible unit of the population should have a known, nonzero chance of being selected. In the Ukrainian context, accurate pre-full-scale-war population statistics, specifically population estimates for electoral districts and precincts (from 2019), can serve as a convenient sampling frame. Since the start of a full scale invasion significant population changes have occurred across these districts and precincts which is crucial to recognize in the design.

Incorporating a longitudinal aspect into a panel adds a time dimension to the population definition. In peaceful circumstances, this involves accounting for mortality and incorporating newborn children into the study. Migration considerations include people who have left the population and, through boosts, recent immigrants (including population rejoiners). In conflict situations, these aspects still need to be addressed, but additional challenges arise. Solutions are required to represent newly deoccupied territories when they become safe, to include households when they rejoin, to account for substantial future internal movements when parts of the country become deoccupied and safe, and to reflect changes in the population when the war is over.

5. Data

We explore two surveys planned for Ukraine to be conducted during the war: Round 11 of the European Social Survey (ESS) and a longitudinal household survey (UKRAINS).

The ESS is a cross-national survey conducted biannually in 20-30+ countries. It is a cross-sectional, face-to-face survey of adults (15+ years old) aiming for an effective sample size of 1,500 per country (or 700 in very small population countries). Its main goal is to provide data on attitudes, beliefs, and behavior patterns for cross-country comparison, enabling time-trend analysis for some core questions (European Social Survey European Research Infrastructure Consortium [ESS ERIC], 2021). Round 11 took place in 2023-2024, with Ukraine participating.

The ESS methodology focuses on cross-country comparability, ensuring consistency in questionnaire design, fieldwork procedures, nonresponse correction, and sample design. The ESS survey in Ukraine is planned within this context.

The second survey planned for Ukraine is the longitudinal household survey (UKRAINS). Its aim is to represent the people of Ukraine longitudinally in their household context, starting in the current situation and following them into the long-term future. The goal is not only to track individuals currently living in Ukraine, but also to be able to represent Ukrainian population as it changes in the future, and to represent household structure in its novel temporary state, where some households are separated.

The planned sample size is 4,000 households in wave 1, with around 12,800 adult (18+) individual interviews.

6. Population definition and noncoverage in the context of conflict

The ESS employs a standard population definition across all participating countries, including all persons aged 15 or older living in private dwellings. This definition, essential for cross-national and time-series comparability, presents several challenges in the current Ukrainian context:

- Ukrainians abroad: Ukrainians who are abroad are not part of the population definition. An estimated 6 million people left Ukraine in early 2022. This substantial proportion, relative to the pre-war population of around 40 million, includes many who intend to return but are currently not in Ukraine. These individuals are excluded from the Ukrainian ESS population but may be included in ESS surveys in their current countries of residency.
- Servicemen Not Living at Home: Between 600,000 and 1 million people are estimated to be in service, often not living at home. These individuals, part of the Ukrainian population more broadly, are now excluded from ESS definition as they do not reside in private dwellings.
- Unsafe Areas for Face-to-Face Interviews: Around 20% of Ukrainian territory is currently occupied by Russians, including Crimea and major cities like Luhansk, Donetsk, and Mariupol. Unsafe areas also include frontline and nearby areas, which are reached by Russian shelling. Interestingly, the population definition would technically include many Russians who moved into the vacated homes of Ukrainian families who fled the war following the invasion. In Crimea, it is estimated that 400,000 Russians have relocated since the Russian occupation, accounting for 20% of Crimea's 2013 population. However, due to non-coverage, it is not possible to study this subgroup.

When the war ceases, the data will quickly become outdated in terms of population representation. Comparison to future ESS rounds will be problematic as changes in opinions will be confounded with population shifts. However, the ESS sample design aims to represent individuals aged 15+ currently living in private dwellings in Ukraine.

UKRAINS aims to represent the current and future Ukrainian population. It considers the return of people currently abroad, servicemen rejoining their families post-war, and internal population movements following de-occupation. To achieve this, UKRAINS plans for:

- Sample refreshments to include de-occupied territories;
- Sample refreshments across Ukraine to include returning households;

- Include servicemen as panel members, initially through proxy interviews;
- Include family members currently abroad in split households, if at least one member remains in a private dwelling in Ukraine.

