Volume 26, Number 4, December 2025 ISSN 1234-7655

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

Swanson D. A., Tayman J., On using ARIMA model confidence intervals applied to population projections based on the components of change: a case study for the world population

Girul A., Factors determining the formation of degraded areas in local government units and the effectiveness of revitalisation activities

Khazaei S., Bohlourihajjar S., Bayesian nonparametric model for weighted data using mixture of Burr XII distributions

Kumar J., Nain A., Exploring new mixtures of distribution to model a skewed and heavy tailed data

Kubus M., Mach Ł., Misiurski P., Application of the nonlinear splines model to forecast changes in the construction costs index

Agu F. I., The truncated Schröter recursive algorithm for the computation of aggregate claim amounts

Ogunde A. A., Nymphas E. F., Inverse Power Lomax Poisson distribution: properties and applications in modelling negatively-skewed reliability data

Wesołowska M., Exploring variation in data on income inequality across databases and measures in post-socialist countries

Kończak G., On a new goodness of fit test for multivariate normality with fixed parameters based on David-Hellwig test idea

Falkiewicz E., The concept of behavioural model of decision-making under risk

Thakur N. S., Chaurasia S., Bhayare U., Mean estimation based on the factor-type estimator under adaptive cluster sampling design

EDITORIAL OFFICE **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

Submission information for authors III

From the Editor VII

Original Research Papers

Swanson D. A., Tayman J., On using ARIMA model confidence intervals applied to population projections based on the components of change: a case study for the world population 1

Girul A., Factors determining the formation of degraded areas in local government units and the effectiveness of revitalisation activities 15

Khazaei S., Bohlourihajjar S., Bayesian nonparametric model for weighted data using mixture of Burr XII distributions 35

Kumar J., Nain A., Exploring new mixtures of distribution to model a skewed and heavy tailed data 55

Kubus M., Mach Ł., Misiurski P., Application of the nonlinear splines model to forecast changes in the construction costs index 69

Agu F. I., The truncated Schröter recursive algorithm for the computation of aggregate claim amounts 89

Ogunde A. A., Nymphas E. F., Inverse Power Lomax Poisson distribution: properties and applications in modelling negatively-skewed reliability data 111

Wesołowska M., Exploring variation in data on income inequality across databases and measures in post-socialist countries 127

Conference Papers

XXXXII Multivariate Statistical Analysis 2024, Lodz, Poland

Kończak G., On a new goodness of fit test for multivariate normality with fixed parameters based on David-Hellwig test idea 151

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

Falkiewicz E., The concept of behavioural model of decision-making under risk 167

Research Communicates and Letters

Thakur N. S., Chaurasia S., Bhayare U., Mean estimation based on the factor-type estimator under adaptive cluster sampling design 185

About the Authors 201

Acknowledgments to Reviewers 205

Index of Authors (2025) 211

STATISTICS IN TRANSITION new series*, December 2025*
Vol. 26, No. 4, 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*, December 2025*
Vol. 26, No. 4, 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

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, December 2025
Vol. 26, No. 4, pp. VII–X

From the Editor

It is with great pleasure that we present our readers the last issue of our quarterly in 2025, consisting of 11 articles covering a wide spectrum of topics discussed by 19 authors from USA, Poland, Iran, India, Slovakia and Nigeria. Let us take this opportunity to wish you Happy Holiday Season and A Happy New Year 2026.

Original Research Papers

In the first paper, ***On using ARIMA model confidence intervals applied to population projections based on the components of change: a case study for the world population***, **David A. Swanson** and **Jeff Tayman** discuss how measures of uncertainty derived from a standard time series model (ARIMA) can be applied to an existing population projection. Specifically, how to apply the uncertainty measures to a world population forecast based on the Cohort-Component Method. The results are compared to the Bayesian probabilistic world forecast developed by the United Nations and found to be similar but showing more uncertainty. The results are followed by a discussion suggesting that this new method is well-suited for developing probabilistic world, national, and sub-national population forecasts.

Agata Girul's paper entitled ***Factors determining the formation of degraded areas in local government units and the effectiveness of revitalisation activities*** focuses on identifying the factors that determine the occurrence of deprived areas requiring revitalization. The article uses data from the survey SG-01 report Municipal Statistics – Revitalisation. The PROFIT (PROperty FITting) multidimensional scaling program for local level territorial units was applied. The program takes into account the delimitation of rural areas based on the typology of Functional Urban Areas. Calculations and figures were made in Statistica 13. The focus on the problematic areas revealed the variety of challenges faced by local territorial units in their respective revitalisation programs. The comprehensive analysis of the causes-and-effects of the degradation and revitalisation processes may prove useful tools in developing effective local development strategies while improving the quality of life of residents of local communities.

In the next paper, ***Bayesian nonparametric model for weighted data using mixture of Burr XII distributions***, **Soleiman Khazaei** and **Soghra Bohlourihajjar** discuss a Bayesian nonparametric approach for analyzing weighted survival data using the Dirichlet Process Burr XII Mixture Model (DPBMM) to estimate the underlying density and survival functions. Parameters are inferred using Markov Chain Monte

Carlo (MCMC) methods, and the Metropolis-Hastings algorithm is applied to obtain de-biased samples from the weighted observations. Numerical illustrations are provided using both simulated and real lifetime data, including the presence of censored observations. The performance of the proposed method is compared with classical kernel density estimates to demonstrate its flexibility in modeling complex and heavy-tailed distributions.

Jitendra Kumar's and **Anuj Nain's** article ***Exploring new mixtures of distribution to model a skewed and heavy tailed data*** presents two distributions – Log-logistic and Inverse Weibull distribution – as an extension of all those proposed by Miljkovic and Grün (2016) due to adding and examining K-component finite mixtures of two more distributions. The EM algorithm has been employed for parameter estimation and then best model has been selected using three model selection criterions, namely NLL, AIC and BIC. The risk measures such as VaR and TVaR have been also computed and compared with their empirical counterparts to assess the goodness of fit of our proposed models at the extreme quantiles. It was found that K-component mixture distribution of Log-logistic and Inverse Weibull works better than competent models. To get more generalized view on the theory of mixture distribution the simulation was carried out providing satisfactory results.

The paper by **Mariusz Kubus**, **Łukasz Mach**, and **Przemysław Misiurski** entitled ***Application of the nonlinear splines model to forecast changes in the construction costs index***, presents an innovative approach to forecast the construction costs index (CCI), which is an important macroeconomic indicator. Due to the long-term nature of the investments in the construction market, the authors tested the model in a ten-month ahead period. Except minor disruptions, which were likely related to COVID-19, they obtained promising results, which definitely outperformed the classical ARIMA and its variant with nonlinear autocorrelation functions modeled with neural network. The achieved forecast results will enable both the demand and supply in the construction market to be in market equilibrium and minimize the formation of speculative bubbles in the market.

In the article entitled ***The truncated Schröter recursive algorithm for the computation of aggregate claim amounts***, **Friday I. Agu** introduces and evaluates the truncated Schröter recursive algorithm for computing aggregate claim amounts in the insurance sector. The algorithm addresses the limitations in the existing methods by incorporating truncation at 1, which is crucial for an accurate modelling of insurance claims where the events leading to a claim are pivotal. Using the AutoCollision dataset, the study compares the truncated Schröter algorithm with the Panjer and Schröter recursion algorithms, focusing on computational efficiency and accuracy. Furthermore, the descriptive statistics revealed substantial variability and risk factors, such as higher claim severity for business-use vehicles and young drivers aged 17–20. The results demonstrate that the truncated Schröter algorithm substantially reduces the execution time while maintaining high accuracy, thus making it a superior tool for risk management and premium setting.

In the next paper, ***Inverse Power Lomax Poisson distribution: properties and applications in modelling negatively-skewed reliability data***, **Adebisi A. Ogunde** and **Emmanuel F. Nymphas** propose a new, four-parameter distribution with increasing, decreasing, bathtub-shaped and a unimodal failure rate, called the Inverse Power Lomax Poisson (IPLP) distribution. The new distribution combines Inverse Power Lomax (IPL) and Poisson distributions along with several properties of the new distribution: its probability density function, its reliability and failure rate functions, the quantiles, the stress-strength parameter, complete and incomplete moments, the moment generating function, the probability weighted moment, Rènyi and q-entropies, and order statistics were derived. The study presents the estimation of the model's parameters based on the maximum likelihood method. The applications of the new distribution are presented using two real data sets, showing its flexibility and potential in modelling lifetime data.

In the paper entitled ***Exploring variation in data on income inequality across databases and measures in post-socialist countries***, **Monika Wesołowska** examines the consistency of income inequality data in post-socialist countries in Central and Eastern Europe and Central Asia across common measures and databases. Such analyses were carried out for single measures, databases, or selected countries, aimed at identifying the research gap for a selected group of countries. The study indicates a high level of consistency in income inequality trends over the long term and highlights strong correlations between different data sources for the same measures. However, they are inflated by the high consistency of data for EU countries, which is why only for this subgroup it would be possible to truly confirm the existence of consistent trends. The ranking of countries is most consistent in the context of extreme equality or inequality, and between measures from the same database, while the occurrence of full consistency in the values of individual measures practically does not occur.

Conference Papers

XXXXII Multivariate Statistical Analysis 2024, Lodz, Poland

Grzegorz Kończak's article ***On a new goodness-of-fit test for multivariate normality with fixed parameters based on David-Hellwig test idea*** presents a proposal for a goodness-of-fit test for multivariate normality. The idea of this test is based on the concept of empty cells. In conducting an empty cells test, it is crucial to partition the area of variation into disjoint cells, which is particularly significant in multivariate analysis. In the proposed testing procedure, cells are defined by confidence ellipsoids and are further segmented along the eigenvectors of the variance-covariance matrix.

This division of the area of variation ensures equal probabilities of observations in the cells under H0. The proposed test validates the conformity of the empirical distribution to a preset, precisely specified, multivariate normal distribution. Accordingly, it reacts not only to changes in the parameters of the distribution, but also to deviations from normality, in particular to the occurrence of a distribution type other than normal.

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

In the paper ***The concept of behavioral model of decision-making under risk*,** **Ewa Falkiewicz** outlines the decision-making process under risk and considers the psychological aspects of the decision-maker. The aim is to construct a principle of optimal decision choice for a single decision-maker. A finite, discrete set of acceptable decisions, a set of possible world states, and a system of probabilities of these states, whose probabilities are either known or subjectively estimated by the decision-maker, and a utility matrix are considered. A process of optimizing the decision-making is suggested, taking into account not only the rationality of the decision-maker but also emotional aspects. Taking into account two emotions important to decision-making – regret at making a decision bringing less utility than possible in given conditions and satisfaction with a choice that is better than the weakest – constitute the behavioral part of the model.

Research Communicates and Letters

In the paper entitled ***Mean estimation based on the factor-type estimator under an adaptive cluster sampling design*** by **Narendra Singh Thakur**, **Shubhangi Chaurasia**, and **Unnati Bhayare,** some estimators are discussed with their properties using the concept of large sample approximations in adaptive cluster sampling. This manuscript emphasizes the use of the factor-type estimator designed for population mean of the variable under study using the data of highly correlated auxiliary (supplementary) variable under adaptive cluster sampling. The bias, mean squared error and optimum mean squared errors up to the first order are obtained and a simulation study is performed for comparison purpose. The condition of optimality is derived as well.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 1–14, https://doi.org/10.59139/stattrans-2025-036
Received – 30.12.2025; accepted – 09.06.2025

On using ARIMA model confidence intervals applied to population projections based on the components of change: a case study for the world population

David A. Swanson[1], **Jeff Tayman**[2]

Abstract

This paper shows how measures of uncertainty from a standard time series model (ARIMA) can be applied to an existing population projection based on components of change using the world as a case study. The measures of forecast uncertainty are relatively easy to calculate and meet several important criteria used by demographers who routinely generate population forecasts. This paper applies the uncertainty measures to a world population forecast based on the Cohort-Component Method. This approach links the probabilistic world forecast uncertainty to the fundamental demographic equation, the cornerstone of demographic theory, which is an important consideration in developing accurate forecasts. The results are compared to the Bayesian probabilistic world forecast developed by the United Nations and found to be similar but show more uncertainty. The results are followed by a discussion suggesting that this new method is well-suited for developing probabilistic world, national, and sub-national population forecasts.

Key words: ARIMA, Bayes, Espenshade-Tayman method, forecast uncertainty, super population.

1. Introduction

Alkema *et al.* (2015) describe a Bayesian approach that links probabilistic uncertainty to a world population forecast based on the Cohort-Component Method (CCM). It proceeds by assembling a large sample of future trajectories for an outcome such as the total population size. The point projection in a given year is the median outcome of the sample trajectories. Other percentiles are used to construct prediction intervals (Alkema *et al.*, 2015). More details on this seminal approach are found in Raftery, Alkema, and Gerland (2014), and a general overview of probabilistic population forecasting can be found in Raftery and Ševčíková (2023).

[1] Department of Sociology, University of California Riverside, 900 University Avenue, Riverside, CA 92521, USA, E-mail: dswanson@ucr.edu. ORCID: https://orcid.org/0000-0003-4284-9478.

[2] Tayman Demographics, 2142 Diamond Street, San Diego, CA 92109, USA., E-mail: jtayman@san.rr.com. ORCID: https://orcid.org/0000-0003-3572-209X.

Because the Bayesian approach described by Alkema *et al.* (2015) is based on the CCM, its measures of uncertainty are linked to the "fundamental equation", whereby a population at a given point in time, P_{t+k}, is equal to the population at an earlier point in time, P_t, to which is added the births and in-migrants that occur between time t and time t+k and to which is subtracted the deaths and out-migrants that occur during this same time period (Baker *et al.*, 2017, pp. 251–252). The fundamental equation is the cornerstone of demographic theory and is the foundation upon which the CCM rests (Baker *et al.*, 2017, pp. 22–23; Burch, 2018; Verma, 2023). A probabilistic approach to population forecasting based on this theoretical foundation yields benefits not found in methods lacking this foundation (e.g., Burch, 2018; Land, 1986). This observation is also consistent with one made by Swanson *et al.* (2023), who argue that a given population forecasting method's strengths and weaknesses largely stem from four sources: (1) its correspondence to the dynamics by which a population moves forward in time; (2) the information available relevant to these dynamics; (3) the time and resources available to assemble relevant information and generate a forecast; and (4) the information needed from the forecast. The Bayes CCM approach comes with strengths. However, it also comes with weaknesses. Goodwin (2015) finds Bayesian inference difficult, effortful, opaque, and even counter-intuitive. Along with the weaknesses described by Goodwin (2015) are implied ones, including being not easy to apply or explain and having a low face validity and high production costs in that a Bayes CCM approach is very data- and analytically intensive.

2. Objective

In view of the facts described in the preceding section, we offer an approach for constructing uncertainty measures that is relatively simple and linked directly to the CCM approach. Importantly, unlike Bayesian inference, we believe it is likely to meet important evaluation criteria used by demographers who routinely develop population forecasts (Smith, Tayman, and Swanson, 2013, pp. 301–322); low production costs (particularly staff time); easy to apply and easy to explain; a high level of face validity; and intuitive.

3. Methods and Data

In describing this new approach, we use a world population forecast with a horizon of 2060. It is found at the *International Data Base* (IDB) site of the U.S. Census Bureau (https://www.census.gov/data-tools/demo/idb/#/table?COUNTRY_YEAR=2024&COUNTRY_YR_ANIM=2024&menu=tableViz). The data and methods are documented in U.S. Census Bureau (2020).

The approach we suggest employs the ARIMA (Auto-Regressive Integrated Moving Average) time series method in conjunction with work by Espenshade and Tayman (1982), whereby we can translate the uncertainty information found in the ARIMA method's forecast to the population forecast provided by the CCM approach. We describe neither the ARIMA (Box and Jenkins, 1976) nor the CCM approach (Smith, Tayman, and Swanson, 2013, Chapter 7) in detail because they are widely known and used. However, as described by Smith, Tayman, and Swanson (2001, pp. 172–176), an ARIMA model attempts to uncover the stochastic processes that generate a historical data series. The mechanism of this stochastic process is described—based on the patterns observed in the data series—and that mechanism forms the basis for developing forecasts. Up to three processes can represent the stochastic mechanism: autoregression, differencing, and moving average. The most general ARIMA model is usually written as ARIMA (p, d, q), where p is the order of the autoregression, d is the degree of differencing, and q is the order of the moving average.

In regard to this case study, the patterns of the autocorrelation (ACF) and partial autocorrelation functions (PACF) were used to find the correct values for p and q (Brockwell and Davis, 2016: Chapter 3). The ARIMA model shown here had random residuals and the smallest possible values for p, d, or q, as determined by the Ljung-Box test (Ljung and Box, 1979). We chose an "adequate" ARIMA model using these criteria. We note that there may be other versions that also are "adequate" and that further refinement of the selection process can be done (e.g., using the augmented Dickey-Fuller test (Dickey and Fuller, 1979) to identify the amount of differencing required to achieve a stationary time series). Because our aim here is heuristic and not definitive, we did not pursue further refinement of the ARIMA model we present beyond determining it to be adequate.

Before turning to a description of the new method, we first clarify our use of the term "confidence interval" in regard to forecast uncertainty. It is more common to use the term "forecast interval" or "prediction interval" in the context of forecasting because a "confidence interval," strictly speaking, applies to a sample (Swanson and Tayman, 2014, p. 204). However, underlying the approach we describe herein is the concept of a "super-population," which, as discussed later, describes a population that is but one sample of the infinity of populations that will result by chance from the same underlying social and economic cause systems (Deming and Stephan, 1941). The concept of viewing a forecast as a sample leads us to choose the term "confidence interval" rather than forecast interval or prediction interval.

We use annual world historical data of total population and land area in square meters to compute population density annually from 1950 to 2020 found at the IDB site to implement the ARIMA model found in the NCSS statistical package (NCSS, 2024) and launch from the annual world forecasts found at the same site for 2021–2060.

Exhibit 1 contains the NCSS output and report on the ARIMA model we use. We use "density" because the Espenshade-Tayman (1982) method for translating uncertainty information does so from an estimated "rate," which in this case is the "rate" of population density. Thus, the 95% confidence intervals generated by the ARIMA world "density" forecasts are translated to the CCM-based world population forecast. Other denominators could be used in developing this "rate" such as the ratio of the population to housing units. However, using the land area as the denominator provides a virtually constant denominator over time, thereby reducing the effort in assembling the "rate" data. It also serves as a stabilizing element regarding the use of ARIMA in that it dampens the effect of short-term population fluctuations more effectively than, say, housing units, which also can fluctuate over time and not always in concert with population fluctuations. As should be obvious, the data assembled to develop the ARIMA density forecast should encompass the base data used to develop the population projection in terms of the total population numbers. The case study we present meets this condition in that the ARIMA model covers the annual period from 1950 to 2020 and the population projection data use the total 2020 population, supplemented by earlier data in the examination of trends.

Exhibit 1 About Here

Here is an example of this process using the 2050 world population projection result found at the IDB site.

Let P = projected world population (at time t_i)

Let D = forecasted world population density obtained from ARIMA at time t_i, and

Let A = land area of the world (131, 821, 645 square kilometers).

The 2050 ARIMA density forecast shows 73.02, 76.81, and 80.60 persons per square kilometer, respectively, for the land area of the world as a whole (95% Lower Limit of forecasted D, forecasted D, and 95% Upper Limit of forecasted D, respectively).

The relative widths of the Lower and Upper Limits are -0.04938 and 0.04938, respectively.

The 2050 world population projection found at IDB is 9.7 billion.

Multiplying 9.75 billion by -0.04938 and adding this product to 9.75 billion yields 9.27 billion, the 95% Lower Limit, and adding the product 9.75 billion × 0.04938 to 9.7 billion yields 10.23 billion, the 95% Upper Limit of the 2050 world population forecast found at IDB.

Putting it all together, we can state that we are 95% certain that the 2050 world forecast found at IDB is between 9.27 billion and 10.23 billion.

As alluded to earlier, underlying the Espenshade-Tayman method is the idea that a sample is taken from a population of interest. In this case, the ARIMA results represent the sample, and the CCM forecasts represent the population. This interpretation is derived from the idea of a "super-population" (Hartley and Sielken, 1975; Sampath, 2005; Swanson and Tayman (2012, pp. 32–33). This concept can be traced back to Deming and Stephan (1941), who observed that even a complete census, for scientific generalizations, describes a population that is but one of the infinity of populations that will result by chance from the same underlying social and economic cause systems. It is a theoretical concept that we use to simplify the application of statistical uncertainty to a population forecast that is considered a statistical model in this context. This approach is conceptually and mathematically different from the classical frequentist theory of finite population sampling (Hartley and Sielken (1975)), but as pointed out by Ding, Li, and Miratrix (2017), in practical terms, these two approaches result in identical variance estimators. As such, we believe that our approach is on solid statistical ground. Before moving on, we also note that using the Espenshade-Tayman method (1982) here is not new. In addition to being employed by Espenshade and Tayman (1982), it has been used by Swanson (1989) and Roe, Swanson, and Carlson (1992) in demographic applications.

4. Results

Table 1 provides population forecast uncertainty measures for the world population from 2020 to 2060 by decade. The table contains three panels. Panel A provides the ARIMA-generated world population density point forecast and their 95% lower and upper limits. Panel B provides the relative widths of these intervals, which measure the proportionate differences between the upper and lower bounds and the point forecast. Panel C provides the IDB world point forecast along with the 95% CIs the Espenshade-Tayman approach has generated for them.