The definition for UKRAINS is: people living in private dwellings in Ukraine and their household members.

Non-coverage of unsafe territories, as described for ESS, also applies to UKRAINS. However, in a longitudinal context, the definition of unsafe territories may change with frontline movements and de-occupation. People returning to de-occupied territories, both IDPs and those returning from abroad, will need to be represented. Some individuals may have remained in their homes throughout the occupation and thus would not have a chance to be included in the initial panel selection. These individuals would need to join the panel upon de-occupation. To represent all three categories of people, a separate sample from de-occupied territories is necessary. This approach should account for the double selection probabilities of returning IDPs or other people, who might be selected initially in safer areas and again upon their return. A sampling statistician may consider excluding such people based on their previous chance of selection. However, reselecting them through a refreshment sample can mitigate the effects of attrition potentially heightened by shifts in life circumstances or geographic relocation. Consequently, it may be more practical to conduct a representative sampling in recently de-occupied territories, adjusting for the dual selection probabilities associated with internally displaced persons (IDPs) or other individuals migrating from previously safer regions of Ukraine. A simple representative sample in de-occupied territories also avoids challenges of household eligibility definition, particularly in cases where only some members of a household relocated from other parts of Ukraine, while others remained during the occupation. In such situation a selection of the household would be automatic into a sample, but sample selection probabilities would be calculated separately for each household member, reflecting their previous circumstances at the time of the original panel selection.

Since Russia's full-scale invasion in February 2022, many Ukrainians have returned home for various reasons, including the reduced likelihood of Kyiv, Central, and Northern Ukraine being occupied and better protection from Russian air threats. However, safety is not guaranteed, and larger population movements are expected when the war ends. Therefore, two types of refreshments are needed: during the war to account for those returning home from abroad, and post-war to account for a potentially larger internal movement and further returns from abroad.

In terms of the timeline UKRAINS aims to represent life events since the beginning of the full-scale invasion. As the panel has not been in field at that time, retrospective data collection through an event history calendar (EHC) is planned. This will gather detailed information on major life events for all panel members since February 2022,

regardless of their start time in the panel. This approach will allow for the study of war experiences and their influence on future life decisions.

A separate group of interest includes IDPs. Those residing in private dwellings within government-controlled areas can be selected through an address-based sample. This includes IDPs who reside on their own or with relatives and / or friends in safer parts of Ukraine. However, representing IDPs in temporary shelters (such as hotels, hostels, and schools) presents a challenge. When practical restrictions prevent interviewing these individuals, future refreshment samples may include them once their circumstances stabilize.

In the current context of Ukraine, defining a household presents unique challenge, partly due to the presence of multiple households cohabiting in the same dwelling. This arrangement often arises as families host relatives or friends from less secure regions, or as households join relatives in rural areas with alternative heating sources, particularly during winter, in response to Russian targeting of Ukrainian infrastructure. Given these dynamics, different researchers may adopt distinct household definitions tailored to their specific research questions. In this context, a practical approach for a sampling statistician may involve selection of all individuals residing at the same dwelling (i.e. the same postal address) and collecting data on key household dynamics. Questions addressing shared meals, financial arrangements, and prior household compositions would provide flexibility, enabling researchers to apply various household definitions according to their analytical needs.

The population definition related to split households is discussed in the following section.

7. Household survey of split households

ESS primarily focuses on individual opinions, beliefs, and behavior, rendering household context relatively insignificant. Conversely, UKRAINS is a household panel study that examines life decisions and choices, where household context is crucial. The standard household definition—comprising individuals who live together and share finances and meals—may not accurately reflect the current situation in Ukraine, where many households are temporarily split.