The relative bounds shown in Panel B, ignoring the sign on the LL95%, are analogous to the half-width that measures interval width (Tayman, Smith, and Lin, 2007), but half-widths are expressed as percentages and not proportions. As the name suggests, a half-width is ½ the width of the confidence interval, and we define this measure as half-width = ((MOE/2) / (projection)) × 100. As expected, the confidence intervals around the IDB point forecast increase in width as the forecast horizon length increases. The half-width for the 10-year horizon (2030) is 1.6% and rises steadily, reaching 6.3% for the 40-year horizon (2040).

Table 1 About Here

As shown in Table 1, the 2050 world forecast found at IDB is 9.75 billion with a 95% lower bound of 9.27 billion and a 95% upper bound of 10.23 billion. Under its medium scenario, the United Nations (UN) expects a world population in 2050 of 9.7 billion, with a 95% lower bound of 9.4 billion and a 95% upper bound of 10.0 billion (United Nations, 2022a: 28). The IDB-based point forecast for 2050 is 0.5% higher than the UN's medium scenario point forecast for 2050, where 0.5% = 100 × (9.75 – 9.70) / 9.70. The lower bound for the IDB-based 2050 forecast differs by 1.4% from the lower bound of the UN's medium scenario 2050 forecast, where 1.4% = 100 × (9.40 – 9.27) / 9.27 and by -2.25% from the upper bound, where -2.25% = 100 × (10.0 – 10.23) / 10.23. In other words, for 2050, the UN 95% intervals are narrower than the intervals produced from the IBF using the ARIMA process. The range in 2050 for the UN interval is six hundred million persons, 37 percent lower than the range from the IDB forecast paired with the ARIMA density model (960 million persons).

Although the specific 95% confidence intervals have not been discussed for years other than 2050 in the UN report on its probabilistic world population forecasts, they are available in an Excel file (UN, 2024). We used them to produce "half-widths" for the 2030, 2040, 2050, and 2060 UN forecasts. They are, respectively, 0.53%, 1.49%, 2.51%, and 3.76%; the half widths that we constructed of the IDB forecasts for 2030, 2040, 2050, and 2060 are, respectively, 1.60%, 3.35%, 4.94%, and 6.28%. Although the difference narrows between 2030 and 2060, the uncertainty intervals we constructed for the IDB forecasts are wider than those found for the UN forecasts. However, like the UN uncertainty intervals, our constructed ones increase over time. Between 2030 and 2060, the uncertainty interval we constructed for the IDB forecasts increases almost fourfold while the UN's increase is sevenfold.

5. Discussion

As is the case with the Bayesian approach described by Alkema *et al.* (2015), the new approach we propose can be linked directly to the CCM method (as well as forecasts produced by other methods such as the Cohort Change Ratio (CCR) approach, which is algebraically equivalent to the CCM approach, but requires less input (Baker et al., 2017, pp. 251–252)). Unlike the approach found in Swanson and Beck (1994), neither the CCM nor the CCR approach is inherently conjoined with a method for generating statistical uncertainty. Thus, we believe this linkage represents a step toward generating probabilistic forecasts based on the fundamental population equation. Notably, the ARIMA method is widely available in the software packages generally used by demographers.

Considering the discussion of data assembly and analysis (Alkema *et al.*, 2015; United Nations, 2022; Yu *et al.*, 2023), it is clear that far more time and resources are required for a Bayesian probabilistic forecast than for the probabilistic forecasting method we have described here. In addition, as characterized by Goodwin (2015), Bayesian inference is difficult, effortful, opaque, and even counter-intuitive, none of which applies to the method we have described in this paper. Beyond Goodwin's (2015) observations, Green and Armstrong (2015) discuss simple versus complex methods in terms of forecasting, which applies here in that the approach we describe falls more into the simple methodological category rather than the complex category. They suggest that while no evidence shows complexity improves accuracy, complexity remains popular among (1) researchers because they are rewarded for publishing in highly ranked journals, which favor complexity; (2) methodologists because complex methods can be used to provide information that supports decision makers' plans; and (3) clients who may be reassured by incomprehensibility.

The ARIMA model (ARIMA, (1,1,0)) we selected (see Exhibit 1) is one of several "adequate models" we examined. We found that the other two, ARIMA (1,1,1) and ARIMA (1,2,0), produced uncertainty intervals that varied from the model we selected but were consistent overall in that their 95% confidence intervals also increased over time, which brings up a point of interest. Swanson and Tayman (2014) suggest that 66% intervals may be preferable to 95% confidence intervals because the latter produces intervals that may be too wide to be useful.

The approach we propose does not produce the uncertainty intervals by age and gender, as does the Bayes CCM approach described by Alkema *et al.* (2015), Yu *et al.* (2023, p. 934) and the CCR approach discussed by Swanson and Tayman (2014). The Bayes CCM approach also produces intervals for births, death, and migration. However, neither the Bayes CCM nor our approach take into account uncertainty in the input data themselves. However, as Yu et al. (2023, p. 934) implied, these are not likely to be among the most important sources of uncertainty for data in the United States and other countries where population forecasts are routinely produced.

In regard to our approach not providing uncertainty intervals by age and gender, Deming's (1950, pp. 127–134) "error propagation" was used to translate uncertainty in age group intervals found in the regression-based CCR forecasts reported by Swanson and Tayman (2014) to the total populations in question. In different forms, "error propagation" has been used by Alho and Spencer (2005), Espenshade and Tayman (1982), and Hansen, Hurwitz, and Madow (1953), among others. It may be possible to reverse-engineer error propagation and develop uncertainty measures by age and gender using our approach. The validity of this could be explored to determine if it is viable. As an approximation, one could generate age uncertainty intervals by controlling "low" and "high" numbers in a given forecast series to their corresponding

95% lower and upper limits, respectively, found using our proposed approach. The United Nations, for example, publishes not only probabilistic population forecasts but also a medium, low, and high series (United Nations, 2022).

Another possibility is to generate ARIMA forecasts of the ratio of land area to the population in a given age group for all age groups, which would generate probability intervals by age. These could be summed using the error propagation method to obtain a probability interval for the total population. In turn, this "bottom-up" result could be compared to the probability generated for the population as a whole.

6. Conclusion

Smith, Tayman, and Swanson (2001, p. 373) opined that future research would focus increasingly on measuring uncertainty in population forecasts and noted that while such research may not directly improve forecast accuracy, it will enhance our understanding of the uncertainty inherent in population forecasts. They went on to note that this change will imply a shift from "population projections" to "population forecasts," a guideline we have followed in this paper. If one is trying to decide between a Bayesian approach to developing a probabilistic forecast and the approach described in this paper, the strengths and weaknesses discussed here need to be considered carefully and in the context of a particular forecasting environment: if it is resource-challenged, constrained by deadlines, and exposed to stakeholders who are not trained in statistical methods, the Bayesian approach may not be suitable.

In closing, we argue that the approach we propose and have described in this paper is well-suited to generate probabilistic world population forecasts and national and subnational population forecasts where CCM and CCR methods are routinely used to produce them. In the case of national and sub-national population projections, remember that migration will play a role, which, in conjunction with smaller populations, will likely lead to higher levels of uncertainty.

References

Alho, J., Spencer, B., (2005). Statistical demography and forecasting. *Springer B. V*, Press. Dordrecht, Heidelberg, London, and New York

Alkema, L., Garland, P., Raftery, A. and Wilmoth, J., (2015). The United Nations probabilistic population projections: An introduction to demographic forecasting with uncertainty. *Foresight (Colch)*, 37, pp. 19–24.

Baker, J., Swanson, D. A., Tayman, J. and Tedrow, L., (2017). Cohort change ratios and their applications. *Springer B.V. Press*, Dordrecht, Heidelberg, London, and New York.

Box, G., Jenkins, G., (1976). Time Series Analysis – Forecasting and control, San Francisco, *CA: Holden-Day.*

Brockwell, P. J., Davis, R. A., (2016*).* Introduction to time series and forecasting, 3rd edition, *Springer Texts in Statistics*, Switzerland.

Burch, T., (2018). Model-based demography: Essays on integrating data, technique, and theory. Demographic Research Monographs. *Springer B.V. Press*, Dordrecht, Heidelberg, London, and New York

Deming, W. E., (1950). Some theory of sampling. New York, *NY: Dover Publications.*

Deming, W. E., Stephan, F., (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36(213), pp. 45–49.

Dickey, D. A., Fuller, W. A., (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366), pp. 427–431.

Ding, P., Li X. and Miratrix, L., (2017). Bridging finite and super population causal inference. *Journal of Causal Inference*, 5(2), 20160027, https://doi.org/10.1515/jci-2016-0027.

Espenshade, T., Tayman, J., (1982). Confidence intervals for postcensal population estimates. *Demography*, 19(2), pp. 191–210.

Goodwin, P., (2015). When simple alternatives to Bayes formula work well: Reducing the cognitive load when updating probability forecasts. *Journal of Business Research*, 68, pp. 1686–1691.

Green, K., Armstrong, J., (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, 68, pp. 1678–1685.

Hansen, N., Hurwitz, W. and Madow, W., (1953). Sample survey methods and theory, Volume I, methods and applications. New York, *NY: John Wiley and Sons* (re-published in 1993).

Hartley, H., Sielken, R., (1975). A “super-population viewpoint” for finite population sampling. *Biometrics*, 31(2), pp. 411–422.

Land, K., (1986). Methods for National population forecasts: A review. *Journal of the American Statistical Association*, 81 (December), pp.888–901.

Ljung, G., Box, G., (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), pp. 297–303.

McNown, R., Rogers, A. and Little, J., (1995). Simplicity and complexity in extrapolative population forecasting models. *Mathematical Population Studies*, 5, pp. 235–257.

NCSS, (2024). ARIMA (Box-Jenkins), https://www.ncss.com/software/ncss/time-series-and-forecasting-in-ncss/#ARIMA.

Pflaumer, P., (1992). Forecasting U.S. population totals with the Box-Jenkins approach. *International Journal of Forecasting*, 8, pp. 329–338.

Raftery, A., Ševčíková, H., (2023). Probabilistic population forecasting: Short to very long term. *International Journal of Forecasting*, 39, pp. 73–97.

Raftery, A., Alkema, L. and Gerland, P., (2014). Bayesian population projections for the United Nations. *Statistical Science*, 29(1), pp. 58–68.

Roe, L., Swanson, D. A. and Carlson, J., (1992). A variation of the housing unit method for estimating the population of small, rural areas: A case study of the local expert procedure. *Survey Methodology*, 18(1), pp. 155–163.

Sampath, S., (2005). Sampling theory and methods. *Alpha Science International Ltd. Harrow*, England.

Smith, S., Tayman, J. and Swanson, D. A., (2001). State and local population projections: Methodology and analysis. *Kluwer Academic/Plenum Press: New York.*

Smith, S., Tayman, J. and Swanson, D. A., (2013). A practitioner's guide to state and local population projections. *Springer B.V. Press.* Dordrecht, Heidelberg, London, and New York.

Swanson, D. A., (2019). Hopi tribal population forecast. Report prepared for the Apache County Superior Court of the state of Arizona, in re: The General Adjudication of all rights to use water in the Little Colorado River System and Source. *CIVIL NO.* pp. 6417–203.

Swanson, D. A., (1989). Confidence intervals for postcensal population estimates: A case study for local areas. *Survey Methodology*, 15(2), pp. 271–280.

Swanson, D. A., Beck, D., (1994). A new short-term county level projection method. *Journal of Economic and Social Measurement*, 20, pp. 25–50.

Swanson, D.A., Tayman, J., (2012). Subnational population estimates. *Springer B.V.* Press. Dordrecht, Heidelberg, London, and New York.

Swanson, D.A., Tayman, J., (2014). Measuring uncertainty in population forecasts: A new approach, pp. 203–215 in Marco Marsili and Giorgia Capacci (eds.)

Proceedings of the 6th EUROSTAT/UNECE Work Session on Demographic Projections. *National Institute of Statistics*, Rome, Italy.

Swanson, D. A., Bryan, T., Hattendorf, M., Comstock, K., Starosta, L. and Schmidt, R., (2023). An example of combining expert judgment and small area projection methods: Forecasting for Water District needs. *Spatial Demography*, 11(8), https://doi.org/10.1007/s40980-023-00119-3.

Tayman, J., Smith, S. and Lin, J., (2007). Precision, bias, and uncertainty for state population forecasts: An exploratory analysis of time series models. *Population Research and Policy Review*, 26(3), pp. 347–369.

United Nations, (2022). *World Population Prospects 2022: Summary of Results*. Department of Economic and Social Affairs, Population Division, UN DESA/POP/2022/TR/NO. 3.

United Nations, (2024). File PPP/POPTOT: Probabilistic projection of total population (both sexes combined) by region, subregion, country or area, 2024–2100 (thousands), Department of Economic and Social Affairs, Population Division, UN, https://population.un.org/wpp/Download/Standard/MostUsed/.

U.S. Census Bureau, (2020). International Data Base: Population estimates and projections methodology. International Programs, Population Division, https://www2.census.gov/programs-surveys/international-programs/technical-documentation/methodology/idb-methodology.pdf.

Verma, R., (2023). Review of the Quality of Population Estimates and Projections at Sub-national Level in India Using Principles of Applied Demography. *Indian Journal of Population and Development*, 3(2), pp. 319–336, http://www.mlcfoundation.org.in/#assets/ijpd/2023-2/V_3_2_5.pdf.

Yu., C., Ševčíková, H. Raftery, A. and Curran, S., (2023). Probabilistic county-level population projections. *Demography*, 60(3), pp. 915–937.

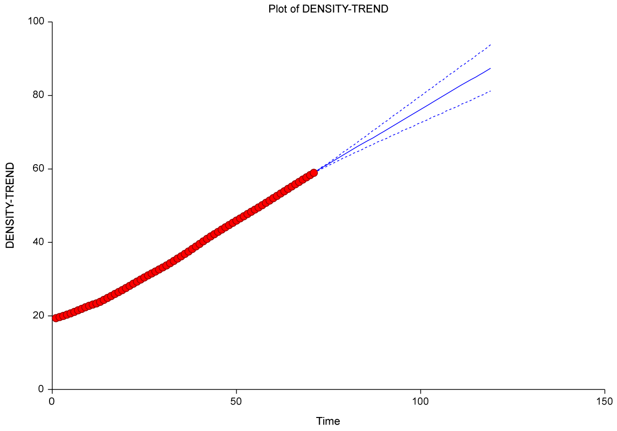
Zakria, M., Muhammad, F., (2009). Forecasting the population of Pakistan using ARIMA models. *Pakistan Journal of Agricultural Sciences*, 46(3), pp. 214–223.

Appendix

EXHIBIT 1. NCSS ARIMA (1,1,0) Report

```
Dataset            ...\WORLD DENSITY 1950-2060.NCSS
Filter             YEAR<2021
Variable           DENSITY-TREND


Minimization Phase Section
______________________________________________________________________
Normal convergence.


Model Description Section
______________________________________________________________________
Series                          DENSITY-TREND
Model                           Regular(1,1,0)    Seasonal(No seasonal
parameters)
Trend Equation                  (16.53608)+(0.5865935)x(date)

Observations                    71
Missing Values                  None
Iterations                      19
Pseudo R-Squared                99.999524
Residual Sum of Squares         0.04902241
Mean Square Error               0.0007104698
Root Mean Square                0.02665464


Model Estimation Section
______________________________________________________________________
Parameter     Parameter       Standard                       Prob
Name          Estimate        Error           T-Value        Level
AR(1)         0.9672837       0.01841017      52.5407        0.000000
```

Forecast and Data Plot



Autocorrelation Plot Section

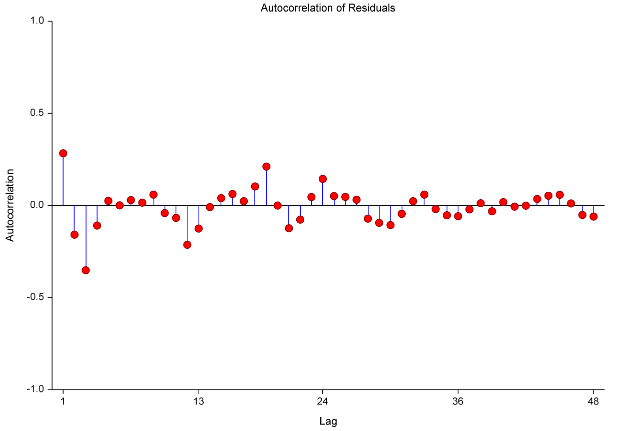


Table 1. Measures of Uncertainty for World Population Forecast, 2030–2060

Panel A. ARIMA Population Density Forecast

2030			2040			2050			2060		
LL95%	Forecast	UL95%	LL95%	Forecast	UL95%	LL95%	Forecast	UL95%	LL95%	Forecast	UL95%
63.87	64.92	65.96	68.50	70.88	73.25	73.02	76.81	80.60	77.53	82.72	87.92

Panel B. ARIMA Upper and Lower 95% Bounds Relative to Forecast[a]

2030		2040		2050		2060	
LL95%	UL95%	LL95%	UL95%	LL95%	UL95%	LL95%	UL95%
-0.01603	0.01603	-0.03355	0.03355	-0.04938	0.04938	-0.06277	0.06277

Panel C. ARIMA 95% Intervals Applied to IDB Forecast (in Billions)

2030			2040			2050			2060		
LL95%	IDB Forecast	UL95%	LL95%	IDB Forecast	UL95%	LL95%	IDB Forecast	UL95%	LL95%	IDB Forecast	UL95%
8.36	8.50	8.64	8.86	9.17	9.48	9.27	9.75	10.23	9.59	10.23	10.87

[a] (LL95% - IDB Forecast) / IDB Forecast and (UL95% - IDB Forecast) / IDB Forecast.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 15–34, https://doi.org/10.59139/stattrans-2025-037
Received – 23.03.2024; accepted – 18.02.2025

Factors determining the formation of degraded areas in local government units and the effectiveness of revitalisation activities

Agata Girul[1]

Abstract

Modern local government units form important links in the socio-economic structure of the country and their development is closely related to the occurrence of degraded areas. This study focused on identifying the social, economic and environmental factors that determine the occurrence of degraded areas requiring revitalisation in Polish local government units. The article used unit data from a Statistics Poland survey based on the SG-01 report: Municipal Statistics – Revitalisation[2]. The PROFIT (PROperty FITting) multidimensional scaling program for local level territorial units was applied. The program takes into account the delimitation of rural areas based on the typology of Functional Urban Areas. The results were visualised through perception maps. Calculations and figures were made in Statistica 13. The focus on the problematic areas revealed the variety of challenges faced by local government units in their revitalisation activities. The survey was thus complemented by an analysis of the results of the undertaken revitalisation projects. The comprehensive analysis of the factors causing the degradation of areas in local government units and the effects of the revitalisation may prove important tools for rural and urban policy makers and planners in developing effective local development strategies and may have an impact on the quality of life of residents.

Key words: revitalisation, degraded areas, PROFIT, quality of life.

1. Introduction

Each territorial unit has degraded areas that require revitalisation and restoration activities. The diagnosis of degraded areas, including the identification of the factors influencing their occurrence, is one of the first mandatory activities included in the revitalisation processes, because an important element of the revitalisation process is

[1] Wroclaw University of Economics and Business, Wroclaw, Poland. E-mail: agata.girul@ue.wroc.pl & Statistical Office in Wroclaw, Wroclaw, Poland. E-mail: a.girul@stat.gov.pl. ORCID: https://orcid.org/0000-0003-0172-2315.

[2] The report implemented by the municipalities is available at the following link: https://form.stat.gov.pl/formularze/2023/passive/SG-01-5.pdf

the precise identification of the intervention areas that would require changes first, given the limited financial resources and organisational capacity of the territorial units.

Revitalisation intended to counteract the phenomenon of degradation is a comprehensive process aimed at infrastructural and socio-economic revitalisation of units (Kowalczyk, 2017), but also economic revitalisation combined with activities aimed at solving social, spatial-functional, technical and environmental problems (The Law of ..., 2015). It can also be said that revitalisation is a complex process of multiple transformations aimed at improving the quality of space in terms of historical, architectural and environmental values (Leshchenko and Gulei, 2024). Thus, as part of revitalisation activities, local governments are tasked with, among other things, restoring degraded areas, reconstructing old buildings, promoting employment and economic activity by attracting new investors or expanding existing businesses, creating new jobs in crisis areas, and improving environmental quality (Stepina and Pelše, 2022). By carrying out revitalisation activities in cooperation with the local community, it is possible to bring degraded areas out of crisis through projects that integrate undertakings for the well-being of the local community, the area and the economy (Sych, 2020).

Revitalisation activities are an important initiative due to the fact that the existence of degraded areas is a common problem and a barrier to the socio-economic development of local units and has a negative impact on the environment (Kowalczyk, 2017, Stepina and Pelše, 2022). Reliable and high-quality diagnosis of degraded areas, followed by corrective activities against them, remains an important element of local development policy (Raszkowski and Sobczak, 2018). In addition, methods and tools to obtain information on the degradation and rehabilitation of areas, including land and the environment, are needed to support policies for sustainable ecosystem management and environmental protection to restore biodiversity and maintain its sustainability (Mao et al., 2018, Mebrat, 2015, Lamb et al., 2005, Xie et al., 2019, Ferreira et al., 2018). Therefore, there is a strong need for systematic and spatial measurement of land and soil degradation (Wessels et al., 2004). This is because the degradation of land results in its lower productivity (Prince et al., 2009).

Research on diagnosing the drivers of land degradation is becoming increasingly important due to global population growth and the expansion of urban areas, which is affecting the faster degradation of land. Increasing numbers of urban residents are migrating to rural areas, which also creates opportunities and challenges for rural revitalisation (Zheng et al., 2024, Turek et al., 2018). Abandoned and dilapidated land or degraded areas also emerge during deindustrialisation, leaving potentially contaminated and underutilised land and buildings in the hands of local governments (Kalnina and Pelše, 2023). Therefore, it is necessary to combat land and soil degradation at different levels and scales around the world, not only for food security and ecological health, but also to ensure global sustainable development (Jie, et al., 2002, Zambon et al., 2017) and to achieve a better quality of life for the inhabitants of territorial units (Kalnina and Pelše, 2023). It can be assumed that revitalisation should certainly be carried out in accordance with the principle of sustainable development (Makuch, 2020). Thus, it should meet the needs of society, respecting the requirements of environmental protection, without endangering the survival of future generations.

It is common for the revitalisation process in local government units to be carried out on the basis of a strategic document, such as a revitalisation programme, which sets out a framework for action and a schedule of activities. This programme has a significant impact on the development of policies to promote sustainable development, so it can be assumed that regeneration will become an important element of sustainable development.

The purpose of this article was an attempt to answer the research questions:

1) Which factors (intervention areas) most significantly determine the occurrence of degraded areas in certain types of municipalities in Poland?
2) What revitalisation activities are carried out by Polish local government units by area of intervention?

Without a clear diagnosis of the areas for intervention at the level of territorial units, it will not be possible to start the process of revitalising areas in crisis, nor will it be possible to achieve the strategic objectives of sustainable development. These considerations are a contribution to the discussion on the reasons for the formation of degraded areas in certain types of municipalities and to the analysis of the effects of revitalisation activities.

In order to answer the above research questions, individual data were applied from a survey conducted by Statistics Poland in 2023 on the basis of the SG-01 report: Municipal Statistics – Revitalisation[3]. The PROFIT (PROperty FITting) analysis was also used – one of the many multidimensional scaling programs that appeared on the market already in the first half of the 1980s (Gatnar, Walesiak, 2004), although in the Statistica 13 package, which was used for preparing statistical analyses and data visualisation, it was implemented in the Kit Plus add-on at the end of 2012. The application of the chosen method allowed the visualisation of the results by means of perception maps and the modelling of the properties of the objects (types of municipalities) in the context of intervention areas and negative phenomena that determine the occurrence of degraded areas due to a specific problem area.

[3] The report implemented by the municipalities is available at the following link: https://form.stat.gov.pl/formularze/2023/passive/SG-01-5.pdf.

2. Data and methods

In order to answer the research questions, the article used unit data from the survey conducted by Statistics Poland using the SG-01 report: Municipal Statistics – Revitalisation. The data collected in the survey made it possible not only to diagnose the causes of the formation of a degraded area, but also to assess the activity of the local government, the enterprises, the effects in the field of revitalisation and the planned financial resources for revitalisation initiatives. Information on degraded areas was provided by all municipalities, regardless of whether they were carrying out revitalisation activities in their area.

Using the report SG-01: Municipal Statistics – Revitalisation, the measurement of the causes of the formation of the degraded area, taking into account the five areas of intervention (social, economic, environmental, spatial-functional, technical) and the negative phenomena determining their occurrence, was determined by a 4-stage scale of intensity of problems:

- there were no problems of a certain type – value 0 was assigned,
- a low degree of difficulties was found to qualify the area as degraded – value 1,
- medium scale problems occurred to qualify the area as degraded – value 2,
- high degree of difficulties was found to qualify the area as degraded – value 3.

The definition of three thresholds of importance for the problems that occurred in the municipality made it possible to clearly identify which of these problems mostly contributed to the creation of a degraded area and which were of marginal importance.

The evaluation of negative factors leading to the occurrence of degraded areas is presented on the basis of data from 2022. The survey units were all municipalities in Poland (2477). Of these, 1488 provided information on the designation of a degraded area and 1440 municipalities carried out revitalisation activities to enhance a degraded area on the basis of a revitalisation document (municipal revitalisation programme, revitalisation programme, other strategic document).

Municipal units were analysed according to the type of municipality, but also according to the delimitation of rural areas, taking into account the typology of Functional Urban Areas (FUA). The following groups of areas were distinguished in the delimitation:

1. Agglomeration – rural areas within the FUAs of provincial cities or within the FUAs of other cities with at least 150,000 inhabitants:
   - high density agglomeration (Adg) – higher than the average population density in Poland,
   - low density agglomeration (Amg) – equal to or less than the average population density in Poland.

2. Non-agglomeration – rural areas outside the FUA boundaries of provincial cities or outside the FUA boundaries of other cities with a population of 150,000 or more:
   - high density non-agglomeration (pAdg) – with a population density greater than 1/3 of the population density in Poland,
   - low density non-agglomeration (pAmg) – population density equal to or less than 1/3 of the population density in Poland.

In the case of urban municipalities, a distinction was made between provincial cities and other cities with more than 150,000 inhabitants (M) and other cities with up to 150,000 inhabitants (pM). Twenty-five units were included in the group of provincial cities and other cities with more than 150,000 inhabitants. The average population density of Poland in 2022 was 120.8 inhabitants/km^2, and the threshold of 1/3 of the average population density was assumed to be 40.3 inhabitants/km^2.

The analysis, which took into account the delimitation of rural areas, did not consider urban-rural municipalities, as the SG-01 report did not provide data on urban and rural parts in urban-rural municipalities.

To answer the research questions, the PROFIT (PROperty FITting) analysis was applied, which uses two statistical techniques: multidimensional scaling to construct a classic perceptual map and multivariate regression. The main purpose of multidimensional scaling is to graphically represent the structure of similarity (or dissimilarity) between analysed objects with respect to a selected set of variables. Such a map, usually 2-dimensional or 3-dimensional (2-dimensional in this analysis), has a very simple interpretation. It is assumed that the smaller the distance between the objects studied, the more similar they are to each other. The result of multidimensional scaling is a plane (a space) on which the objects of interest are distributed. In the present analysis, these objects were the types of municipalities (Adg, Amg, pAdg, pAmg, M, pM) that characterize urban and rural municipalities.

In PROFIT analysis, multidimensional scaling aims to arrange objects in a way that simultaneously reduces the number of dimensions and reproduces the originally observed distances between objects as closely as possible.

The quality of the fit of the reconstructed data to the input data is measured by the STRESS function, which is most often defined as the square root of the standardised sum of squares of the residuals between the input distances and the distances reconstructed by multidimensional scaling. It takes the form:

$$\phi = \sqrt{\frac{\sum\sum(d_{ij}-f(\delta_{ij}))^2}{\sum\sum d_{ij}^2}} \quad (1)$$

where d_{ij} – denotes the reconstructed distance between points *i* and *j* on the perception map, δ_{ij} – denotes the distance between points *i* and *j* on the input data (observed distances), $f(\delta_{ij})$ – is a function defined on the input data where in the metric of multidimensional scaling it is assumed that $f(\delta_{ij}) = \delta_{ij}$.

The STRESS function is an indicator of the fit of the reconstructed data in the perception map to the input data. The smaller its value, the better the match between the reconstructed distance matrix and the observed distance matrix. It can be assumed that the perception map perfectly shows the observed distances when the STRESS function is close to 0.

In order to answer the question of how and in what direction the objects are arranged on a plane (in a space) due to the intensity of each of the input variables in the PROFIT analysis, multiple regression and estimation of model parameters were used, relating each variable to the position or the coordinates of objects on the perceptual map. These variables in the analysis presented in this paper were the intervention areas and the negative phenomena determining the occurrence of degraded areas due to a specific problem range (Appendix: Table A1).

The regression models were built on the basis of input data, which were averaged assessments of the causes of the degraded area, taking into account the assumed intervention areas and the negative phenomena that determine their occurrence. The coordinates assigned to the objects on the perception map were treated as independent variables in the regression model, and the averaged values of the individual variables for the given objects were treated as dependent variables. The number of regression equations constructed was therefore equal to the number of variables of the objects studied. After carrying out the regression analysis, the coordinates of the direction coefficients were superimposed on the previously constructed perceptual map. By projecting the points representing the individual objects (in this analysis, the types of municipalities) onto the vectors of the variables, it was possible to determine the position of the objects in relation to the intensity of these variables and thus to establish a preference series. The vector on the perception map pointed out the direction of increasing values of the variables analysed. Interestingly, the distance of a given object from the straight line on which the vector was located did not matter. What mattered was the ranking of the projections of the objects on these lines (Jabkowski, 2010).

When assessing the extent to which the ranking of objects in relation to the value of a given variable was explained by the position of these objects on the plane, the coefficients of determination of the regression equations were considered. The closer the value of the coefficient of determination is to 1, the better the fit.

The advantage of using the PROFIT scaling programme in this analysis was certainly the ability to present the results using perception maps. Although the PROFIT analysis is not the only method allowing to visualise the results on a diagram called a biplot, as such a diagram is also possible in the more popular Principal Component Analysis (PCA), it is worth noting that PROFIT is more oriented towards modelling specific variables and properties of objects in the context of data analysis, and not necessarily towards dimension reduction or identifying principal components as in PCA. On the other hand, both methods are useful for the hidden patterns' recognition.

All calculations and perception maps were made in Statistica 13 using the Kit Plus add-on and the PROFIT multidimensional scaling programme implemented in it.

3. Results

Using the PROFIT analysis, it was possible to diagnose the factors that have the greatest influence on the occurrence of degraded areas in certain types of local government units, based on the intervention areas included in the analysis. Perception maps were obtained with values of the STRESS function close to 0 and with acceptable levels of coefficient of determination for most of the variables greater than 0.7, due to the delimitation of rural areas taking into account the typology of Functional Urban Areas (Appendix: Table A2). The level of the coefficient of determination was considered acceptable to draw preliminary conclusions on the research topic in the cross-section of municipality types.

The results of the PROFIT analysis confirmed that the municipal units, due to the nature of the municipality, were not similar in their assessment of the impact of the analysed factors – the scatter of objects in the perception maps presented below for these objects was quite large.

The analysis also showed that in urban municipalities, social factors had the greatest impact on the classification of the area as a degraded area. This is particularly evident in provincial cities and other large cities with more than 150,000 inhabitants (the objects with symbols M and pM projected to the straight line Ob._S are ranked highest in the order hierarchy, Figure 1). The social sphere as a significant cause of the occurrence of degraded areas (rating 3 – high influence) was reported on average by 39% of all Polish local government units that identified a degraded area, and by 45% of urban local government units. In urban municipalities, environmental, technical and economic areas were important for the identification of degraded areas, in addition to the social area. These three areas were particularly problematic in cities with up to 150,000 inhabitants.

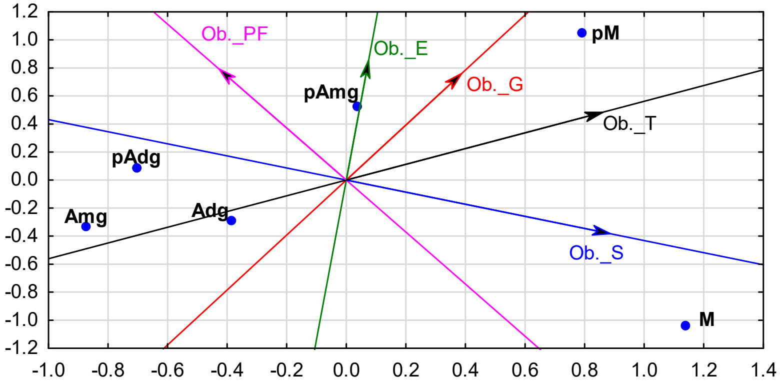


Figure 1. Perception map showing the results of the PROFIT analysis, including the areas of intervention and the delimitation of rural areas and the typology of functional urban areas

Source: own work based on the results of the SG-01 study: Municipal Statistics – Revitalisation of Statistics Poland.

The environmental, technical and economic areas were identified as having a high impact on the occurrence of degraded areas by about 25% of the urban agglomerations. Interestingly, the spatial-functional area had a greater impact on the identification of degraded areas in rural (agglomeration and non-agglomeration) municipalities than in urban ones – especially for cities with up to 150,000 inhabitants. For large provincial cities and other cities with 150,000 inhabitants or more, the spatial-functional area had little influence on the identification of degraded areas. In rural non-agglomeration municipalities (with low population density), environmental and economic factors were also determinants of the existence of degraded areas.

Based on the PROFIT analysis, it was also possible to diagnose specific factors that more or less influenced the qualification of an area as a deprived area in the corresponding types of municipalities.

In the case of the social area in urban municipalities, aspects of problems related to unemployment, poverty, crime, but also poor demographics, were the determining factors for the occurrence of degraded areas. These factors were more pronounced in provincial cities and other large cities with more than 150,000 inhabitants than in smaller cities with up to 150,000 inhabitants (Figure 2). In rural municipalities, on the other hand, the factors that were linked to the existence of degraded areas were issues related to the level of participation in public and cultural life, as well as the level of social activity and education. The last factor, the level of education, also influenced the identification of degraded areas in urban municipalities, but to a lesser extent than in rural municipalities.

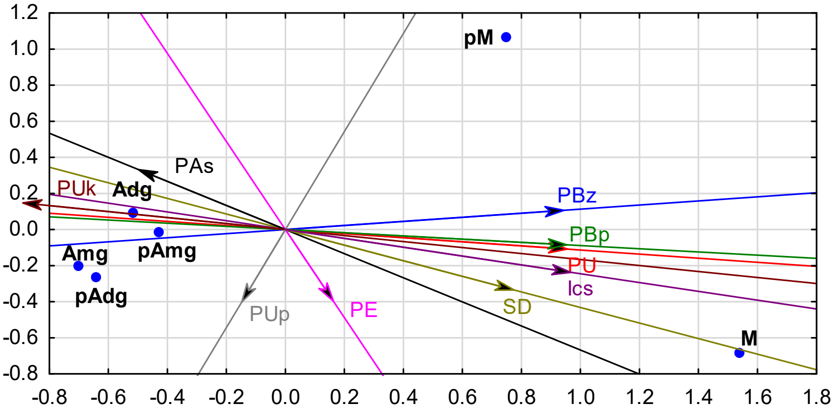


Figure 2. Perception map showing the results of the PROFIT analysis by social area of intervention and the delimitation of rural areas and the typology of functional urban areas

Source: own work based on the results of the SG-01 study: Municipal Statistics – Revitalisation of Statistics Poland.

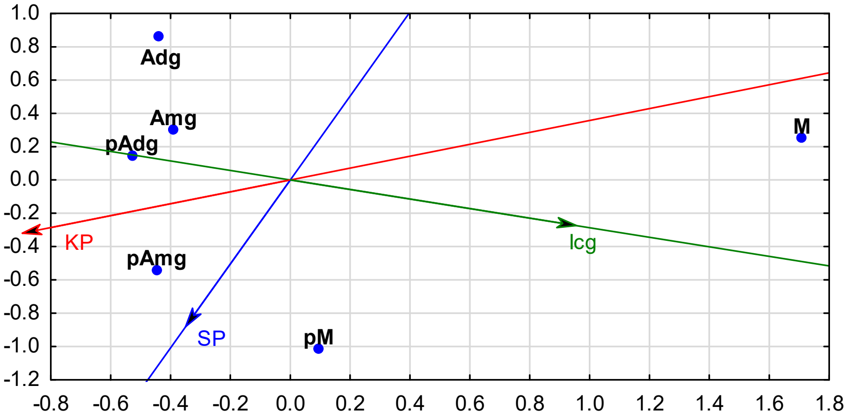


Figure 3. Perception map showing the results of the PROFIT analysis by economic area of intervention and the delimitation of rural areas and the typology of functional urban areas

Source: own work based on the results of the SG-01 study: Municipal Statistics – Revitalisation of Statistics Poland.

In the economic area, as in the social area, it was possible to observe different factors determining the presence of degraded areas in urban and rural municipalities. In urban municipalities, the degree of entrepreneurship turned out to be the most problematic economic factor, while in rural municipalities – the condition of local businesses. The problem related to the level of entrepreneurship was particularly acute in smaller towns of up to 150,000 inhabitants and in non-agglomeration municipalities with low population density, while the problem related to the condition of businesses was more pronounced in rural non-agglomeration municipalities than in agglomeration municipalities (Figure 3).

The environment was also an important area of intervention that could not be omitted from the analysis. In this area, it was found that in urban agglomerations, problems related to the exceedance of air quality and environmental quality standards, which affect smaller towns more than larger ones, were a factor in the occurrence of degraded areas in this area, while in rural agglomerations it was the presence of waste posing a threat to life, health or the environment (Figure 4).

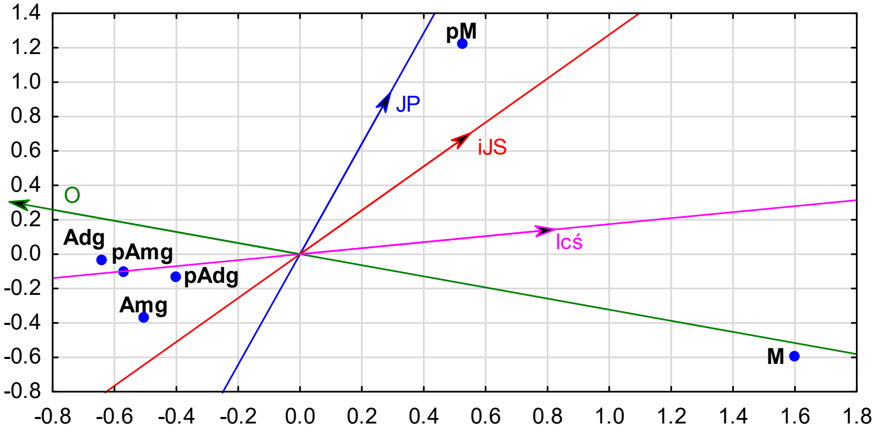


Figure 4. Perception map showing the results of the PROFIT analysis by environmental intervention area and the delimitation of rural areas and the typology of functional urban areas

Source: own work based on the results of the SG-01 study: Municipal Statistics – Revitalisation of Statistics Poland.

The spatial functional area was characterised by the greatest intensity of influence on the designation of a degraded area in rural municipalities. On the basis of the analysis of individual factors, it also turned out that the spatial functional aspects determining the classification of an area as degraded in rural municipalities were, in particular, issues related to the degree of accessibility of technical and social infrastructure and the level of adaptation of urban solutions to the current functions of the area, the quality and availability of public areas but also the re-development of brownfield sites. The latter type of factor was particularly important for the qualification of degraded areas in rural non-agglomeration municipalities with low population density. In urban municipalities, on the other hand, among the spatial functional factors influencing the occurrence of degraded areas, other unspecified spatial-functional factors can be identified, both for large and smaller cities with up to 150,000 residents (Figure 5).

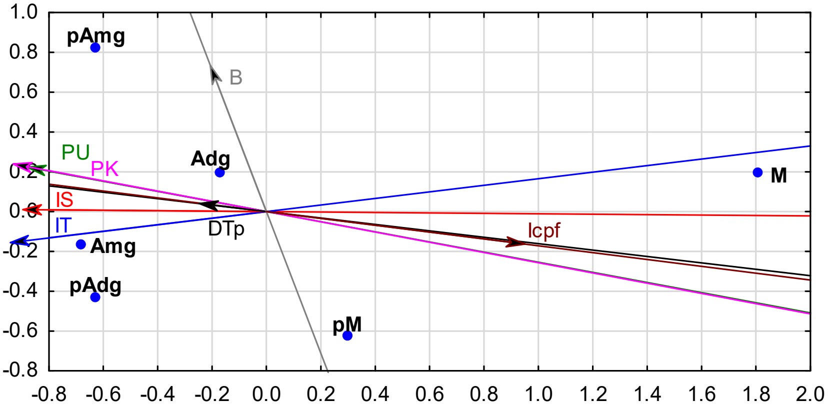


Figure 5. Perception map showing the results of the PROFIT analysis taking into account the spatial functional intervention area and the delimitation of rural areas and the typology of functional urban areas

Source: own work based on the results of the SG-01 study: Municipal Statistics – Revitalisation of Statistics Poland.

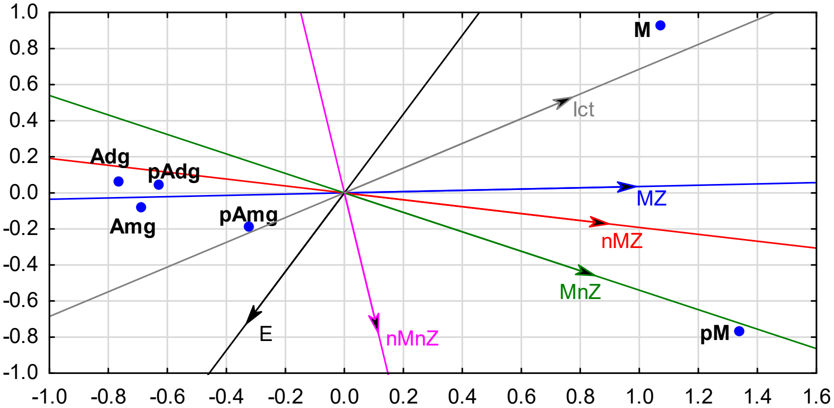


Figure 6. Perception map showing the results of the PROFIT analysis by technical area of intervention and the delimitation of rural areas and the typology of functional urban areas

Source: own work based on the results of the SG-01 study: Municipal Statistics – Revitalisation of Statistics Poland.