For example, a husband serving at the frontline for over a year, while his wife and children remain at their family home, still supports his family financially, communicates frequently, and plans to return home post-war. Despite the physical separation, they consider themselves one household. Another example involves a wife and children who have relocated to a safer European country, while the husband remains in Ukraine, unable to leave due to wartime restrictions. They maintain financial ties, communicate daily, and intend to reunite once it is safe. We refer to these as temporarily split households.

At the data collection stage, we could define households in several ways: as those who lived together before February 2022, as those who currently share finances, or based on individuals' own definitions of their household as they see it today. However, it is critical to include all household members at the panel's start to fully understand the household context, encompassing members not currently residing at the dwelling. UKRAINS plans to include servicemen who belong to the household as panel members, conduct brief proxy interviews while they are serving, and conduct full interviews upon their return. The study also intends to interview household members abroad via online video calls, continuing these interviews over time. This approach will enable the study of household dynamics including when members from abroad return to Ukraine.

While some servicemen and Ukrainians living abroad will be represented in the study, not all of them will be included. Households without a member residing in a private dwelling will be excluded from the initial wave, for example where a single servicemen lived alone before deployment or households where all members have moved abroad. However, they will be represented in future refreshments if they return to live in a private dwelling in Ukraine.

8. Vacant and demolished houses

Ukraine does not have a named register available to sampling statisticians, making an address-based sample a viable option. For the ESS sample design, 2019 electoral precincts are used as clusters, each containing an average of around 1,500 individuals aged 18 and over, with a range of about 300 to 4,000 per precinct. By estimating average household sizes, we can approximate the number of households per precinct in 2019. However, many dwellings are now vacant or demolished, and the population distribution has changed significantly.

The most efficient sample allocation is an equal probability sample, which, at the cluster selection stage, implies a proportional to population size (PPS) allocation. This ensures that clusters with larger populations have a higher chance of being selected, reflecting their relative size. Ideally, the most recent population size information would be used, including only current residents and excluding those who have temporarily left their homes. Unfortunately, high-quality, up-to-date statistics at the precinct level are not available.

To address this, the plan is to select all dwellings—vacant or not—and then screen for occupancy among the selected dwellings. Since the exact number of vacant dwellings is unknown, the selection must include both occupied and vacant dwellings in proportions representative of the cluster. Importantly, as the screening will follow the selection of clusters, the PPS should reflect the total number of dwellings, including both

occupied and vacant ones. The 2019 statistics provides a good estimate of the total number of dwellings even at the time of the war. Thus, clusters can be selected using PPS according to the total number of dwellings.

There is no high-quality list of dwellings within each electoral precinct. After cluster selection, enumeration is necessary, involving enumerators listing all dwellings in the cluster. During this stage, demolished dwellings and non-private buildings can be excluded. A set number of dwellings (e.g. 25) is then randomly selected in each PSU, and interviewers can determine occupancy during visits. Some vacant dwellings may not be detected and might be classified as non-contacts. Nevertheless, this sampling design ensures equal selection probabilities for occupied dwellings, and their PPS representation.

9. Challenges during fieldwork

Several challenges can arise during the fieldwork stage. The situation may change between the drawing of the sample and the data collection. For example, in our experience with the ESS, the Ukrainian-controlled part of the Donetsk region was deemed safe at the time of drawing the sample, and two clusters were selected in this region. However, by the time of enumeration, the safety situation had deteriorated. While the overall sample size remained unchanged and enumeration proceeded in many other PSUs, we decided to randomly allocate additional dwellings from the Donetsk region across the rest of the sample. Some flexibility in the design and ability to adapt to a rapidly changing situation is therefore necessary, and planning for different scenarios in advance can be beneficial.

Another challenge is related to males aged 25-60 who can be conscripted into military service. While many men volunteer for service or join when called, some avoid service and hide in their homes. This can create difficulties for interviewers listing household members, as these men may not be reported to avoid detection. This issue can be partially mitigated through interviewer training and assurances of confidentiality, but some proportion of this subgroup may still be omitted during the war. To address this, we plan to reassess household composition at the end of the war, allowing for updates and corrections to previous responses, including the addition of any household members not initially listed.

We also anticipate that the reported dwelling vacancy rate may be underrepresented. In unsafe regions, a neighbor might look after multiple vacant dwellings and may be reluctant to report these vacancies due to fear that such dwellings could be targeted by criminals. While this issue does not affect population representation, it can

lower apparent response rates, as genuinely vacant houses will be counted as non-contacts. Again, interviewer training and assurances of confidentiality can help mitigate this problem.

10. Discussion and conclusions

While theoretical sampling concepts such as population and household definitions are relatively straightforward in many survey contexts, conflict situations necessitate a much more careful consideration of these definitions, especially for longitudinal and household surveys. Population movement within a country, emigration and return patterns at the onset of conflict, and the continued movement as some areas become safer or more dangerous, occupied, or deoccupied, create a fluid population that still needs to be accurately represented at each point in time through high-quality sample design.

This paper provides practical solutions to sampling in a conflict situation in Ukraine during a full-scale war. A cross-sectional study, such as the ESS, is easier to plan since its purpose is to capture a snapshot of the population at a single point in time. Although the population described in such a study may quickly become outdated, it remains relevant to a critical historical moment in the country's life. However, analysts must be aware of missing subgroups: IDPs living in non-private households, servicemen (who now constitute a larger proportion of the population than in non-conflict circumstances), and individuals who are abroad unless specifically covered.

For a household longitudinal study, additional considerations are necessary to ensure continuous representation of the population, reflecting all its changes. While it is challenging to plan for the unknown, different scenarios can be considered. This includes tracking people who flee for safety and through sample refreshment accounting for those who return to their old homes as areas become safe or are deoccupied. Incorporating split households into the design can involve including household members who are currently away (even long-term) if one member of the household can be interviewed. These away members can be interviewed via proxy or video interviewing to provide a complete household context.

Finally, post-stratification may be challenging or significantly limited compared to times of peace, as earlier statistics may no longer accurately reflect the current population due to drastic changes. Even fundamental demographic characteristics, such as age by gender distribution, have undergone substantial shifts during the full-scale war. A large number of males aged 25-60 are in military service, while many young females and children have moved abroad. Additionally, in less safe regions, there are significantly fewer children than before.

Unlike in many other surveys where post-stratification can help correct for non-coverage, the current context in Ukraine lacks reliable statistics for territories occupied

by Russia or frontline areas. Therefore, using poststratification to adjust for noncoverage may not be an option.

Although this paper describes two studies in Ukraine in the context of 2023–2024, the challenges outlined are similar to those encountered in other conflict situations. Our sampling experience may serve as a valuable starting point for planning sample designs in similar circumstances elsewhere.

References

Aguilera, A., Krishnan, N., Muñoz, J., Russo Riva, F., Sharma, D. and Vishwanath, T., (2020). Sampling for Representative Surveys of Displaced Populations. In J. Hoogeveen & U. Pape (Eds.), *Data Collection in Fragile States* (Chapter 8). International Bank for Reconstruction and Development/The World Bank.

Box, K., Thomas, G., (1944). The Wartime Social Survey. *Journal of the Royal Statistical Society*, 107(3/4), 151-189. Wiley for the Royal Statistical Society. Stable URL: https://www.jstor.org/stable/2981213.

Cafiero, C., Yassin, F., Hopkins, C. and Battistella, E., (2021). Food Security & Livelihoods Assessment in Eastern Ukraine. FAO. Retrieved from https://fscluster.org/sites/default/files/documents/eastern_ukraine_gca_food_security_and_livelihood_final-ready_to_print.pdf.

European Social Survey European Research Infrastructure Consortium (ESS ERIC) Director, in collaboration with the Core Scientific Team (CST), (2021). *Round 11 Survey Specification for ESS ERIC Member, Observer and Guest Countries* (v1), November 29, 2021. Retrieved from https://www.europeansocialsurvey.org/sites/default/files/ 2023-06/ESS11_survey_specification.pdf.