The last area of research was the technical area, which was particularly important for the qualification of degraded areas in urban municipalities, due to the significant influence of factors related to the structural condition of both residential and non-residential buildings, with or without historic monument status. Interestingly, the factor related to the functional technical solutions that enable the efficient use of buildings, e.g. in terms of energy efficiency and environmental protection, proved to have an impact on the diagnosis of degraded areas not only in urban municipalities, especially in relation to smaller cities with up to 150,000 inhabitants, but also in rural agglomeration and non-agglomeration municipalities (Figure 6).

Based on the analysis, it was possible to observe an ordering of factors for urban and rural municipalities on the two opposite sides of the coordinate axis, while the units characterizing urban-rural municipalities showed characteristics of both types of municipalities. This is due to the fact that urban-rural municipalities have one of the settlements in their area that has urban status and the rest of the area is rural. It can, therefore, be said that, in terms of the classification of degraded areas and the factors that determine the occurrence of intervention areas, urban-rural municipalities have similar characteristics to rural municipalities in social, economic, environmental and spatial functional areas, and to urban municipalities in the technical area.

4. Discussion and conclusion

Interest in the revitalisation of degraded areas has been growing steadily for several years. This is due to the desire of local governments to take advantage of the attractive locations of degraded areas, but also to the lack of a methodology for the revitalisation process that takes into account the specificity of degraded areas and the needs of the local community (Turek et al., 2018). Consequently, there is also a growing interest in spatial analysis or the use of satellite remote sensing to identify local causes of area degradation (Mao et al., 2018, Prince et al., 2009, Xie et al., 2019).

In the context of the process of enhancing degraded areas, it is also crucial for urban planners and policy makers, and especially for the socio-economic development of local government units, to have a thorough understanding of the factors influencing the occurrence of such areas on the territory of the municipalities concerned. The method applied in this study using the PROFIT multivariate analysis software proved to be a method that provides insight into the factors determining the qualification of areas as degraded for different types of municipalities. The identification of specific factors is a key step in developing solutions for individual units in terms of effective management of degraded areas and implementation of revitalisation processes on their territory. It is necessary to carry out research and analysis of the current situation in order to assess strengths and weaknesses, as well as to identify possible solutions for future actions in the process of infrastructural and socio-economic revitalisation of the units (Kalnina and Pelše, 2023).

On the basis of the survey carried out, it was possible to learn that urban and rural municipalities face a large number of identical problems in the context of the areas of intervention, but they tend to target different problem factors. Although the social area is mainly a problem of urban municipalities, especially large cities where the accumulation of problems such as high unemployment, poverty, crime or unfavourable demographic trends is enormous, some deficits in this field can also be observed in rural municipalities. In rural areas, there is an unsatisfactory level of social activity and participation in the public and cultural life of the inhabitants, which, according to local authorities, has an impact on the occurrence of degraded areas in these areas. It seems that social inclusion through social innovation activities could help to effectively combat rural marginalisation as a panacea to counteract social inequalities between urban and rural areas (Bock, 2016).

Interestingly, the largest number of revitalisation activities is in the social sector. In 2022, Polish municipalities with a revitalisation programme or a municipal revitalisation programme planned more than 32,000 revitalisation projects worth about PLN 80 billion, of which about 12,000 projects concerned the social area of intervention, worth about PLN 12 billion. The revitalisation effects related to the social sphere of intervention in 2022 in local government units included, among other things: almost 134,000 inhabitants were provided with social assistance, e.g. in the area of nutrition; vocational activation courses were held, from which almost 11,000 people benefited; more than 60,000 courses were held, which were addressed to various groups of recipients (the elderly, children and young people). The second area in which most revitalisation activities were carried out is the spatial-functional area. In 2022, municipalities planned about 9 thousand revitalisation activities in this area for a sum more than twice as high as in the social sector – more than PLN 27 billion.

In terms of different factors, the spatial and functional area proved to be important in the classification of degraded areas, especially for the self-governments of rural municipalities. It turned out that in rural areas, the lack of infrastructure or its poor technical condition, the lack of land development according to its intended use, the low level of transport services and the lack of urban planning solutions or the underutilisation of brownfield sites were elements that required revitalisation activities. The underutilisation of brownfield land is an aspect that has proved to be particularly important for rural non-agglomeration municipalities with low population density. In the case of urban municipalities, where the spatial-functional area was not characterized by a significant intensity of factors influencing the occurrence of degraded areas, only other spatial functional factors, previously unspecified, determining the occurrence of degraded areas were detected in larger and smaller cities. Although public space and its accessibility are changing for the better, especially in terms of green spaces and parks (Liu et al., 2024), there are still evident gaps in the satisfaction of needs in this area, especially observed in rural areas. In this intervention area, the municipal results of revitalisation projects in 2022 included 536 km of repaired and constructed roads, 138 km of constructed cycle paths, 209 ha of revitalised green spaces or 36 ha of revitalised brownfield sites, including 16 ha in rural municipalities.

Issues relating to the economic development of territorial units are a key element that should not be overlooked. Entrepreneurship is a driving force of the economy, but also an important factor in counteracting unfavourable socio-economic processes such as unemployment, exclusion or marginalisation. On the basis of the analysis it could be seen that in urban municipalities – in smaller towns – the factor influencing the diagnosis of degraded areas was too low a level of entrepreneurship, while in rural municipalities – mainly in units located outside towns – the poor condition of local businesses was the factor influencing the diagnosis of degraded areas. Business environment institutions, local governments should therefore stimulate and support entrepreneurs operating in rural areas. Cooperation between entrepreneurs and the implementation of development projects focusing on promoting regional diversity (e.g. tourism) and expanding the market for products and services should also play a key role (Yang et al., 2021). It is interesting to note that in 2022, municipalities planned 2,500 revitalisation projects from the economic intervention area with a total value of around PLN 6 billion. The effects of the revitalisation activities carried out in this area include, among others, the creation of almost 2,000 jobs or almost 2,000 business entities that have started operations in business premises in the revitalisation area.

Although there are studies confirming that local authorities in Poland do not perceive negative environmental phenomena as important factors in the assessment of problems in degraded areas, municipalities do include environmental issues in their local strategic programmes and perceive the need for revitalisation activities in this sphere (Jadach-Sepioło et al., 20–21). On the basis of this study, it was found that the environmental area is more relevant to urban municipalities than to rural ones, which does not mean that in rural areas it is not taken into account in the diagnosis of the establishment of intervention areas. In the case of urban municipalities, the factors determining the occurrence of degraded areas in relation to environmental problems were found to be aspects related to exceeding air quality standards or other environmental quality standards, and in relation to rural municipalities – the presence of waste that poses a threat to life, human health or the state of the environment. Rural municipalities should pay particular attention to activities that promote the implementation of proper waste management by raising the awareness of the local population of the need to protect the environment from waste.

In addition, local government units planned a total of more than 2,000 environmental revitalisation activities worth PLN 5.5 billion in 2022. Rural municipalities have planned more activities in this area (40% of the total) than urban municipalities (25%). The environmental effects of revitalisation include more than 3,000 buildings where asbestos was removed, including almost 2,000 in rural communes, and almost

5,000 flats where heat sources were replaced (e.g. by solar panels, gas heating), including almost 2,000 in rural units.

The last problem area included in the analysis was the technical area, where most factors had a greater influence on the delimitation of degraded areas in urban municipalities than in rural ones. Among the determinants influencing the occurrence of degraded areas in urban municipalities, especially those with up to 150,000 inhabitants, the poor technical condition of residential and non-residential buildings, including those with historical status, could be mentioned, as well as the non-functioning of technical solutions enabling the effective use of buildings, e.g. in terms of energy efficiency and environmental protection. The latter factor had a significant impact on the identification of degraded areas in rural municipalities. Once again, there is a need for municipalities in rural areas to take action to protect the environment. In 2022, more than 6,500 revitalisation projects were planned in the technical area, amounting to approximately PLN 16 billion, and among the revitalisation effects, the municipalities carried out works aimed at, among other things, improving the energy efficiency of 775 buildings or adapting about 300 buildings to the needs of people with disabilities.

The results of this study showed that local authorities are active in carrying out different types of regeneration in different areas of intervention. The study also pointed out the problematic factors influencing the qualification of degraded areas. Furthermore, it provided an insight into where financial resources and projects should be directed in order to mitigate the impact of the intensity of specific problem factors influencing the occurrence of degraded areas.

Finally, it should also be emphasised that the results of this study on the basis of unit data from the survey conducted by Statistics Poland using the SG-01 report: Municipal Statistics – Revitalisation, provided interesting conclusions, which encourage further research on this topic. The approach proposed in the article to identify problematic factors for the diagnosis of degraded areas, using the PROFIT analysis, may be of great importance for the implementation of effective revitalisation activities. It also seems reasonable to adapt the revitalisation process to the specificities of the territorial units. However, measures should be taken so that the units included in the SG-01 study for the urban-rural part are also presented with a breakdown between the rural part and the city. This would allow for more precise information, taking into account the delimitation of rural areas and the typology of Functional Urban Areas (FUAs). Research on the diagnosis of degraded areas in local government units should also be extended to include an analysis of the participatory activities of their inhabitants. The involvement of the public in the acceptance of changes in the intervention areas is an important element of revitalisation activities. The social reflections of the inhabitants of municipalities on the possibilities of renovating dilapidated or abandoned spaces can contribute to their quality of life and their civic attitude. Developing research in this area can influence local, national and European local development policies.

References

Bock, B., (2016). Rural marginalisation and the role of social innovation: a turn towards endogenous development and rural reconnection. *Sociologia Ruralis*, Vol. 56(4), pp. 552–573.

Ferreira, C., Walsh, R. and Ferreira, A., (2018). Degradation in urban areas. *Current Opinion in Environmental Science & Health*, Vol. 5, pp. 19–25.

Gatnar, G., Walesiak, M., (2004). Metody statystycznej analizy wielowymiarowej w badaniach marketingowych. *Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu*, Wrocław.

Jabkowski, P., (2010). O korzyściach wynikających z zastosowania analizy PROFIT. w: Praktyczna analiza danych w marketingu i badaniach rynku. *Wydawnictwo StatSoft Polska*.

Jadach-Sepioło, A., Olejniczak-Szuster, K. and Dziadkiewicz, M., (2021.) Does environment matter in smart revitalization strategies? Management towards sustainable urban regeneration programs in Poland. *Energies*, Vol. 14(15), No. 4482.

Jie, C., Jing-Zhang, C., Man-Zhi, T. and Zi-Tong, G., (2002). Soil degradation: a global problem endangering sustainable development. *Journal of Geographical Sciences*, Vol. 12(2), pp. 243–252.

Kalnina, M., Pelše, M., (2023). Overview of the current situation on degraded area management in Latvia. *23rd SGEM International Multidisciplinary Scientific GeoConference 2023*, Vol. 23, No. 5.1. pp. 439–446.

Kowalczyk, J., (2017). Polish model of urban renewal: formal and legal aspects of revitalisation in Polish cities. *Anuarul Institutului de Cercetări Socio-Umane, C.S. Nicolăescu-Plopşor*, No. XVIII/2017, pp. 145–158.

Lamb, D., Erskine, P., D. and Parrotta, J., A., (2005). Restoration of Degraded Tropical Forest Landscapes. *Science*, Vol. 310, pp. 1628–1632.

Leshchenko, N., Gulei, D. (2024). Complex revitalization of historically formed industrial territories in Kyiv in post-war recovery. *Journal of Architecture and Urbanism*, Vol. 48(1), pp. 1–10.

Liu, S., Tan, C., Deng, F., Zhang, W. and Wu, X., (2024). A new framework for assessment of park management in smart cities: a study based on social media data and deep learning. *Scientific Reports*, Vol. 14(1), No. 3630.

Makuch, K., (2020). The impact of the sustainable development principle on the revitalisation process under the act of 9 October 2015 on revitalization. *Nieruchomości@*, Vol. III(III), pp. 83–93.

Mao D., Wang, Z., Wu, B., Zeng, Y., Luo, L. and Zhang, B., (2018). Land degradation and restoration in the arid and semiarid zones of China: Quantified evidence and implications from satellites. *Land Degradation & Development*, Vol. 29(11), pp. 3841–3851.

Mebrat, W., (2015). Natural regeneration practice in degraded high lands of Ethiopia through area enclosure. *International Journal of Environmental Protection and Policy*, Vol. 3(5), pp. 120–123.

Prince, S., D., Becker-Reshef, I. and Rishmawi, K., (2009). Detection and mapping of long-term land degradation using local net production scaling: application to Zimbabwe. *Remote Sensing of Environment*, Vol. 113(5), pp. 1046–1057.

Raszkowski, A., Sobczak, E., (2018). Delimitation procedure of degraded areas and the area targeted for revitalisation. *Economics of the 21st century*, Vol. 2(18), pp. 30–38.

Stepina, M., Pelše, M. (2022), European Union funding support to Latvian municipalities for degraded areas revitalization. *Research for Rural Development 2022*, Vol. 37, pp. 233–239.

Sych, O., (2020). Revitalization as a component of urban strategy. *Bulletin of V. N. Karazin Kharkiv National University Economic Series*, No. 99, pp. 66–73.

The Law of 9 X 2015 on revitalization (uniform text *Journal of Laws 2024*, item 278).

Turek, A., Salach, A., Markiewicz, J., Maciejewska, A. and Zawieska, D., (2018). An example of multitemporal photogrammetric documentation and spatial analysis in process revitalisation and urban planning. *GISTAM 2018 – Proceedings of the 4th International Conference on Geographical Information Systems Theory, Applications and Management*, pp. 223–230.

Wessels, K., J., Prince, S., D., Frost, P., E., Zyl and D., V., (2004). Assessing the effects of human-induced land degradation in the former homelands of northern South Africa with a 1 km AVHRR NDVI time-series. *Remote Sensing of Environment*, 91, pp. 47–67.

Xie, Z., Phinn, S., R., Game, E., T., Pannell, D., J., Hobbs, R. J., Briggs, P., R. and McDonald-Madden, E., (2019). Using Landsat observations (1988–2017) and Google Earth Engine to detect vegetation cover changes in rangelands – a first step towards identifying degraded lands for conservation. *Remote Sensing of Environment*, Vol. 232, No. 111317.

Yang, J., Yang, R., Chen, M-H., Su, J., Zhi, Y. and Xi, J., (2021). Effects of rural revitalization on rural tourism. *Journal of Hospitality and Tourism Management*, Vol. 47(4), pp. 35–45.

Zambon, I., Colantoni, A., Carlucci, M., Morrow, N., Sateriano A. and Salvati, L., (2017). Land quality, sustainable development and environmental degradation in agricultural districts: A computational approach based on entropy indexes. *Environmental Impact Assessment Review*, Vol. 64, pp. 37–46.

Zheng, N., Wang, S., Wang, H. and Ye, S., (2024). Rural settlement of urban dwellers in China: community integration and spatial restructuring. *Humanities and Social Sciences Communications*, Vol. 11(1), No. 188.

Appendix

Table A1. Areas of intervention and factors identifying the degraded area

Full name	Abbreviation
SOCIAL AREA	**OB._S**
unemployment rate	PBz
poverty level	PU
the level of crime or other elements of public security	PBp
level of education	PE
level of social activity	PAs
level of participation in public life	PUp
the level of participation in cultural life	PUk
demographic situation	SD
other	Ics
ECONOMIC AREA	**Ob._G**
degree of entrepreneurship	SP
condition of local enterprises	KP
other	Icg
ENVIRONMENTAL AREA	**Ob._E**
exceeding of air quality standards	JP
exceeding of other environmental quality standards (e.g. noise)	iJS
presence of waste that poses a threat to life, human health or the state of the environment	O
other	Ics
SPATIAL FUNCTIONAL AREA	**Ob._PF**
degree of equipment in technical infrastructure and/or its technical condition	IT
the degree of equipment with social infrastructure and/or its technical condition	IS
the level of adaptation of urban planning solutions to the current functions of the area	PU
the level of communication services	PK
the quality of public areas and their accessibility	DTp
re-development of brownfield sites (post-industrial and other areas)	B
other	Icpf
TECHNICAL AREA	**Ob._T**
technical condition of residential buildings with monument status	MZ
technical condition of non-residential buildings with monument status	nMZ
technical condition of residential buildings without monument status	MnZ
technical condition of non-residential buildings without monument status	nMnZ
the functioning of technical solutions for the efficient use of buildings (in particular in terms of energy efficiency and environmental protection)	E
other	Ict

Source: own work based on the results of the SG-01 study: Municipal Statistics – Revitalisation of Statistics Poland.

Table A2. Results of the regression analysis and STRESS function

Variables	Results of the regression analysis				STRESS function
	constant term	coefficient of regression		coefficient of determination	
		β_1	β_2		
	by the delimitation of rural areas and the typology of FUA				
INTERVENTION AREA					
OB._S	1.88*	0.13*	-0.06*	0.93	0.0000041
Ob._G	1.23*	0.03	0.07*	0.73	
Ob._E	0.94*	0.01	0.09*	0.73	
Ob._PF	1.40*	-0.03*	0.07*	0.81	
Ob._T	1.15*	0.11*	0.07*	0.98	
SOCIAL AREA					
PBz	2.56*	0.16*	0.03	0.90	0.0000144
PU	2.50*	0.28*	-0.05	0.93	
PBp	2.12*	0.45*	-0.06	0.92	
PE	1.84*	0.01	-0.02	0.28	
PAs	1.96*	-0.07	0.07	0.45	
PUp	1.83*	-0.02	-0.07	0,28	
PUk	1.67*	-0.17*	0.04	0.81	
SD	2.12*	0.15*	-0.10	0.71	
Ics	0.27*	0.24*	-0.09*	0.99	
ECONOMIC AREA					
SP	2.07*	-0.04	-0.14*	0.90	0.0000014
KP	1.44*	-0.16*	-0.07	0.90	
Icg	0.17*	0.19*	-0.07*	0.99	
ENVIRONMENTAL AREA					
JP	1.32*	0.05	0.22*	0.97	0.000000
iJS	1.07*	0.03	0.05	0.80	
O	1.19*	-0.29*	0.13*	0.98	
Icś	0.20*	0.12	0.03	0.70	
SPATIAL FUNCTIONAL AREA					
IT	1.98*	-0.15*	-0.05	0.99	0.0000046
IS	1.84*	-0.08*	0.00	0.80	
PU	1.44*	-0.11*	0.05	0.82	
PK	1.57*	-0.13*	0.06	0.92	
DTp	1.71*	-0.02	0.01	0.26	
B	1.05*	-0.02	0.13	0.58	
Icpf	0.18*	0.13*	-0.04	0.93	
TECHNICAL AREA					
MZ	1.30*	0.24*	0.01	0.98	0.0000044
nMZ	1.25*	0.12*	-0.04	0.83	
MnZ	1.40*	0.18*	-0.17	0.93	
nMnZ	1.36*	0.01	-0.11	0.61	
E	1.42*	-0.04	-0.16	0.64	
Ict	0.20*	0.11*	0.13	0.88	

* Parameters statistically significant at the significance level $\alpha = 0.05$.

Source: own work based on the results of the SG-01 study: Municipal Statistics – Revitalisation of Statistics Poland.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 35–54, https://doi.org/10.59139/stattrans-2025-038
Received – 03.09.2024; accepted – 05.06.2025

Bayesian nonparametric model for weighted data using mixture of Burr XII distributions

Soleiman Khazaei[1], **Soghra Bohlourihajjar**[2]

Abstract

In this paper, we develop a Bayesian nonparametric approach for analyzing weighted survival data. Specifically, we employ the Dirichlet Process Burr XII Mixture Model (DPBMM) to estimate the underlying density and survival functions when the observed data are weighted. Parameters are inferred using Markov chain Monte Carlo (MCMC) methods, and the Metropolis-Hastings algorithm is applied to obtain de-biased samples from the weighted observations. Numerical illustrations are provided using both simulated and real lifetime data, including the presence of censored observations. The performance of the proposed method is compared with classical kernel density estimates to demonstrate its flexibility in modeling complex and heavy-tailed distributions.

Key words: Bayesian nonparametric, weighted data, Dirichlet process, mixture model, Burr XII distribution, survival data.

1. Introduction

Building upon the foundational ideas introduced by Fisher (1934), the concept of weighted distributions has been further developed. Rao (1985) and Rao et al. (1915) recognized the importance of a unified framework and identified a variety of sampling scenarios that could be effectively described using weighted distributions Patil (1978). Consider a non-negative random variable X with a natural density function $f(x;\theta)$, where $\theta \in \Theta$ denotes the natural parameter and Θ is the parameter space. A new random variable is then defined with a density function $g(x;\theta)$, specified as follows:

$$g(x;\theta) = \frac{w(x;\theta)f(x;\theta)}{E[w(X;\theta)]}, \quad E[w(X;\theta)] < \infty, \quad x \geq 0, \tag{1}$$

This new variable is referred to as a weighted random variable with respect to X, and $g(x;\theta)$ is called the weighted density function corresponding to $f(x;\theta)$. The function $w(x;\theta)$ is a non-negative function of x, and $E[w(X;\theta)]$ denotes its mathematical expectation under the distribution of X.