ILO Decent Work Technical Support Team and Country Office for Central and Eastern Europe, (2016). *Employment needs assessment and employability of internally displaced persons in Ukraine: summary of survey findings and recommendations* / Budapest: ILO. Retrieved from https://www.ilo.org/wcmsp5/groups/public/---europe/---ro-geneva/---sro-budapest/documents/publication/wcms_457535.pdf.

IOM UN migration, (2024). *Ukraine & neighbouring countries 2022-2024: 2 years of response*. Retrieved from https://www.iom.int/sites/g/files/tmzbdl486/files/documents/2024-02/iom_ukraine_neighbouring_countries_2022-2024_2_years_of_response.pdf.

Khoury, R. B., (2020). Hard-to-survey populations and respondent-driven sampling: Expanding the political science toolbox. *American Political Science Review,* 18(2). https://doi.org/10.1017/S1537592719003864.

Kish, L., (1965). *Survey Sampling*. John Wiley and Sons Inc., New York.

Kulu, H., Christison S., Liu C. and Mikolai J., (2022). The war and the future of Ukraine's population. *MigrantLife, Working paper* 9.

Libanova, E., Pozniak, O., (2023). War-driven wave of Ukrainian emigration to Europe: An attempt to evaluate the scale and consequences (the view of Ukrainian researchers). *Statistics in Transition new series and Statistics of Ukraine, A New Role for Statistics: Joint Special Issue*, 24(1), pp. 259–276. ISSN 1234-7655.

Lubbad, I., (2024). Presentation at the inaugural meeting: Revision of the United Nations Handbooks related to Household Surveys. *Household surveys in conflict and humanitarian settings*. Cluster 4: Statistics, Information Society and Technology, Economic and Social Commission for Western Asia.

Lynn, P., (2004). Development of a sampling method for household surveys in post-war Bosnia and Herzegovina. *Statistics in Transition,* 6(6), pp. 953–977.

Mikheieva, O., Kuznetsova, I., (2023). Internally displaced and immobile people in Ukraine between 2014 and 2022: Older age and disabilities as factors of vulnerability. *Migration Research Series,* No. 77. International Organization for Migration (IOM), Geneva.

Mneimneh, Z. N., Axinn, W. G., Ghimire, D., Cibelli, K. L. and Alkaisy, M. S., (2014). Conducting surveys in areas of armed conflict, 7, pp. 134–156. *Cambridge University Press*. https://doi.org/10.1017/CBO9781139381635.010.

Myroshnychenko, V., (2023). The shifting demography of the Occupied Territories (2022-2023). Retrieved from https://t4pua.org/en/2119.

Organization for Security and Co-operation in Europe (OSCE), (2014). *Protocol on the results of consultations of the Trilateral Contact Group, signed in Minsk, 5 September 2014*, OSCE, retrieved 3, July 2024, https://www.osce.org/files/f/documents/a/a/123258.pdf.

Organization for Security and Co-operation in Europe (OSCE), (2015). *Package of Measures for the Implementation of the Minsk Agreements, 12 February 2015*, OSCE, retrieved 3, July 2024, https://www.osce.org/files/f/documents/5/b/ 140221.pdf.

Ostiller, N., and The Kyiv Independent news desk, (2023). Up to 800,000 Russians have reportedly moved to Crimea since occupation in 2014. *The Kyiv Independent*. Retrieved from https://kyivindependent.com/media-around-800-000-russians-have-moved-to-occupied-crimea-since-illegal-annexation-in-2014/.

Paniotto, V., (2022). Challenges of surveys in Ukraine under conditions of war. In Workshop *The future of social research in and on Russia and Ukraine*. Hanse-Wissenschaftskolleg Delmenhorst.

Raifman, S., DeVost, M. A., Digitale, J. C., et al., (2022). Respondent-Driven Sampling: a Sampling Method for Hard-to-Reach Populations and Beyond. *Curr Epidemiol Rep,* 9, pp. 38–47. https://doi.org/10.1007/s40471-022-00287-8.

Reichel, D., Morales, L., (2017). Surveying immigrants without sampling frames – evaluating the success of alternative field methods. *CMS,* 5, 1. https://doi.org/10.1186/s40878-016-0044-9.

Rickard, K., Toal, G., Bakke, K. M. and O'Loughlin, J., (2023). How Reliable Are Polls In Wartime Ukraine? *PONARS Eurasia Policy Memo No. 830*. United Nations University, Virginia Tech, University College London, Peace Research Institute Oslo (PRIO), University of Colorado Boulder.

Rogoza, J., (2023). Ukraine in the face of a demographic catastrophe. OSW Commentary, Centre for Eastern Studies, 524.

Sarioglo, V., Ogay, M., (2023). Approach to population estimation in Ukraine using mobile operators' data. *Statistics in Transition new series and Statistics of Ukraine, A New Role for Statistics: Joint Special Issue,* 24(1), pp. 131–144. ISSN 1234-7655.

Shannon, H. S., Hutson, R., Kolbe, A., Stringer, B. and Haines, T., (2012). Choosing a survey sample when data on the population are limited: A method using Global Positioning Systems and aerial and satellite photographs. *Emerging Themes in Epidemiology,* 9(5). http://www.ete-online.com/content/9/1/5.

Spagat, M., (2012). Estimating the Human Costs of War: The Sample Survey Approach. In M. R. Garfinkel & S. Skaperdas (Eds.), *The Oxford Handbook of the Economics of Peace and Conflict*. Oxford Handbooks. https://doi.org/10.1093/oxfordhb/9780195392777.013.0014.

UNHCR, Regional Bureau for Europe. (2024, February). *Ukraine refugee situation: Population movements | Factsheet #1*. Retrieved from https://data.unhcr.org/en/documents/details/106707.

UNICEF, (2024). *Humanitarian action for children: Ukraine and refugee response*. Retrieved from https://www.unicef.org/media/150011/file/2024-HAC-Ukraine.pdf

West, P. W., (2016). Simple random sampling of individual items in the absence of a sampling frame that lists the individuals. *N. Z. j. of For. Sci.,* 46, 15. https://doi.org/10.1186/s40490-016-0071-1.

Wikipedia contributors, (2024). 2001 Ukrainian census. In *Wikipedia, The Free Encyclopedia*. Retrieved 10, 30, July 3, 2024, from https://en.wikipedia. org/w/index.php?title=2001_Ukrainian_census&oldid=1216156907.

STATISTICS IN TRANSITION new series, March 2026
Vol. 21, No. 1, pp. 221–224

About the Authors

Alizadeh Morad received the Bachelor's degree in Statistics from Shahid Chamran University, Ahvaz, Iran, in 2003, the Master's degree in Statistics from Ferdowsi University of Mashhad, Mashhad, Iran, in 2005, and the PhD degree in statistics from the Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, in 2019. Since 2025, he has been an Associate Professor in the Faculty of Intelligent Systems Engineering and Data Sciences, Persian Gulf University, Bushehr, Iran. He has authored or coauthored more than 180 research papers in scientific journals. His research interests include distribution theory, lifetime distributions, discrete distributions, and related areas. Dr. Alizadeh serves as a member of the editorial board for several journals.

Arasan Jayanthi is an Associate Professor at the Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia. She has obtained her DPhil in Statistics from the University of Oxford. Her main research interests include survival analysis, reliability modelling, and computational statistics. She has authored more than 70 journal publications, 4 books and academic modules, and holds 7 intellectual property copyrights, including statistical software innovations. She has led more than seven nationally funded research projects and is a member of the editorial boards of several international statistical journals. She has received the Vice-Chancellor's Fellowship Award for Teaching Excellence at Universiti Putra Malaysia.

Ben Ali Mohamed is a Professor of higher education authorized to direct research at the Department of Mechanical Engineering at the Higher School of Technology of Casablanca, Hassan II University. He specializes in industrial engineering. His main areas of interest include: industrial quality and safety, production management, food safety, bacteriology, parasitology, multivariate statistical analysis, statistical modeling, biostatics, multivariate functional data analysis, statistical inference, and data analysis. He can be contacted at benali8mohamed@gmail.com.