If $w(x;\theta) = x$, the resulting weighted distribution is known as the length-biased distribution. For instance, studies involving family size as a sampling factor often produce length-biased samples. In Zelen and Feinleib (1969), this distribution is applied to the early

[1] Department of Statistics, Razi University, Kermanshah, Iran. E-mail: s.khazaei@razi.ac.ir. ORCID: https://orcid.org/0000-0003-2537-9232.

[2] Department of Statistics, Razi University, Kermanshah, Iran. E-mail: bohlurihajjar.soghra@razi.ac.ir. ORCID: https://orcid.org/0000-0001-9803-5823.

detection of breast cancer. Similarly, Patil and Rao (1977) used the length-biased distribution to study human family structures and wildlife populations. Later, Patil and Rao (1978) introduced a broader class of distributions of the form given in Equation 1, incorporating arbitrary non-negative weight functions $w(x;\theta)$, and provided several practical examples. For additional examples of weighted distributions and their applications, see Blumenthal (1967), Gupta and Kirmani (1990), Mahafoud and Patil (1982), Patil and Rao (1977), Patil and Rao (1978). This paper adopts a Bayesian nonparametric framework to model weighted data derived from such distributions. Specifically, we use the Dirichlet Process Mixture Model (DPMM), a popular Bayesian nonparametric approach, and apply it to survival analysis.

Burr (1942) introduced a family of distributions, from which twelve distinct types (named Burr distributions Type I to XII) can be derived as special cases. Among them, the Burr Type XII (Burr XII) distribution is widely used in survival studies.

Let $\kappa_B(t|c,k)$ and $\mathscr{K}_B(t|c,k)$ denote the probability density function (p.d.f.) and cumulative distribution function (c.d.f.) of the Burr XII distribution, respectively. These functions, which will be employed in our mixture model, are defined as:

$$\kappa_B(t|c,k) = ck\frac{t^{c-1}}{(1+t^c)^{k+1}}, \quad c,k>0, \quad t>0 \tag{2}$$

$$\mathscr{K}_B(t|c,k) = 1-(1+t^c)^{-k} \tag{3}$$

with the parameter space

$$\Theta = \{(c,k); 0<c<\infty, 0<k<\infty\}.$$

In Hatjispyros et al. (2017) , the Dirichlet Process Mixture Model (DPMM) is employed for density estimation under length-biased data. The authors use a log-normal distribution as the kernel, with a fixed distribution assigned to its shape parameter.

In contrast, we consider a DPMM with the Burr Type XII (Burr XII) distribution as the kernel, which includes two shape parameters treated as random variables in the model. Despite the increasing adoption of nonparametric methods in data analysis, many existing approaches struggle to handle weighted data effectively. Key limitations include inflexibility in capturing complex distributional shapes, poor modeling of heavy-tailed behavior, and a lack of adaptability to hidden heterogeneity. Traditional parametric and classical nonparametric models often fail to accurately represent the underlying structure of such data, particularly when distributions exhibit skewness or heavy tails. The Burr XII distribution is highly flexible, making it well-suited for modeling diverse distributional forms, especially those with heavy tails. However, its integration into mixture models, particularly within a Bayesian nonparametric framework, has received limited attention. This research addresses this methodological gap by introducing a Bayesian nonparametric model based on a Dirichlet Process Mixture of Burr XII distributions. The proposed framework offers enhanced flexibility and robustness for analyzing weighted data, enabling more accurate characterization of the complex, heterogeneous structures frequently encountered in real-world applications.

The Burr XII distribution has support on $\mathbb{R}^+$ and serves as a generalization of both the log-normal and Weibull distributions. These characteristics make it particularly suitable for survival analysis (Bohlouri Hajjar and Khazaei (2018), Lanjoni et al. (2016), Rao et al. (2015) and Rodriguez (1977)).

In Bohlouri Hajjar and Khazaei (2018), the Burr XII distribution is used as the kernel in a DPMM framework, where the survival function and hazard rate are computed for both simulated and real-world datasets. In Joudaki et al. (2024), a Dirichlet Process Mixture Model (DPMM) with a three-parameter Burr XII kernel was considered. Their study investigates survival analysis using this flexible modeling approach, demonstrating its effectiveness in capturing complex features of survival data. Bayesian estimation methods for hybrid censored data from the Burr XII distribution using various loss functions were discussed in Hassan (2021). These methods were particularly valuable for managing complex censoring scenarios. In Nurul et al. (2024), a DPMM-based approach was proposed for clustering mixed-type data with cluster-specific covariance matrices, effectively addressing intricate data structures. Moreover, Michael et al. (2023) applied DPMMs to longitudinal data involving repeated attempts, successfully modeling the complexities inherent in such datasets.

In the next section, we present the preliminary concepts and methodology, followed by a detailed introduction of the proposed model. Section 4 describes the use of Gibbs sampling to estimate the original (unweighted) distributions from their corresponding weighted forms. Section 5 demonstrates the application of our approach to both simulated and real datasets. Finally, Section 6 summarizes our findings and conclusions.

2. Preliminary and Methodology

We aim to estimate the density and survival functions by considering a general case of the weight function $w(x;\theta)$. To avoid computing the often intractable normalizing constant, our strategy is to model $g(x;\theta)$ directly and then infer $f(x;\theta)$, using the fact that $g(x;\theta) \propto w(x;\theta)f(x;\theta)$.

If we assume that $f(x;\theta)$ belongs to a parametric family, then both $f(x;\theta)$ and $g(x;\theta)$ are known up to the normalizing constant, which may not be analytically tractable.

Let $w(\cdot;\theta)$ be a general weight function; an essential condition for modelling $F(\cdot;\theta)$ through $G(\cdot;\theta)$ $\big(F(\cdot;\theta)$ and $G(\cdot;\theta)$ denote the distribution functions corresponding to $f(\cdot;\theta)$ and $g(\cdot;\theta)$, respectively$\big)$ is

$$\int_0^\infty w(x)^{-1}g(x)dx < \infty, \tag{4}$$

because f is a distribution function.

Through the invertibility implied by Equation (4), it becomes possible to reconstruct the distribution function F from G. In the Bayesian nonparametric framework, we place a suitable nonparametric prior on g, relying on the relationship defined by Equation (4).

The key question, however, is how the posterior structure derived from modeling g directly can be transformed into the corresponding posterior structure for f.

The first step involves developing a method to convert a weighted sampler into an unweighted sample. Once this conversion is achieved, inference about the posterior distributions can be made.

The Markov Chain Monte Carlo (MCMC) approach is an indirect method for simulating samples from complex probability distributions. One of the key MCMC methods is the Metropolis-Hastings algorithm Hatjispyros et al. (2017), which generates samples from a target distribution by utilizing its full joint density function along with proposal distributions for each of the variables of interest.

Algorithm 1: Metropolis-Hastings algorithm

```
 1. Initialize with $x^{(0)} \sim q(x)$
    2. for $i = 1,2,\dots$ do
       Propose $x^{cand} \sim q(x^{(i)}|x^{(i-1)})$
       Calculate the acceptance probability:
       $\alpha(x^{cand}|x^{(i-1)}) = \min\left\{1, \frac{q(x^{(i-1)}|x^{cand})\pi(x^{cand})}{q(x^{cand}|x^{(i-1)})\pi(x^{(i-1)})}\right\}$
       Generate $u \sim \mathrm{Uniform}(0,1)$
        if $u < \alpha$ then
          $x^{(i)} \leftarrow x^{cand}$
        else
          $x^{(i)} \leftarrow x^{(i-1)}$
        end if
   end for.
```

Here, $q(x)$ represents the weighted distribution. Hatjispyros et al. (2017) demonstrated how the Metropolis-Hastings algorithm can be used to convert a length-biased sample into an unbiased one. Following a similar strategy, we aim to apply this algorithm using a general weight function, as defined in Equation (4), to transform samples from a weighted distribution into their unweighted counterparts. This methodology is particularly important when dealing with complex models where the weighted distribution arises due to inherent sampling bias.

Suppose that $y_1, y_2, \dots, y_N$ denote a random sample from g. The Metropolis-Hastings algorithm is used to convert this sample into a sample from $f(x;\theta) \propto w(x;\theta)^{-1} g(x;\theta)$. In the algorithm, we assume that $g(\cdot)$ is replaced by $q(\cdot)$ in Algorithm 1 with the acceptance probability

$$\min\left\{1, \frac{w^{-1}(y_{j+1})}{w^{-1}(x_j)}\right\}.$$

If x_j denotes the current sample from $f(x)$, then

$$\begin{aligned} x_{j+1} &= y_{j+1} \quad \text{with probability} \quad \min\left\{1, \frac{w^{-1}(y_{j+1})}{w^{-1}(x_j)}\right\}, \\ x_{j+1} &= x_j \quad \text{otherwise.} \end{aligned} \tag{5}$$

The transition density is

$$P(x_{j+1}|x_j) = \min\left\{1, \frac{w^{-1}(y_{j+1})}{w^{-1}(x_j)}\right\} g(x_{j+1}) + \{1 - r(x_j)\} 1(x_{j+1} = x_j),$$

where

$$r(x) = \int \min\left\{1, \frac{w^{-1}(x^*)}{w^{-1}(x)}\right\} g(x^*) dx^*.$$

We can outline the general methodology as follows:

1. *Sample Generation*: Consider $(y_1, \ldots, y_n)$ as a sample from g, to which we assign a suitable nonparametric prior.

2. *Posterior Inference*: Using MCMC methods, posterior samples from the random measure $\Pi(dg|y_1, \ldots, y_n)$ and other relevant parameters are obtained. Consequently, a sequence $\{y^l_{n+1}\}, l = 1, 2, \ldots$ from the posterior predictive density $g(y|y_1, \ldots, y_n)$ will be generated.

3. *Weighted Proposal Values*: The sequence $\{y^l_{n+1}\}$ serves as proposal values in a Metropolis-Hastings chain whose stationary distribution is the weighted posterior predictive, i.e.,
$$\{y^l_{n+1}\} \propto w(y)^{-1} g(y|y_1, \ldots, y_n).$$
Using Equation (5), we generate the corresponding values $\{x^l_{n+1}\}$ at iteration l.

4. *Final Sample*: The resulting sequence $\{x^l_{n+1}\}$ constitutes a sample from the posterior predictive distribution f, which corresponds to the unweighted density.

3. The model and inference

In this section, we aim to model $g(x;\theta)$. Modeling the weighted distribution $g(x)$ within the Bayesian nonparametric framework is based on an infinite mixture model Lo (1984), which takes the following form:

$$g_P(y) = \int \kappa(y;\theta) P(d\theta), \tag{6}$$

where P is a discrete probability measure and $\kappa(y;\theta)$ is a kernel density defined on $(0,\infty)$ for all θ in the parameter space. This kernel satisfies the condition

$$\int_0^\infty w^{-1}(y;\theta)\kappa(y;\theta) dy < \infty.$$

By choosing the Burr(XII) density (with parameters c and k) as the kernel of the mixture model, we obtain

$$g_{c,k,P}(y) = \int_{\mathbb{R}} \kappa_B(y|c,k)P(dc,dk),$$

where $\kappa_B(y|c,k)$ is the Burr(XII) density and P is a discrete random probability measure. Suppose that

$$P \sim DP(\upsilon, P_0),$$

where $DP(\upsilon, P_0)$ denotes the Dirichlet process with precision parameter $\upsilon > 0$ and base measure P_0 Ferguson (1983). We refer to this mixture model as the Dirichlet Process Burr(XII) Mixture Model (DPBMM).

The hierarchical representation of the DPBMM can be expressed as follows:

$$\begin{aligned} y|c,k &\sim \kappa_B(y|c,k), \\ (c,k)|P &\sim P, \\ P|\upsilon,P_0 &\sim DP(\upsilon,P_0). \end{aligned} \tag{7}$$

Suppose that the base distribution P_0 is the prior distribution for the joint distribution of c and k. By choosing Burr(XII) distribution as the kernel, P_0 that yields a closed-form expression for $\int \kappa_B(.|c,k)P_0(dc,dk)$ is not available. Moreover, we choose multiple distributions of Uniform$(0,\phi)$ and Exponential with the parameter γ for P_0, i.e.,

$$P_0(c,k|\phi,\gamma) = \text{Unif}(c|0,\phi) \times \text{Exp}(k|\gamma). \tag{8}$$

This choice achieves the modeling goals. Considering the hyperparameters γ and ϕ as random, we choose prior distributions $Pareto(a_\phi,b_\phi)$ and $IGamma(a_\gamma,b_\gamma)$ for them, respectively. We set $a_\phi = a_\gamma = d$ and chose $d = 2$, which makes the variance of the Pareto distribution infinite. This allows the distribution to accommodate a wide range of values. The parameters b_ϕ and b_γ are determined by the data Bohlouri Hajjar and Khazaei (2018).

Finally, for any t_i, $i = 1,...,n$, representing lifetime data in a sample of n observations, by considering DPBMM and selecting priors for parameters of the model we have

$$\begin{aligned} t_i|c_i,k_i &\sim \kappa_B(t_i|c_i,k_i), \quad i=1,...,n, \\ (c_i,k_i)|P &\sim P, \\ P &\sim DP(\nu,P_0), \\ P_0|\gamma,\phi &\sim \text{Unif}(c|0,\phi)\times\text{Exp}(k|\gamma), \\ \nu,\gamma,\phi &\sim \text{Gamma}(a_\nu,b_\nu)\times IGamma(a_\gamma,b_\gamma)\times Pareto(a_\phi,b_\phi). \end{aligned} \tag{9}$$

After determining the model, we want to formulate how to sample from DPMMs by Gibbs sampling. According to Kottas (2006), Gibbs sampling for drawing a sample from

$$[(\theta_1,\ldots,\theta_n),\, \upsilon,\, \ldots \mid t]$$

is based on the following full conditional distributions (brackets are used to indicate conditional and marginal distributions):

$$\begin{array}{ll} (1) & [(\theta_i)|(\theta_{-i},z_{-i}),\upsilon,...,t], \quad \text{for } i=1,...,n \\ (2) & [(\theta_j^*)|z,n^*,\upsilon,...,t], \quad \text{for } j=1,...,n^* \\ (3) & [\upsilon|\{(\theta_j^*),j=1,...,n^*\},n^*,t],[...|\{(\theta_j^*),j=1,...,n^*\},n^*,t]. \end{array} \tag{10}$$

Here, t is the vector of failure time data. The θ_i's are the parameters of the kernel in DPMMs that will be analyzed.

Model (8) and the discreteness property of the Dirichlet process exhibit clustering in the θ's. We present n^* as the number of clusters among the θ_i's, denoted by θ_j^*'s. The vector of indicators $z=(z_1,...,z_n)$ indicates the clustering configuration such that $z_i=j$ when $\theta_i=\theta_j^*$.

Also, θ_{-i}, which is used in (8), is defined as $\theta_{-i}=(\theta_1,\theta_2,...,\theta_{i-1},\theta_{i+1},...,\theta_n)$. Model (12) is the same as model (8), with the difference that the vector of indicators $z=(z_1,...,z_n)$ indicates the clustering configuration such that $z_i=j$ when $\theta_i=\theta_j^*$. Also, θ_{-i}, which is used in (8), is defined as $\theta_{-i}=(\theta_1,\theta_2,...,\theta_{i-1},\theta_{i+1},...,\theta_n)$.

4. Modeling

We apply the following algorithm to model the unweighted density $f(x)$ from the weighted density $g(x)$. First, to generate a sample from $g(x)$, the model's parameters need to be estimated. To this aim, we draw a sample from (c_i,k_i) and update z_i for each t_i.

In simulation-based parameter estimation, we use the Gibbs sampler, which includes two steps to reach the goal.

Algorithm 2 : Gibbs sampler

```
1. Initialize with $\theta^{(0)} \sim f(\theta)$
2. For $i = 1,2,...$ do
    $\theta_1^{(i)} \sim f(\theta_1|\theta_2^{(i-1)},\theta_3^{(i-1)},...,\theta_d^{(i-1)},D)$,
    $\vdots$
    $\theta_d^{(i)} \sim f(\theta_d|\theta_1^{(i)},\theta_2^{(i)},...,\theta_{d-1}^{(i)},D)$,
```

where $\theta_1,...,\theta_d$ are model parameters, and D is the vector of observations. The values at iteration i are sampled from the conditional distributions using the most recent values of the other parameters.

Now, the model will be applied to lifetime data with right-censored observations, which are very common in survival studies. To calculate the related distributions, the data are divided into uncensored and censored observations.

1- Uncensored data: For uncensored data (t_{io}), the conditional posterior density of (c_i, k_i) is a mixture distribution Neal (2003):

$$f(c_i, k_i \mid \{(c_{i'}, k_{i'}); i \neq i'\}, \nu, \gamma, \phi, t_{io}) = \frac{q_0^o h^o(c_i, k_i \mid \phi, \gamma, t_{io}) + \sum_{j=1}^{n^{*(i)}} n_j^{*(i)} q_j^o \delta_{c_j^*, k_j^*}}{q_0^o + \sum_{j=1}^{n^{*(i)}} n_j^{*(i)} q_j^o},$$

where $q_j^o = k_B(t_{io} \mid c_j^*, k_j^*)$, and

$$\begin{aligned} q_0^o &= \nu \int_0^\phi \int_0^\infty k_B(t_{io} \mid c,k) G_0(c,k)\, dc\, dk \\ &= \frac{\nu}{\phi} \int_0^\phi \frac{c t_{io}^{c-1}}{(1+t_{io}^c)} \left(\int_0^\infty \frac{k e^{-k/\gamma}}{(1+t_{io}^c)^k} dk \right) dc \\ &= \frac{\nu}{\phi} \int_0^\phi \frac{c t_{io}^{c-1}}{(1+t_{io}^c)\left(\ln(1+t_{io}^c) + \frac{1}{\gamma}\right)} dc, \end{aligned}$$

where the last integration can be computed numerically, and

$$h^o(c_i, k_i \mid \gamma, \phi, t_{io}) \propto k_B(t_{io} \mid c_i, k_i) P_0(c_i, k_i \mid \gamma, \phi) \propto [c_i \mid \gamma, \phi, t_{io}][k_i \mid c_i, \gamma, \phi, t_{io}],$$

with

$$[c_i \mid \gamma, \phi, t_{io}] \propto c_i t_{io}^{c_i - 1} I_{(0,\phi)}(c_i), \quad i = 1, \dots, n,$$

and

$$[k_i \mid c_i, \gamma, \phi, t_{io}] \propto \text{Gamma}\left(\cdot \mid 2, \frac{1}{\frac{1}{\gamma} + \ln(1 + t_{io}^{c_i})} \right).$$

2- Right censored data: For right-censored data (t_{ic}), the conditional posterior density of (c_i, k_i) is

$$f(c_i, k_i \mid \{(c_i, k_i); i \neq i'\}, \nu, \gamma, \phi, t_{ic}) = \frac{q_0^c h^c(c_i, k_i \mid \phi, \gamma, t_{ic}) + \sum_{j=1}^{n^{*(i)}} n_j^{*(i)} q_j^c \delta_{c_j^*, k_j^*}}{q_0^c + \sum_{j=1}^{n^{*(i)}} n_j^{*(i)} q_j^c},$$

where $q_j^c = 1 - K_B(t_{ic} \mid c_j^*, k_j^*)$, and

$$\begin{aligned} q_0^c &= \nu \int_0^\phi \int_0^\infty \big(1 - K_B(t_{ic} \mid c,k)\big) G_0(c,k)\, dc\, dk \\ &= \frac{\nu}{\phi\gamma} \int_0^\phi \int_0^\infty \frac{e^{-k/\gamma}}{(1+t_{ic}^c)^k} dk\, dc \\ &= \frac{\nu}{\phi\gamma} \int_0^\phi \left(\frac{1}{\gamma} + \ln(1+t_{ic}^c) \right) dc, \end{aligned}$$

where the last integration can be computed numerically. Using the property of censored data, we have

$$\begin{aligned} h^c(c_i,k_i \mid \gamma,\phi,t_{ic}) &\propto \big(1-K_B(t_{ic}\mid c_i,k_i)\big)G_0(c_i,k_i) \\ &\propto [c_i \mid \gamma,\phi,t_{ic}][k_i \mid c_i,\gamma,\phi,t_{ic}] \\ &= \frac{I_{(0,\phi)}(c_i)}{\phi\gamma}\frac{1}{\frac{1}{\gamma}+\ln(1+t_{ic}^{c_i})}k_i\exp\left\{-k_i\Big(\frac{1}{\frac{1}{\gamma}+\ln(1+t_{ic}^{c_i})}\Big)\right\} \\ &= \frac{I_{(0,\phi)}(c_i)}{\phi\gamma}\frac{1}{\frac{1}{\gamma}+\ln(1+t_{ic}^{c_i})}\times \text{Gamma}\Big(k_i \mid 2,\frac{1}{\frac{1}{\gamma}+\ln(1+t_{ic}^{c_i})}\Big). \end{aligned}$$

We use the slice sampling method to sample from the first part of the above expression. Therefore, using this MCMC approach, a sample from $h^c(c_i,k_i \mid \phi,\gamma,t_{ic})$ can be obtained. Now, for both observed and censored data, (c_i,k_i) for $i=1,\dots,n$ can be updated iteratively and improved. In a general form, (c_j^*,k_j^*) can be updated conditional on ϕ,γ, and t as follows:

$$\begin{aligned} f(c_j^*,k_j^* \mid \phi,\gamma,t,n^*) &\propto G_0(c_j^*,k_j^* \mid \gamma,\phi)\prod_{\{io:s_{io}=j\}}k_B(t_{io}\mid c_j^*,k_j^*)\prod_{\{ic:s_{ic}=j\}}\big(1-K_B(t_{ic}\mid c_j^*,k_j^*)\big) \\ &\propto [c_j^* \mid \gamma,\phi,t_{io}][k_j^* \mid c_j^*,\gamma,\phi,t_{ic}]\prod_{\{ic:s_{ic}=j\}}\frac{1}{(1+t_{ic}^{c_j^*})^{k_j^*}} \\ &\propto c_j^{*n_j^o}I_{(0,\phi)}(c_j^*)\prod_{\{io:s_{io}=j\}}\frac{t_{io}^{c_j^*-1}}{1+t_{io}^{c_j^*}}\times\text{Gamma}(n_j^o+1,B^*), \qquad (11) \end{aligned}$$

where

$$B^*=\sum_{\{io:s_{io}=j\}}\left(\frac{1}{\gamma}+\ln(1+t_{io}^{c_j^*})\right)+\sum_{\{ic:s_{ic}=j\}}\ln(1+t_{ic}^{c_j^*}),$$

and n_j^o is the number of observed data points in cluster j.