Białek Jacek is a Full Professor at the University of Łódź in Poland, where he has been a long-time faculty member in the Department of Statistical Methods. His primary research interests focus on the theory and application of price indices, with particular expertise in the measurement of the Consumer Price Index (CPI) and the Harmonized Index of Consumer Prices (HICP). Additionally, he works at Statistics Poland in the Department of Price and Services, specializing in the analysis of scanner and web-scraped data. Professor Białek has authored over 100 scientific publications and developed the *PriceIndices* R package, a tool widely used for analyzing scanner data and

calculating price bilateral and multilateral indices. He is also a member of the Editorial Board of the *International Journal of Statistics and Probability*. For his scientific achievements, Professor Jacek Białek has received numerous prestigious awards, including the Minister's Award, the City of Łódź Award, the "Łódzkie Eureka" distinction, as well as several Rector's Awards from the University of Łódź.

Choczyńska Agnieszka is a Research and Teaching Assistant at the Department of Applications of Mathematics in Economics, Faculty of Management, AGH University of Krakow. Her main areas of interest include economic sentiment indicators, text mining and time-series models of financial markets. She is also a part of an international project ODDEA (Overcoming Digital Divide in Europe and Southeast Asia), funded by Horizon Europe.

Eftekharian Abbas is an Associate Professor in the Department of Statistics, Faculty of Basic Sciences, at the University of Hormozgan. His primary research interests include ordered data analysis, nonparametric inference, ranked set sampling, income inequality, and classification methods. He currently serves as a reviewer for several well-established academic journals.

Fidler Julia studies mathematics at the Nicolaus Copernicus University in Toruń, specializing in applications of mathematics in economics and finance. Her research interests focus on general probability theory, stochastic processes, and game theory. They also include auction mechanisms as well as the analysis of economic preferences and decision-making, including the use of topological methods.

Grzenda Wioletta is an Associate Professor at the Institute of Statistics and Demography at SGH Warsaw School of Economics. She has a PhD in Mathematics. She has received habilitation in Economics and Finance from SGH Warsaw School of Economics for her work on Bayesian modelling of family and occupational careers. She has published papers on the applications of Bayesian and classical statistical methods in the analysis of employment, fertility and probability theory. She is an author and co-author of books on Artificial Intelligence**,** Bayesian statistics, advanced statistical methods, and programming in data analytics.

Kaminska Olena is a survey statistician at the Institute for Social and Economic Research at the University of Essex. She works on the UK Household Panel Study and is also a member of the Sampling and Weighting Panel for the European Social Survey. Her main research areas include sampling and weighting in complex practical situations, panel retention and motivation to participate in surveys, and the overall quality of survey response.

Kharazmi Omid is an Associate Professor in the Department of Statistics at Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran. He received his B.S. degree in Statistics from Shahid Bahonar University, Kerman, Iran, in 2007, and completed his M.S. and Ph.D. degrees in Statistics at the University of Isfahan, Iran, in 2009 and 2019, respectively. He has actively collaborated with researchers both locally and internationally, fostering

a global exchange of knowledge and ideas. His research findings have been disseminated in reputable journals, contributing to the advancement of statistical science. His research interests include applied probability modeling, reliability theory, information theory, data science, and Bayesian analysis.

Marszałek Agnieszka is a graduate of the SGH Warsaw School of Economics, majoring in Data Analysis – Big Data. From 2021 to 2023, she worked as an assistant at the Institute of Statistics and Demography at the SGH Warsaw School of Economics. Her research interests focus on the situation of young people in the labor market and the application of statistical methods and data mining in business analysis. Simultaneously, since 2019, she has been working at PKO Bank Polski SA as a data quality expert, specializing in data quality research and management, including the design of IT solutions in this area.

Matysiak Łukasz is an Assistant Professor at the Institute of Mathematics and Cryptology, Faculty of Cybernetics, Military University of Technology in Warsaw. His main research areas include algebra (with particular focus on factorization), cryptologic and economic applications of algebraic methods, as well as economics, especially preference theory. He is a member of the Editorial Board of Military Technology and Science (formerly *Biuletyn Wojskowej Akademii Technicznej*).