The key task in generating a sample from Equation (13) is drawing from the first part of the equation. Sampling from the gamma distribution is straightforward. To sample from

$$[c_j^* \mid \phi,\gamma,t]\propto c_j^{*n_j^o}I_{(0,\phi)}(c_j^*)\prod_{\{io:s_{io}=j\}}\frac{t_{io}^{c_j^*-1}}{1+t_{io}^{c_j^*}},$$

auxiliary variables $W=\{w_{io}:\{io:s_{io}=j\}\}$ are introduced such that

$$[c_j^*,W \mid \phi,t_{io}]=c_j^{*n_j^o}I_{(0,\phi)}(c_j^*)\prod_{\{io:s_{io}=j\}}I_{\big(0,\frac{t_{io}^{c_j^*-1}}{1+t_{io}^{c_j^*}}\big)}(w_{io}).$$

By marginalization over W, $[c_j^* \mid \phi,t_{io}]$ is obtained for $j=1,\dots,n^*$. Moreover, w_{io} are uniform variables on $(0, \frac{t_{io}^{c_j^*-1}}{1+t_{io}^{c_j^*}})$. Therefore,

$$[c_j^* \mid \phi, t] = c_j^{*n_j^o} I_{(B,\phi)}(c_j^*),$$

where $B = \max\{0, \frac{\ln(w_{io})}{1+t_{io}}\}$. Drawing from $[c_j^* \mid \phi, t]$ is now straightforward.

Subsequently, following the approach in Escobar and West (1995), ϕ, γ, and ν are updated. Introducing a latent variable u such that

$$[u \mid \nu, t] = \text{Beta}(\nu + 1, n),$$

we have

$$[\nu \mid u, n^*, t] = p\,\text{Gamma}(a_\nu + n^*, b_\nu - \ln(u)) + (1-p)\,\text{Gamma}(a_\nu + n^* - 1, b_\nu - \ln(u)),$$

where

$$p = \frac{a_\nu + n^* - 1}{n(b_\nu - \ln(u)) + a_\nu + n^* - 1}.$$

To update ϕ, we have

$$[\phi \mid c^*, k^*] = [\phi][c^*, k^* \mid \phi] = \frac{2b_\phi^2}{\phi^3} I_{(b_\phi,\infty)}(\phi) \prod_{j=1}^{n^*} \frac{1}{\phi} I_{(0,\phi)}(c_j^*) = \frac{2b_\phi^2}{\phi^{n^*+3}} I_{(b^*,\infty)}(\phi),$$

where $b^* = \max\{b_\phi, \max_{1 \le j \le n^*} c_j^*\}$, implying

$$[\phi \mid c^*, k^*] = \text{Pareto}(\phi \mid 2 + n^*, b^*).$$

Repeating this technique, γ is updated as

$$[\gamma \mid c^*, k^*] = [\gamma] \prod_{j=1}^{n^*} [k_j^* \mid \gamma] = \text{IGamma}(n^* + 2, b_\gamma + \sum_{j=1}^{n^*} k_j^*).$$

Thus, all conditional distributions required for Equation (8) can now be computed.

5. Data illustrations

We evaluate the performance of the proposed model by applying it to three simulated datasets, each with a sample size of $n = 200$. For model comparison, we consider alternative approaches from the literature. Gibbs sampling is implemented with a total of $N = 15,000$ iterations, discarding the initial 1,000 as burn-in, and applying thinning by retaining every 20th iteration. Assuming a sufficiently large sample size, we place a $Gamma(1, 0.001)$ prior on the concentration parameter α, allowing the model to infer an appropriate number of clusters based on the data.

To quantify the deviation between estimated and true models, we compute numerical indices using both Euclidean and Hellinger distance metrics that have been widely used in similar studies:

$$d_E(f,\hat{f}) = \sqrt{\sum_{i=1}^{n} \left(f(t_i) - \hat{f}(t_i)\right)^2} \tag{12}$$

$$d_H(f,\hat{f}) = \sqrt{\sum_{i=1}^{n} \left(\sqrt{f(t_i)} - \sqrt{\hat{f}(t_i)}\right)^2}. \tag{13}$$

Here, f denotes the true function, which may be a density, survival, or hazard function, while $\hat{f}$ represents its estimate obtained from the MCMC output.

We generate samples of size $n = 200$ from the following distributions:

(i) Mixture of Burr distributions with 2 parameters (B2M):

$$0.4 \times \text{Burr}(c = 25, k = 10) + 0.6 \times \text{Burr}(c = 7, k = 4).$$

(ii) Mixture of lognormal distributions (LNM):

$$0.2 \times \text{LN}(\mu_1 = 0, \sigma_1^2 = 0.15) + 0.3 \times \text{LN}(\mu_2 = 1, \sigma_2^2 = 0.02) + 0.5$$
$$\times \text{LN}(\mu_3 = 2, \sigma_3^2 = 0.04).$$

(iii) Mixture of Weibull distributions (WM):

$$0.4 \times \text{Weibull}(\alpha = 1, \lambda = 0.25) + 0.6 \times \text{Weibull}(\alpha = 6, \lambda = 0.5).$$

Here, $\text{Burr}(c,k)$ denotes the Burr Type XII distribution with shape parameter c and scale parameter k; $\text{LN}(\mu,\sigma^2)$ denotes the lognormal distribution with scale parameter μ and shape parameter σ; and $\text{Weibull}(\alpha,\lambda)$ refers to the Weibull distribution with shape parameter α and scale parameter λ.

Table 1 presents the computed index values for the density, survival, and hazard functions under the Dirichlet Process Mixture Model (DPMM) with different kernel choices, evaluated across the three simulated datasets. Details of the DPWM, DPLNM, and DPB2M models are provided in references Hassan et al. (2021) and Cheng and Yuan (2013). The results in Table 1 indicate that the DPB2M model achieves a better fit compared to the other models based on the simulated datasets.

Table 1. Deviance metrics d_E ($d_H(\cdot,\hat{\cdot})$) for DPMMs with different kernels. Values in parentheses are Hellinger distances.

Dataset	Metric	DPWM	DPLNM	DPB2M
I	$f(t)$	1.553 (1.126)	1.019 (0.522)	6.307 (4.015)
	$S(t)$	0.602 (0.467)	0.295 (0.213)	3.276 (1.658)
	$h(t)$	23.097 (4.031)	10.160 (1.473)	33.559 (8.834)
II	$f(t)$	1.448 (1.224)	0.373 (0.309)	1.901 (1.709)
	$S(t)$	0.650 (0.471)	0.195 (0.144)	1.692 (1.088)
	$h(t)$	5.542 (2.489)	1.788 (0.668)	7.407 (3.832)
III	$f(t)$	2.218 (0.669)	4.762 (1.184)	2.607 (0.767)
	$S(t)$	0.399 (0.213)	0.321 (0.198)	0.368 (0.206)
	$h(t)$	8.281 (1.053)	13.017 (1.695)	13.809 (1.558)

Note that the values in parentheses represent Hellinger distances.

5.1. Simulated data

In this subsection, we illustrate the capability of the DPB2M model to effectively model weighted data and recover the corresponding unweighted distribution. For this purpose, we consider two types of datasets: simulated data and real data.

Assume we are given a sample $(x_1,\ldots,x_n)$, and our objective is to estimate its density function. To evaluate the goodness of fit of the proposed model, we compare the resulting density estimate with the following two density estimators, which are used in Hatjispyros et al. (2017):

i) **The classical kernel density estimate:**

$$\tilde{g}_h(x;(x_1,\ldots,x_n)) \propto \frac{1}{n}\sum_{j=1}^{n} N(x \mid x_j,h^2)\,1_{(0,+\infty)}(x)$$

ii) **The kernel density estimate for indirect data (Jones' kernel density estimate):**

$$\hat{f}_{J,h}(x;(x_1,\ldots,x_n)) \propto \frac{1}{n}\hat{\mu}\sum_{j=1}^{n} x_j^{-1} N(x \mid x_j,h^2)\,1_{(0,+\infty)}(x)$$

where $\hat{\mu}$ is the harmonic mean of the sample $(x_1,\ldots,x_n)$.

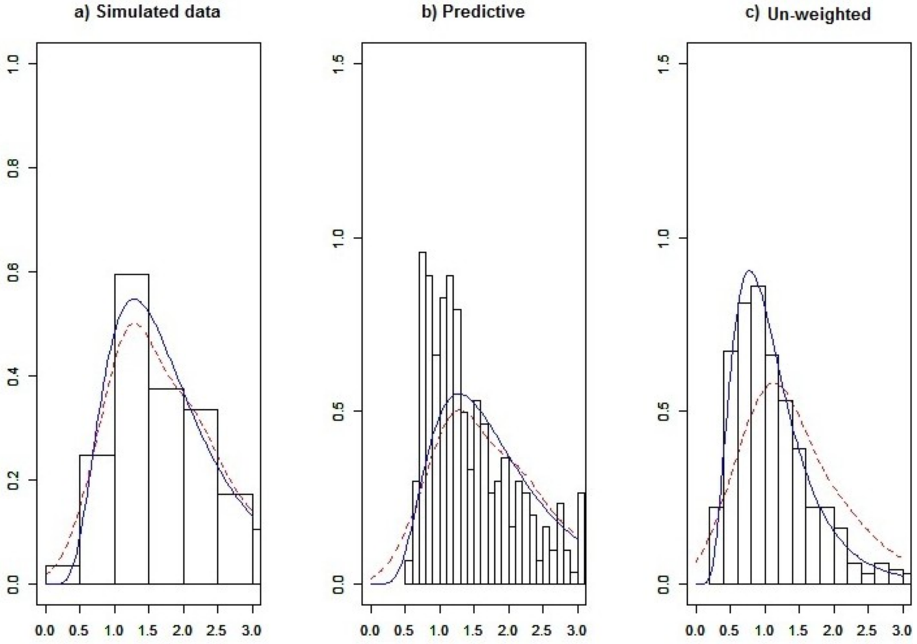


Figure 1. Simulated data from the log-normal distribution with parameters $(0.5, 0.5)$ and a sample size of $n = 100$. In each figure, the true densities are shown with a solid line, and the kernel density estimates $\tilde{g}_h$ (a),(b) and $\hat{f}_{J,h}$ (c) with a dashed line.

These estimators are among the best-known nonparametric methods and provide a good fit for both the weighted and unweighted data in our analysis. For the simulated datasets, the Metropolis-Hastings algorithm was run for over 50,000 iterations, while the Gibbs sampler was executed for 60,000 iterations, with a burn-in period of 10,000 iterations.

5.1.1 Length biased distribution of log-normal

Here, the first dataset is simulated from the log-normal distribution with parameters $(\mu, \sigma^2) = (0.5,\ 0.5)$. We know that the length-biased distribution of a log-normal with parameters $\mu + \sigma^2$ and σ^2 is again a log-normal with parameters μ and σ^2 Patil and Rao (1978).

By choosing the log-normal distribution as the kernel, we can illustrate the model's preference and the algorithm. This model was tested in Hatjispyros et al. (2017) for length-biased data using simulated data from the Gamma distribution with DPMM when the Gamma distribution was considered the kernel.

The results of the simulated data are shown in Figure 1. In panel (a), the histogram of simulated log-normal $(0.5, 0.5)$ data is presented, and the true density curve is depicted with a solid line, while the kernel density estimate $\tilde{g}_h$ is shown with a dashed line. The estimate $\tilde{g}_h$ closely approximates the true underlying density.

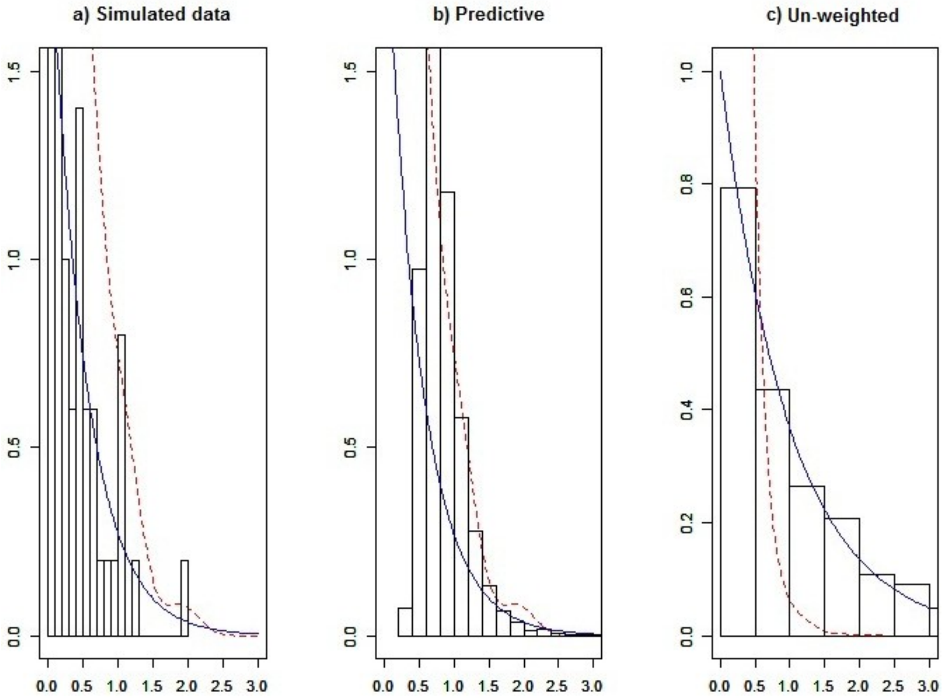


Figure 2. Simulated data from the Weibull distribution with parameters $(\alpha = 1, \lambda = 2)$ and sample size $n = 100$. True densities are shown with a solid line, and the kernel density estimates $\tilde{g}_h$ and $\hat{f}_{J,h}$ with a dashed line.

In panel (b), the histogram of the posterior predictive distribution of the data is shown along with the true density curve (solid line). The real data here follows a log-normal distribution with parameters $(0, 0.5)$, and Jones' density estimate $\hat{f}_{J,h}$ is represented by a dashed line.

Panel (c) depicts the histogram of the transformed data to the unweighted scale, corresponding to the indirect data estimate, which is shown with a dashed line. The distribution of the unweighted data is also close to the true distribution, a log-normal with parameters $(0, 0.5)$, demonstrating the model's ability to recover the underlying density effectively.

5.1.2 Weighted distribution of Gamma

Here, we consider a $Gamma(\alpha, \beta)$ distribution with the weight function $w(x|a,b) = x^a \exp(-x/b)$. By applying the weighting function, we obtain a new distribution that is again a Gamma distribution, but with updated parameters $(\alpha + a, (\beta + b)/(b\beta))$.

The dataset is simulated from a weighted Gamma distribution, specifically $Gamma(1, 2)$ using the weight function $w(x) = \exp(-x)$, which corresponds to $a = 0$ and $b = 1$. The corresponding unweighted version of this distribution is a Gamma distribution with parameters $\alpha = 1$ and $\beta = 1$.

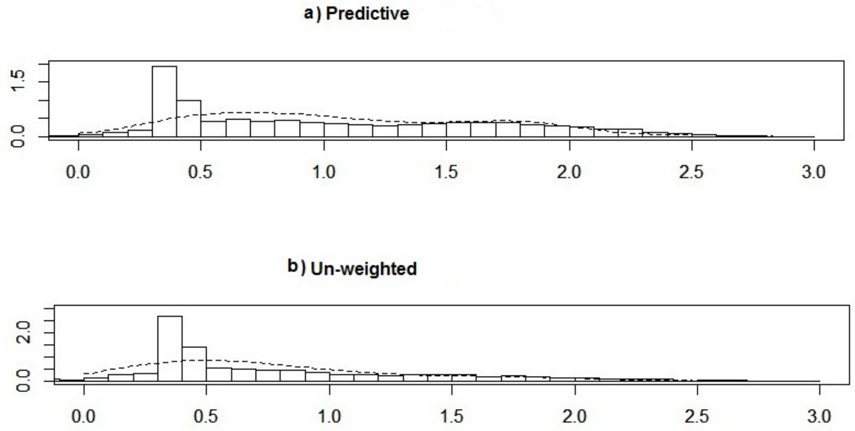


Figure 3. Real dataset of the widths of shrubs with size $n = 46$. (a) Histogram of data and estimated posterior predictive density by DPBMM, and (b) histogram of de-biased data using the Metropolis-Hastings algorithm and $\hat{f}_{J,h}$.

In Figure 2, panel (a), we present the histogram of the simulated data along with its true density curve and the kernel density estimate $\tilde{g}_h$. Panel (b) displays the histogram of the predictive values, the corresponding kernel estimate $\tilde{g}_h$, and the true density curve. In panel (c), we present the histogram of the unweighted distribution obtained using the Metropolis-Hastings algorithm, along with the estimate $\hat{f}_{J,h}$ for this data.

5.2. Real data

In the previous section, Dirichlet Process Bayesian Mixture (DPB2M) models demonstrated a good fit to the simulated data. Therefore, we now apply the flexibility of this Bayesian nonparametric model to real datasets, namely the shrub width data and the bladder cancer data.

5.2.1 Widths of shrubs data

We consider the data that can be found in Muttlak and McDonald (1990) for applications with real data. This data consists of 46 measurements of the width of shrubs that are sampled by line-transect. In this sampling method, the probability of inclusion in the sample is proportional to the width of the shrub, making it a case of length-biased sampling.

We can see the predictive values of the DPB2M model with the histogram and $\tilde{g}_h$ with the dashed line in panel (a) of Figure 3, and also the unweighted version of the data values and $\hat{f}_{J,h}$ depicted in panel (b).

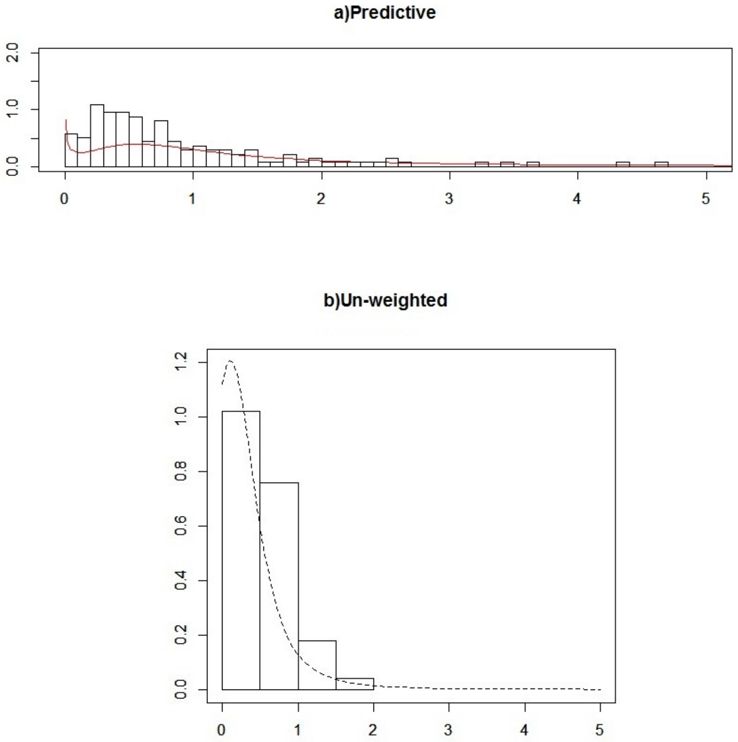


Figure 4. Real dataset of bladder cancer patients with size $n = 137$. (a) Histogram of data and estimated posterior predictive density by DPBMM, and (b) histogram of de-biased data using the Metropolis-Hastings algorithm and $\hat{f}_{J,h}$.

5.2.2 Bladder cancer data

The next real dataset is survival data which includes censored values. This data is taken from Lee and Wang (2003), page 231, which corresponds to remission times (in months) of a random sample of bladder cancer patients. Properties of this data are: the total number of observations 137, censored data 9, largest observation 46.12, and smallest observation 0.08. For easier model fitting, we divided the data into 10 intervals.

In Ahmad et al. (2016), a parametric model was fitted to this dataset without considering censored observations. This model is referred to as the length-biased weighted Lomax distribution. The Lomax distribution is a special case of the Burr(XII) distribution. In this section, we apply the DPBM model, which is a Bayesian nonparametric model with Burr(XII) distribution as the kernel of a mixture model.

In Figure 4, panel (a), we show the histogram of the bladder cancer data, along with the estimated density function based on the DPBM model. Since this dataset comes from a weighted distribution, a histogram of the unweighted values obtained using the Metropolis-Hastings method and the corresponding curve is presented in panel (b) of Figure 4.

6. Conclusion

This article uses the Bayesian nonparametric approach to model weighted data. We use the Dirichlet process mixture model (DPMM) with Burr(XII) distribution as the kernel function in mixing models. We assumed weighted distribution with an arbitrary weight function that satisfies equation (2). Using the Metropolis-Hastings algorithm, the weighted distribution converted to the unweighted one. We fit the DPMM with the different kernels and weight functions for real and simulated data sets as an application. As an application in the survival study, a real lifetime dataset containing censored observations is used, and density and survival functions are estimated.

7. Discussion

While the proposed Bayesian nonparametric model, the Dirichlet Process Mixture of Burr XII distributions, provides substantial flexibility and robustness for modeling weighted data, several limitations should be noted. First, the model's computational complexity can be high, particularly with large datasets or high-dimensional covariates, potentially hindering its scalability in practice. Second, the selection of hyperparameters and prior distributions may significantly affect performance and inference, necessitating careful tuning and sensitivity analysis. Third, although the Burr XII distribution is flexible, there may be cases where other kernel distributions might better capture certain data characteristics.

Future research could address these limitations by extending the model to incorporate covariate information more explicitly, such as through hierarchical or dependent Dirichlet processes, enhancing its utility in complex data settings. Additionally, developing more efficient computational algorithms, such as variational inference or scalable MCMC methods, could improve the model's feasibility for big data applications. Exploring integration with other flexible distributions or developing multivariate extensions may further broaden its scope and practical impact.

Acknowledgements

The author thanks Razi University for facilities provided during this research.

References

Ahmad, A., Ahmad, S. P. and Ahmed, A., (2016). Length-biased weighted Lomax distribution: statistical properties and application. *Pakistan Journal of Statistics and Operation Research*, 12(2), pp. 245–255.

Bohlourihajjar, S., Khazaei, S., (2018). Bayesian nonparametric survival analysis using mixture of Burr XII distributions. *Communications in Statistics-Simulation and Computation*, 47(9), pp. 2724–2738.

Blumenthal, S., (1967). Proportional sampling in life length studies. *Technometrics*, 9(2), 205–218.

Burr, I. W., (1942). Cumulative frequency functions. *The Annals of Mathematical Statistics*, 13(2), 215–232.

Cheng, N., Yuan, T., (2013). Nonparametric Bayesian lifetime data analysis using Dirichlet process lognormal mixture model. *Naval Research Logistics (NRL)*, 60(3), pp. 208–221.

Damien, P., Walker, S., (2002). A Bayesian Non-parametric Comparison of Two Treatments. *Scandinavian Journal of Statistics*, 29(1), pp. 51–56.

Escobar, M. D., West, M., (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430), pp. 577–588.

Ferguson, T. S., (1983). Bayesian density estimation by mixtures of normal distributions. In Recent advances in statistics, pp. 287–302, *Academic Press*.

Fisher, R. A., (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6(1), pp. 13–25.

Ghosh, J. K., Ramamoorthi, R. V., (2003). *Bayesian nonparametrics*. Springer Series in Statistics. Springer-Verlag, New York.

Gupta, R. C., Kirmani, S. N. U. A., (1990). The role of weighted distributions in stochastic modeling. *Communications in Statistics-Theory and Methods*, 19(9), pp. 3147–3162.

Hassan, et al., (2021). E-Bayesian estimation of Burr Type XII model based on adaptive Type-II progressive hybrid censored data. *International Journal of Computing Science and Mathematics*, 14(3), pp. 233–248.

Hatjispyros, S. J., Nicoleris, T., Walker, S. G., (2017). Bayesian nonparametric density estimation under length bias. *Communications in Statistics-Simulation and Computation*, 46(10), pp. 8064–8076.

Kilany, N. M., (2016). Weighted Lomax distribution. *SpringerPlus*, 5(1), p. 1862.

Hajji Joudaki, et al., (2024). Survival Analysis Using Dirichlet Process Mixture Model with a Three-Parameter Burr XII Kernel. *Communications in Statistics - Simulation and Computation*, 53(5), pp. 2406–2424.

Kottas, A., (2006). Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, 136(3), 578–596.

Lanjoni, B. R., Ortega, E. M. and Cordeiro, G. M., (2016). Extended Burr XII regression models: theory and applications. *Journal of Agricultural, Biological, and Environmental Statistics*, 21(1), pp. 203–224.

Lee, E. T., Wang, J., (2003). *Statistical methods for survival data analysis* (Vol. 476). John Wiley & Sons.

Lo, A. Y., (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, pp. 351–357.

Mahafoud, M., Patil, G. P., (1982). On weighted distributions. In *Statistics and Probability: Essays in honor of C. R. Rao*, pp. 383–405.

McLachlan, G., Peel, D., (2004). *Finite mixture models.* John Wiley & Sons.

Michael, J., et al., (2023). Dirichlet Process Mixture Models for the Analysis of Repeated Attempt Designs. *Biometrics*, 79(4), 3907–3915.

Müller, P., Quintana, F. A., (2004). Nonparametric Bayesian data analysis. *Statistical Science*, pp. 95–110.

Muttlak, H. A., McDonald, L. L., (1990). Ranked set sampling with size-biased probability of selection. *Biometrics*, pp. 435–445.

Neal, R. M,. (2003). Slice sampling. *The Annals of Statistics*, 31(3), pp. 705–767.

Nurul, A. B., et al., (2024). Clustering Mixed-Type Data via Dirichlet Process Mixture Models with Cluster-Specific Covariance Matrices. *Symmetry*, 16(6), pp. 712–732.

Patil, G. P., Rao, C. R., (1977). The weighted distributions: A survey of their applications. *Applications of Statistics*, 383.

Patil, G. P., Rao, C. R., (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, pp. 179–189.

Rao, C. R., (1985). Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In *A celebration of statistics*, pp. 543–569. Springer, New York, NY.

Rao, C. R., (1965). On discrete distributions arising out of methods of ascertainment. *Sankhya: The Indian Journal of Statistics, Series A*, pp. 311–324.

Rao, G. S., Aslam, M. and Kundu, D., (2015). Burr-XII distribution parametric estimation and estimation of reliability of multicomponent stress-strength. *Communications in Statistics - Theory and Methods*, 44(23), pp. 4953–4961.

Rodriguez, R. N., (1977). A guide to the Burr type XII distributions. *Biometrika*, 64(1), pp. 129–134.

Sethuraman, J., (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, pp. 639–650.

Walker, S. G., (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1), pp. 45–54.

Zelen, M., Feinleib, M., (1969). On the theory of screening for chronic diseases. *Biometrika*, 56(3), pp. 601–614.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 55–68, https://doi.org/10.59139/stattrans-2025-039
Received – 20.01.2023; accepted – 22.05.2025

Exploring new mixtures of distribution to model skewed and heavy tailed data

Jitendra Kumar[1], **Anuj Nain**[2]

Abstract

The search for relevant models that can describe the loss data has been one of the main interests of researchers for decades. There is limited research on modeling such data using K-component mixture models. An example of that can be Miljkovic and Grün's study (2016) of six distributions where they proposed finite mixtures to model the data. In this paper we study two more distributions, namely log-logistic and inverse Weibull distribution in addition of all those proposed by Miljkovic and Grün (2016). We employed the EM algorithm for parameter estimation and then selected the best model using three model selection criteria, namely NLL, AIC and BIC. We also computed the risk measures such as VaR and TVaR and compared them with their empirical counterparts to assess the goodness-of-fit of our proposed models at the extreme quantiles. We found that K-component mixture distribution of log-logistic and inverse Weibull works better than competent models. To get a more generalized view on the theory of mixture distribution, a simulation was carried out, which gave satisfactory results.

Key words: K-component finite mixture models, EM algorithm, Danish fire insurance losses data set, log-logistic distribution, inverse Weibull distribution.

1. Introduction

Many times, in real-life situations random variables are not generated from a single but from a mixture of several distributions. The mixture distribution is a weighted sum of K distributions where the weights sum up to one. Each weight corresponds to the proportion or contribution of the corresponding component density. Each density in the mixture is characterized by an unknown parameter (or a parameter vector).

In actuarial domain, searching for distributions that can be used to model a data which is heavy tailed, positively skewed, non-Gaussian and multimodal has gained lot

[1] Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India. E-mail: vjitendrav@gmail.com. ORCID: https://orcid.org/0000-0003-4473-4148.

[2] Department of Statistics, Central University of Rajasthan, Bandersindri, Ajmer, India. E-mail: anuj.nain.stats@gmail.com. ORCID: https://orcid.org/0009-0003-3678-5647.

of attention of the researchers. Mixture distribution can be well used to model such data. Mixture models that can cover the humps and the tail provide a good fit for the data, but the literature on such models is limited. A mixture of exponential distributions was proposed by Keatinge (1999). In this model the maximum likelihood (ML) estimation was based on Newton's algorithm. Although the model was applicable in some areas it was not very useful in modelling heavy-tailed data.

Klugman and Rioux (2006) proposed a flexible model that included not only exponential components but also Gamma, Pareto and Log-normal with non-negative weights that sum up to one, provided that either weight associated with Log-normal or Gamma component must be zero. Further, some work regarding this type of modeling was done by Lee and Lin (2010) and Verbelen *et al.* (2015, 2016). They considered finite mixture of Erlang distribution.

Miljkovic and Grün (2016), extended the work beyond Erlang families. They developed K-component mixtures of six models with components from non-Gaussian families of distributions. The six families of distributions they used were Burr, Gamma, Inverse Burr, Inverse Gaussian, Log-normal, and Weibull. In total, there are more than thirty distributions in their paper. They compared their results with composite Weibull models developed by Bakar *et al.* (2015). Their results outperformed the models introduced by Bakar *et al.* (2015). In this paper, we have extended the work of Miljkovic and Grün (2016) by introducing K-component finite mixtures of two more distributions, log-logistic and inverse Weibull.

While the present work focuses on the development and application of finite mixture models using heavy-tailed distributions, it is also worth mentioning that considerable theoretical work has been carried out to address the asymptotic behavior and estimation challenges in mixture models. Traditional likelihood-based methods, such as the likelihood ratio test (LRT), are known to exhibit non-standard asymptotic distributions due to violations of regularity conditions in mixture settings. Pioneering studies by Ghosh and Sen (1985) and Chernoff and Lander (1995) revealed that the asymptotic distribution of LRT in such contexts may involve the supremum of a Gaussian process, rather than a standard chi-square distribution. To address these complexities, Chen and Chen (1998a, b) developed adjusted LRT methods, and Chen *et. al* (2001) further proposed a modified LRT with desirable asymptotic properties and enhanced power under local alternatives. In addition, Chen and Kalbfleisch (1996) introduced penalized minimum-distance estimators, which yield consistent estimation of both the mixing distribution and the number of components. While these works do not directly align with the specific mixture distributions explored in this paper, their contributions to model identifiability, penalization, and asymptotic inference provide valuable methodological context.

It is also to be noted that although the present study builds upon the framework established by Miljkovic and Grün (2016), it is important to acknowledge that further research has been conducted in related modeling beyond 2016. For instance, Lachos Dávila et al. (2018) presented finite mixtures of skew-normal and skewed distributions, focusing on modeling skewness and kurtosis, rather than multimodality or tail-heaviness typical of actuarial data. Similarly, Gagnon and Wang (2024) developed robust heavy-tailed versions of generalized linear models, providing improved outlier resistance within a GLM framework rather than mixture modeling. Additionally, Marambakuyana and Shongwe (2024) explored a vast class of two-component composite and mixture models based on non-Gaussian distributions for insurance claims data; however, their emphasis was on empirical comparisons among predefined combinations, not on extending parametric families within finite mixtures.

While these studies contribute valuable insights, their modeling objectives and methodological approaches diverge from the specific structure of the finite mixture modeling considered in the present work. Therefore, the model proposed by Miljkovic and Grün (2016) remains the most relevant and appropriate foundation for the development pursued herein.

The chronology of the work presented in this study is as follows.

In Section 2.1 we have given a brief introduction to the mixture distribution and EM algorithm. In Section 2.2 and 2.3 we provide theory for the parameter estimates from log-logistic and inverse Weibull family respectively. In Section 2.4 we have discussed model selection criteria and risk measures.

In Section 3.1 we have given description of our data set and its characteristics. In Section 3.2 we provide our results. The mixture of log-logistic distribution is found to perform better than most of the models given by Miljkovic and Grün (2016) whereas those of inverse Weibull distribution completely outperform the best distributions given by Miljkovic and Grün (2016).

In Section 3.3, the K-S test is applied for checking the goodness-of-fit of the proposed mixture density. Plots of empirical density versus K-component mixture density are constructed to show how well the fitted density covers the empirical density, Q-Q plot and P-P Plot are also constructed to justify the goodness-of-fit. In Section 3.4, a simulation study is conducted on 50 samples of size 2492 each. The results of the simulation justify the fit of four component inverse Weibull distribution.

2. Mixture Distribution and EM Algorithm

2.1. Mixture Distribution

Let X_1 X_2 . . . X_n be an independent and identically distributed random sample from the K-component finite mixture of probability distributions. This mixture distribution is represented as

$$f\,(x; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(x;\ \theta_K\ )\ , \tag{1}$$

subject to $\sum_{k=1}^{K} \pi_k = 1$,

where $\boldsymbol{\Theta}$ = (π', θ') = (π_1 , π_2 ,. . . , π_{k-1} , θ_1 , θ_2 , . . . θ_k) is the vector of unknown parameters and $0 < \pi_i \leq 1$. These K distributions may or may not be from the same family. In this paper we assume that for the mixture density given in (1) the component densities $f_k(.)$ are from the same family. Further, we implement the EM algorithm for parameter estimation. For more details one may refer Dempster (1997). Miljkovic and Grün (2016) obtained parameter estimates using EM algorithm for the K-component finite mixture with component densities from the same family. Using their notations and terminology we mention here an important function called Q-function. For the details of the Q-function one may refer to Miljkovic and Grün (2016).

$$Q\,(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}_{old}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}[\log(\pi_k) + \log(f_k(x_i;\ \theta_k\ ))] \tag{2}$$

where w_{ik} is the expected value of the i^th^ weight in the k^th^ component density. This function is maximized to obtain new estimates of $\boldsymbol{\Theta}$ (θ and π). The estimates of π are updated in the s^th^ iteration by

$$\pi_k^{(s)} = \frac{\sum_{i=1}^{n} w_{ik}^{(s)}}{n} \tag{3}$$

For simplification we will write m component finite mixture distribution as m-K followed by the name of the distribution. For example, four component inverse Weibull distribution would be written as 4-K inverse Weibull distribution.

2.2. Parameter estimation in m-K log-logistic distribution.

The log-logistic distribution is widely used in actuarial and financial modelling. Its properties are such attractive that it acts as an alternative distribution to the lognormal and Weibull distribution. Finite mixture models of log-logistic distribution can be used for modelling heavy-tailed data on positive support. In this section we provide the parameter estimates of m-K log-logistic distribution through the E-M algorithm.

The log-logistic distribution has the following probability density function:

$$f(x) = \begin{cases} \dfrac{a\left(\frac{x}{b}\right)^{a-1}}{b\left(\left(\frac{x}{b}\right)^{a} + 1\right)^{2}} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

With shape parameter a > 0, scale parameter b > 0. Introducing the notion of finite mixture, we see that each k^th^ , (k = 1,2 ... m) component density in (1) follows log-logistic distribution with parameters a_k and b_k . Maximization of the Q function (2) with respect to a_k gives the following expression:

$$\sum_{i=1}^{n} w_{ki}\left(\frac{-a_k\left(\frac{x}{b_k}\right)\log\left(\frac{x}{b_k}\right)+a\log\left(\frac{x}{b_k}\right)+\left(\frac{x}{b_k}\right)^{a_k}+1}{a_k\left(\left(\frac{x}{b_k}\right)^{a_k}+1\right)}\right)=0 \tag{4}$$

maximization of the Q function with respect to b_k gives the following expression:

$$\sum_{i=1}^{n} w_{ki}\left(\frac{a_k\left(\frac{x}{b_k}\right)^{a_k}-1}{b_k\left(\left(\frac{x}{b_k}\right)^{a_k}+1\right)}\right) = 0 \tag{5}$$

Solving equations (4) and (5) we get the estimates of a_k and b_k

These equations can be solved using uniroot function in R. The function is useful in searching the root of a nonlinear equation. For details on the function one may refer to Brent (1973).

2.3. Parameter estimation in m-K inverse Weibull distribution

The inverse Weibull distribution is another distribution used in actuarial and income modelling. It has gain interest of the researchers in modeling heavy-tailed data on positive support appearing in finance and actuaries. In this section we provide the parameter estimates of m-K inverse Weibull distribution. The pdf of inverse Weibull distribution is given by

$$f(x) = \begin{cases} \dfrac{a\left(\frac{b}{x}\right)^{a} e^{-\left(\frac{b}{x}\right)^{a}}}{x} & x > 0, \\ 0 & \text{otherwise} \end{cases}$$

shape parameter a > 0, scale parameter b > 0. Introducing the notion of the finite mixture, we see that each k^th^ , (k= 1 ,2 ... m) component density in (1) follows log-logistic distribution with parameters a_k and b_k. Maximization of the Q function with respect to a_k gives the following expression:

$$\sum_{i=1}^{n} w_{ki}\left(-\left(\frac{b_k}{x}\right)^{a_k}\log\left(\frac{b_k}{x}\right)+\log\left(\frac{b_k}{x}\right)+\frac{1}{a_k}\right)=0 \tag{6}$$

Solving it using the numerical technique we get the estimate $\widehat{a_k}$ of a_k.

Maximization of the Q function with respect to b$_k$ gives the following expression:

$$\sum_{i=1}^{n} w_{ki}\left(a_k\left(\frac{1-\left(\frac{b_k}{x}\right)^{a_k}}{b_k}\right)\right) = 0 \tag{7}$$

Solving equations (6) and (7) we get the estimates of a$_k$ and b$_k$.

2.4. Model Selection and Risk Measure

In order to access the goodness-of-fit of the proposed models, the following measures are considered: Negative log-likelihood (NLL), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Let us denote log-likelihood function by L(θ) for a given model, then NLL is computed as -L(θ). AIC is given by −2L(θ) + 2p, where p is the number of parameters present in the model and BIC is given by −2L(θ) + plog(n) , where n is the number of observations.

Let X be the random variable denoting the loss and π_p is the 100pth percentile of the distribution of X. Then according to the notations and definitions by Klugman *et al.* (2012), the Value at Risk of X at the 100p% level, denoted by VaRp (X) is the 100pth percentile of the distribution of X. Let $F_X(x)$ be the cumulative distribution function of X, then for continuous distributions, VaRp (X) can be written as the value of π_p satisfying

$$F_X(\pi_p) = \text{p}, \tag{8}$$

VaRp (X) lacks additive property, therefore we define TVaRp (X), which is the average of all VaRp (X) values above the security level p, Again using the notations and definitions by Klugman *et al.* (2012) . TVaRp (X) is given by

$$\text{TVaR}p(X) = \frac{\int_p^1 \text{VaR}_u(X)du}{1-P}$$

When X is a continuous random variable, the formula simplifies to

$$\text{TVaR}p(X) = E\,[\,X\,|X > \text{VaR}_P(X)]$$

We can use convenient notations [see Miljkovic and Grün (2016)] as

$$\text{TVaR}p(X) = E\,[\,X\,|X > \pi_p\,] = \frac{\int_{\pi_p}^1 xf(x)dx}{1-F_X(\pi_p)} = \frac{\int_{\pi_p}^1 xf(x)dx}{1-P}$$

For the sake of convenience we write VaRp (X) as VaR as TVaR$p(X)$ as TVaR.

3. Analysis

3.1. Data

In this paper, we have used the Danish fire losses data set. This globally recognized dataset has been used by actuaries for actuarial modeling for more than two decades.

The Copenhagen Reinsurance company was responsible for collecting the data. The data set contains a record of 2492 observations (loss claims in Danish Krone) over the period of 1980-1990 [see Miljkovic and Grün (2016)].

One can easily access the Danish fire losses data set (available as *danish*) in R by using the package **SMpracticals** [Davison (2013)]. Figure 1 shows the histogram of the data and Table 1 gives the summary statistics.