Muthukumar Madaswamy is an Assistant Professor in the Department of Statistics at PSG College of Arts & Science, Coimbatore, Tamil Nadu, India, with over 18 years of teaching experience. He specializes in statistical modeling, time series analysis, stochastic processes, applied probability, and survival analysis. His research interests include Hidden Markov Models, count data models, and advanced statistical distributions. He has guided several postgraduate and doctoral research scholars and has published research articles in reputed national and international journals. He has also presented papers at various conferences and actively contributes to academic and research development activities.

Oullada Oumaima is a PhD researcher in industrial engineering at the Laboratory of Processes, Mechanics, Materials and Industrial Engineering (LP2MGI) of the Higher School of Technology and the National Higher School of Electricity and Mechanics, Hassan II University, Casablanca, Morocco. She can be contacted at oumaimaoullada@gmail.com.

Paweł Lula holds a Full Professor position at the Department of Computational Systems at the Krakow University of Economics. He works as a researcher and academic teacher in data analysis, Artificial Intelligence (mainly natural language processing) and decision support systems (multicriteria analysis, graph-based methods, decision trees). He is experienced in design and implementation of data analysis and decision support systems in Python and R. He has worked as a visiting professor at several foreign universities in Serbia, Ukraine, Russia, Romania, Hungary and Italy.

Pumputis Dalius is an Associate Professor in the Department of Mathematical Statistics, Faculty of Fundamental Sciences, at Vilnius Gediminas Technical University (VILNIUS TECH). He is a member of the Lithuanian Mathematical Society and serves on its Audit Committee. He also participates in the international project Baltic-Nordic-Ukrainian Network on Survey Statistics, helping to organize the network's conferences. His main research interests include survey sampling, time series, and estimation of the distribution of L-statistics.

Ranjbar Vahid is an Assistant Professor of Statistics at the Department of Statistics, Faculty of Science of the University of Golestan. His research interests are distribution theory, censored data analysis, statistical inference and data analysis. Dr. Ranjbar has published more than 60 research papers in international/national journals and conferences.

Rifai Said is a Professor of higher education at the Department of Mechanical Engineering at the Higher School of Technology of Casablanca, Hassan II University. He specializes in industrial engineering. His main areas of interest include: industrial quality and safety, statistical modeling, and data analysis. He can be contacted at said57.rifai@gmail.com.

Sassi Abdellah is a PhD researcher in industrial engineering at the Laboratory of Processes, Mechanics, Materials and Industrial Engineering (LP2MGI) of the Higher School of Technology and the National Higher School of Electricity and Mechanics, Hassan II University, Casablanca, Morocco.

Usman Mahamood is an Assistant Professor of Statistics, Data Science and Applied Mathematics at Vellore Institute of Technology (VIT) Vellore, Tamil Nadu, India, in the Department of Mathematics. He has completed his PhD from Indian Institute of Technology (IIT) Dhanbad, India in 2022. He completed his MPhil program in 2016 from Aligarh Muslim University, Aligarh, India. His research interest is in the areas of sampling techniques, applied statistics, machine learning and data science. He has published various research papers (more than 20) in reputed journals (Q1, Q2, SCI indexed, Scopus, etc.) worldwide.

Vyshnavi Muraleedharan is a doctoral researcher in the Department of Statistics at PSG College of Arts & Science, Coimbatore. She has completed her PhD in Statistics from Bharathiar University. Her research focuses on Hidden Markov Models, count time series analysis, and statistical distributions. Her doctoral thesis, titled "Hidden Markov Models for Predicting Dynamic Patterns Using Statistical Distributions and Fuzzification Techniques" contributes to advanced modeling approaches for dynamic data. She has published research articles in reputed journals and presented papers at national and international conferences. Her academic interests include statistical modeling, probability theory, and applied statistics.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* https://sit.stat.gov.pl/ForAuthors.

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **Bold**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper.
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, (**1.1.**, **1.2.** …), **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see https://anglia.libguides.com/harvardctr. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).