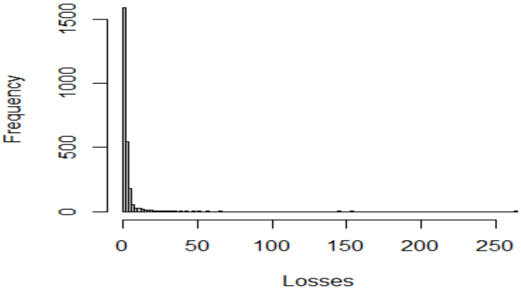


Figure 1. Histogram for Danish fire losses data set

Table 1. Summary Statistic

Min	Q1	Q2	Q3	Max	Mean
0.3134	1.1572	1.6339	2.6445	263.2504	3.0627

It can be observed that the mean is greater than the median and the third quartile (Q3), which indicates high skewness in the data. We consider modelling this data set using finite mixtures of various components for two distributions: log-logistic and inverse Weibull.

3.2. Results and discussions

In this section we have compared our work with work done by Miljkovic and Grün (2016). They considered six families of distributions and modeled their different K-component mixtures (in total there are 33 distributions). These six families were suitable for a data set having characteristics as present in the Danish fire losses data. However, two distributions, namely the log-logistic and the inverse Weibull remained unexplored in their study. We have extended their work by considering these two families, and modeling their K-component mixtures. The results are shown in the following table (Table 2).

Table 2. Model selection criteria for different component mixtures

Mixture	K(component)	NLL	AIC	BIC
log-logistic	1	4280.587	8565.175	8576.816
	2	3962.067	7934.133	7963.238
	3	**3850.580**	**7717.161**	**7753.728**
	4	3952.919	7927.838	7991.868
	5	3953.138	7936.276	8064.163
inverse Weibull	1	3966.83	7937.661	7949.302
	2	3842.314	7694.628	7723.733
	3	3793.387	7602.774	7649.341
	4	**3777.205**	**7576.409**	**7640.439**
	5	3773.208	7574.417	7655.908
	6	3771.172	3572.344	7663.482

It is observed that the proposed 3-K log-logistic distribution performs better than most of the models developed by Miljkovic and Grün (2016) (The parameter estimates have been presented in Figure 2). It performs better than 1-K, 2-K and 3-K models from the Gamma family, 1-K model from the Inverse Burr family, 1-K, 2-K and 3-K model from Inverse Gaussian, 1-K, 2-K and 3-K models from the Log-Normal family and 1-K, 2-K, 3-K, 4-K, 5-K models from the Weibull family.

Also, we observe that proposed 4-K inverse Weibull distribution performs better than all the distributions proposed by Miljkovic and Grün (2016) (The parameter estimates have been presented in Figure 4). (It is to be noted that we have reproduced the results for the six distributions studied by Miljkovic and Grün (2016). As these results were consistent with the original findings, we chose not to present them explicitly in the paper in order to avoid redundancy and to maintain focus on our novel contributions. However, for comparison one can refer to the table from Miljkovic and Grün (2016)).

The following Table 3 gives the VaR and TVaR risk measures.

Table 3. Risk measures

Specification	VaR(0.99)	TvaR(0.99)
Empirical Estimates	24.61	54.60
Proposed Models		
3-K log-logistic	20.33	31.50
4-K inverse Weibull	24.67	70.55

VaR from the proposed 3-K log-logistic (20.33) and that from 4-K inverse Weibull (23.57) are close to the empirical VaR (24.61). This adds to the goodness-of-fit of our proposed models. VaR from 4-K inverse Weibull is better than that from 2-K Burr

(25.02), 5-K Log Normal (26.75) and 3-K Inverse Burr (23.57), which were the best three models proposed earlier by Miljkovic and Grün (2016) [see Miljkovic and Grün (2016)]. This means that 4-K inverse Weibull covers tail of the data better than 2-K Burr, 5-K Log Normal and 3-K Inverse Burr. Due to presence of high skewness at the tail of the distribution it was obvious that TVaR from the distribution would be different from the empirical counterpart.

3.3. Goodness-of-fit

We conducted the Kolmogorov-Smirnov test (known as K-S test) for checking the goodness-of-fit of the proposed models. The null hypothesis states that the proposed distribution (either 3-K log-logistic or 4-K inverse Weibull) fits the data well. The significance level was set to .05. In the case of 3-K log-logistic distribution the test statistic was less than the critical value and hence the null hypothesis was accepted. In the case of 4-K inverse Weibull distribution also, the test statistic was less than the critical value and hence the null hypothesis was accepted. The results are summarized in Table 4.

Table 4. Goodness-of-fit test (H_0:The proposed distribution fits the data well)

Distribution	K-S Test Statistic	Critical Value (at α= 0.05)	Decision
3-K log-logistic	0.014	0.027	H_0 accepted
4-K inverse Weibull	0.009	0.027	H_0 accepted

Figure 2 and Figure 3 are for 3-K log-logistic distribution and show how well 3-K log-logistic distribution covers empirical density. Figure 2 shows the plot of empirical density against fitted density, Figure 3 shows the P-P plot and the Q-Q plot.

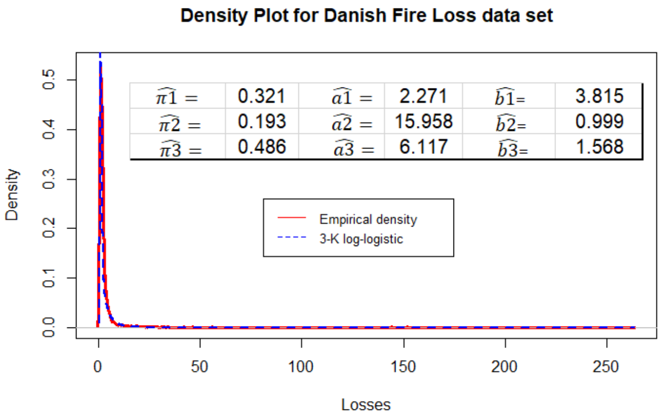


Figure 2. Plot of Empirical density and Fitted 3-K log-logistic density

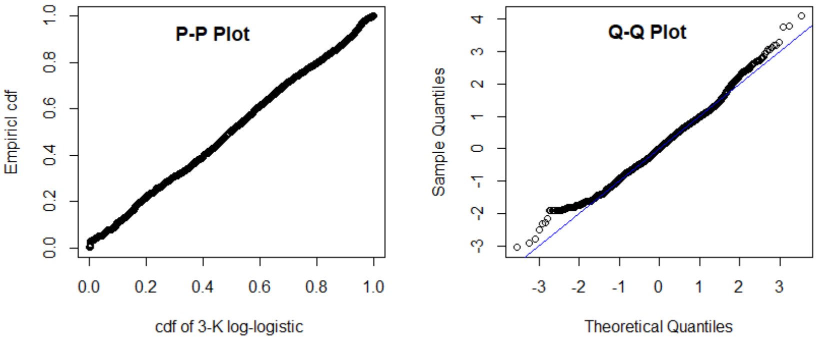
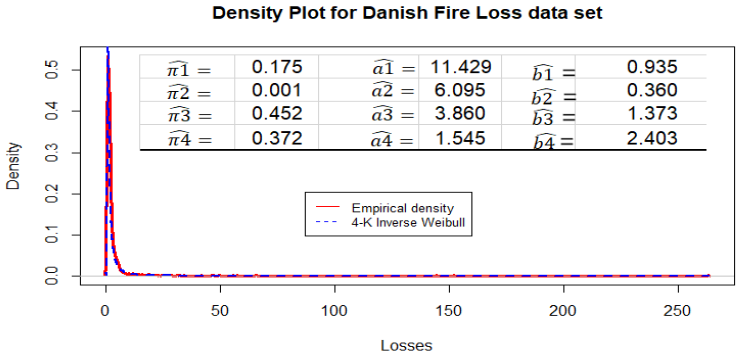


Figure 3. P-P plot and Q-Q Plot for the fitted 3-K log-logistic distribution

Figure 4 and Figure 5 are for 4-K inverse Weibull distribution and show how well the 4-K inverse Weibull density covers the empirical density. Figure 4 shows the plot of empirical density against fitted density, Figure 5 shows the P-P plot and the Q-Q plot.



$\widehat{\pi 1} =$	0.175	$\widehat{a1} =$	11.429	$\widehat{b1} =$	0.935
$\widehat{\pi 2} =$	0.001	$\widehat{a2} =$	6.095	$\widehat{b2} =$	0.360
$\widehat{\pi 3} =$	0.452	$\widehat{a3} =$	3.860	$\widehat{b3} =$	1.373
$\widehat{\pi 4} =$	0.372	$\widehat{a4} =$	1.545	$\widehat{b4} =$	2.403

Figure 4. Plot of empirical density and fitted 4-K inverse Weibull density

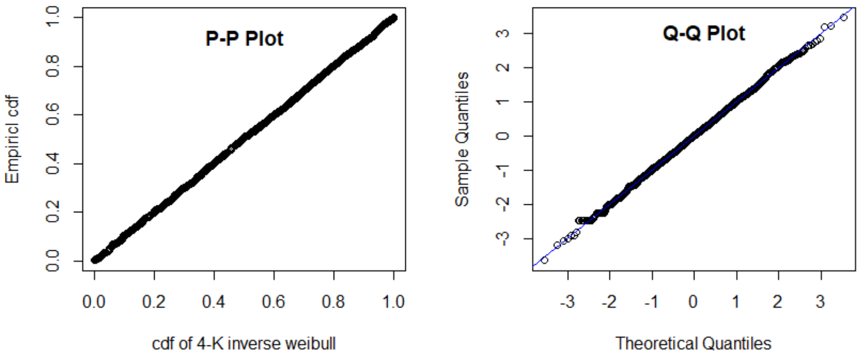


Figure 5. P-P plot and Q-Q Plot for the fitted 4-K inverse Weibull distribution

3.4. Simulation study

We conducted simulation for 4-K inverse Weibull distribution. We generated N=50 random samples from the given data of size n=2492 each, that can mimic the actual Danish fire losses data set. Then estimated the parameters, computed the VaR, constructed the density plot, conducted K-S test for goodness-of-fit and constructed P-P plot and Q-Q plot. We found that **4-K inverse Weibull** distribution fits each of the simulated data set well, as the Empirical VaR and fitted VaR from the distribution were very close to each other and the value of K-S test statistic was less than the critical value for each simulated sample. The table below (Table 5) shows the results for first ten simulated samples.

Table 5. Summary of results for the first ten simulated samples.

Sample No.	Empirical VaR(0.99)	Fitted VaR(0.99)	K-S Test statistic to be compared with 0.0272
1	25.288	23.59	0.012
2	21.989	23.1	0.010
3	25.059	24.33	0.013
4	26.309	29.34	0.020
5	20.873	22.12	0.011
6	25.954	26.03	0.012
7	22.465	21.54	0.012
8	23.900	23.661	0.013
9	20.963	23.09	0.012
10	22.351	21.7	0.011

As an example of simulation result we have shown Density plot, P-P Plot, Q-Q Plot for the first two simulated samples.

Simulated sample 1.

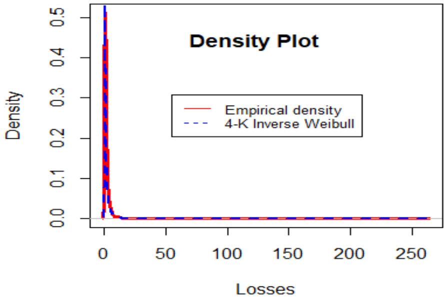


Figure 6. Plot of empirical density and simulated 4-K inverse Weibull density

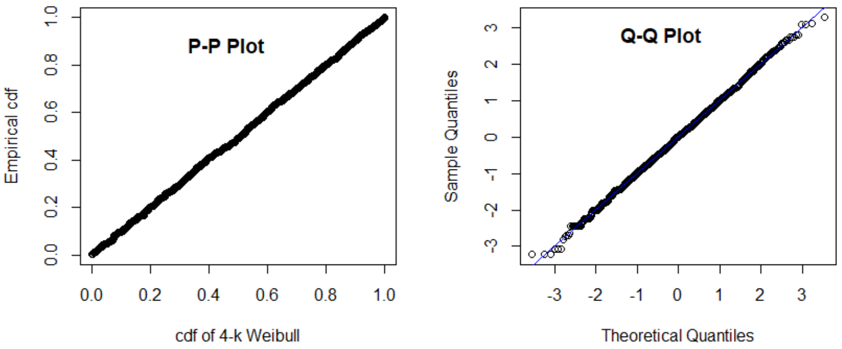


Figure 7. P-P plot and Q-Q Plot for the simulated 4-K inverse Weibull distribution

Simulated sample 2.

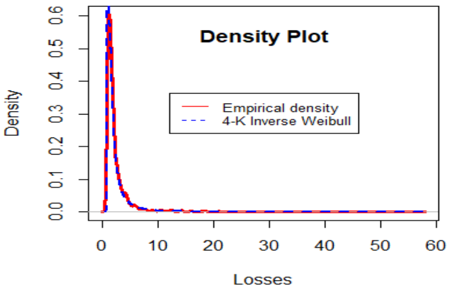


Figure 8. Plot of empirical density and simulated 4-K inverse Weibull density

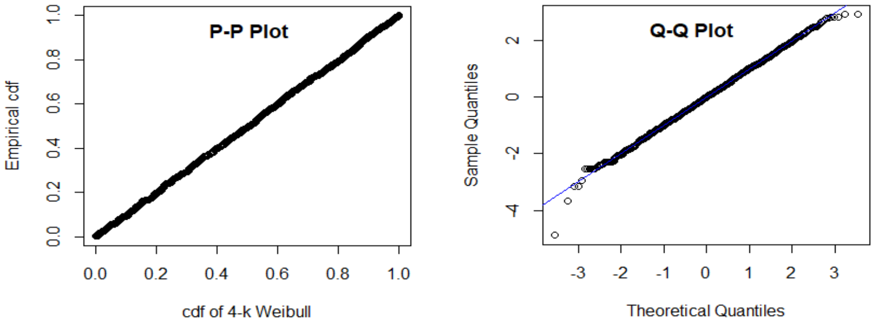


Figure 9. P-P plot and Q-Q Plot for the simulated 4-K inverse Weibull distribution

We conducted two sample t-tests to check the significance of difference between the Empirical VaR and VaR computed from the fitted 4-K inverse Weibull distribution. We formulate the null hypothesis, H_0: There is no significant difference between the empirical VaR and VaR computed from the fitted 4-K inverse Weibull distribution. We see that the null hypothesis is accepted at the significance level of 0.05. The result summary of the test is given in Table 9.

Table 9. Checking the significance of difference between the Empirical VaR and fitted VaR

Specification	
Test statistic	1.755
Critical Value	1.984
P-value	0.082
Level of significance	0.05

4. Conclusion

In this paper, we extended the work done by Miljkovic and Grün (2016) by adding and examining K-component finite mixtures of two more distributions: log-logistic and inverse Weibull.

We found that 3-K log-logistic distribution performed better than most of the distributions given earlier by Miljkovic and Grün (2016), whereas 4-K inverse Weibull distribution outperformed all the distributions given by Miljkovic and Grün (2016).

Hence, we conclude that for skewed and non-Gaussian data on positive support, finite mixtures of log-logistic and inverse Weibull distribution can successfully cover the heavy tail behavior and multimodal nature. To the best of our knowledge, we are among the recent after Miljkovic and Grün (2016) to work on K-component finite mixture model on such type of loss data.

We considered finite mixtures from the same families in our paper. Further work may be done by considering finite mixtures from different families. One may also use the Bayesian approach for parameter estimation.

References

Akaike, H., (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, Vol.19, pp. 716–723.

Bakar, S. A., Hamzah, N. A., Maghsoudi, M. and Nadarajah, S., (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics*, Vol. 61, pp. 146–154.

Brent, R., (1973). Algorithms for Minimization without Derivatives. Englewood Cliffs. NJ: *Prentice-Hall.*

Chen, H., Chen, J., (1998a). The likelihood ratio test for homogeneity in the finite mixture models. Technical Report STAT 98–08. Department of Statistics and Actuarial Science. *University of Waterloo*, Waterloo.

Chen, H., Chen, J., (1998b). Tests for homogeneity in normal mixtures with presence of a structural parameter. Technical Report STAT 98–09. Department of Statistics and Actuarial Science. *University of Waterloo*, Waterloo.

Chen, H., Chen, J. and Kalbfleisch, J. D., (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(1), pp. 19–29.

Chen, J., Kalbfleisch, J. D., (1996). Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, *24*(2), pp. 167–175.

Chernoff, H., Lander, E., (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of Statistical Planning and Inference*, 43(1-2), pp. 19–40.

Dávila, V. H. L., Cabral, C. R. B. and Zeller, C. B., (2018). Finite mixture of skewed distributions. Germany: *Springer International Publishing.*

Davison, A.,(2013). SMPracticals: Practicals for Use with Davison (2003), Statistical Models. *R package version*, pp. 1–4.

Dempster, A. P., Laird, N. M. and Rubin, D. B., (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 19, pp. 1–22.

Gagnon, P., Wang, Y., (2024). Robust heavy-tailed versions of generalized linear models with applications in actuarial science. *Computational Statistics & Data Analysis*, 194, p. 107920.

Ghosh, J. K., Sen, P. K., (1984). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results.

Keatinge, C. L., (1999). Modeling losses with the mixed exponential distribution. In *Proceedings of the Casualty Actuarial Society*, Vol. 86, pp. 654–698.

Klugman, S. A., Panjer, H. H. and Willmot, G. E., (2012). Loss models: from data to decisions. *John Wiley & Sons.*

Klugman, S., Rioux, J., (2006). Toward a unified approach to fitting loss models. *North American Actuarial Journal*, Vol. 10, pp. 63–83

Lee, S. C., Lin, X. S., (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, Vol. 14, pp. 107–130.

Marambakuyana, W. A., Shongwe, S. C., (2024). Composite and mixture distributions for heavy-tailed data—An application to insurance claims. *Mathematics*, 12(2), p. 335.

Miljkovic, T., Grün, B., (2016). Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, Vol. 70, pp. 387–396.

Verbelen, R., Antonio, K. and Claeskens, G., (2016). Multivariate mixtures of Erlangs for density estimation under censoring. *Lifetime data analysis*, Vol. 22, pp. 429–455.

Verbelen, R., Gong, L., Antonio, K., Badescu, A. and Lin, S., (2015). Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin: The Journal of the IAA*, Vol. 45, pp. 729–758.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 69–87, https://doi.org/10.59139/stattrans-2025-040
Received – 13.07.2023; accepted – 21.03.2025

Application of the nonlinear splines model to forecast changes in the construction costs index

Mariusz Kubus[1], **Łukasz Mach**[2], **Przemysław Misiurski**[3]

Abstract

The research presents the application of the nonlinear splines model to forecast the construction costs index (CCI), which is an important macroeconomic indicator. Due to the long-term nature of the investments in the construction market, we tested our model in a ten-month ahead period. Except minor disruptions, which were likely related to COVID-19, we obtained promising results, which definitely outperformed the classical ARIMA and its variant with nonlinear autocorrelation functions modeled with neural network. The achieved forecast results will enable both the demand and supply in the construction market to be in market equilibrium and minimize the formation of speculative bubbles in the market.

Key words: construction market, nonlinear splines model, forecast, non-stationary time series.

1. Introduction

The market economy of each developed country is based on three factors of production – labor, land and capital. Each of these factors should be developed evenly so that sustainable economic development can occur. A comprehensive study of a country's economic development can be carried out by monitoring an important sector for the economy – the construction sector. By its very nature, this sector reflects the condition and state of development of the three aforementioned factors of production. It is natural that in considering investment projects in the area of the construction sector we must consider the stock of land, labor and capital. The integration of production factors in official statistics is expressed through the construction costs index, since it is

[1] Centre of Education and Mathematics Applications, Opole University of Technology, Opole, Poland. E-mail: M.Kubus@po.edu.pl. ORCID: https://orcid.org/0000-0002-6602-2742.

[2] Institute of Economics and Finance, Faculty of Economics, University of Opole, Opole, Poland. E-mail: L.Mach@uni.opole.pl. ORCID: https://orcid.org/0000-0002-8200-4261.

[3] Department of Enterprise Management, E-business and Electronic Economy, Opole University of Technology, Opole, Poland. E-mail: P.Misiurski@po.edu.pl. ORCID: https://orcid.org/0000-0002-7052-8535.

this index that will reflect directly the capital and indirectly the labor and land required to use this capital to implement the project. In order to plan the investment projects successfully, it is necessary to use data analysis methods capable of forecasting this index. Therefore, the research carried out in this article has tested and demonstrated the usefulness of the nonlinear splines method as a utilitarian forecasting tool, both for stationary and non-stationary time series.

The literature on the construction market and its importance for the economy is not as rich as it might seem. It is dominated mainly by country-specific reports and analyses. On the other hand, Asian researchers seem to dominate theoretical considerations, which seems justified due to the intensive development of construction in these countries. Three-level definition of construction as an economic activity was presented by researchers from Singapore (Pheng & Hou 2019). The issue of the role of construction in the economy was described by Jorge Lopes in the chapter: *Construction in the economy and its role in socio-economic development: role of construction in socio-economic development* (Lopes 2011). An attempt to define the scope of the construction sector was made by Andrew Foulkes and Les Ruddock of the British University of Salford. The latter edited a book on the economic aspects of modern construction (Ruddock & Ruddock 2008). There have also been studies on the importance of the construction market in the economy of specific countries. One can mention here, for example, the publication of scientists from Malaysia regarding Turkmenistan: *Role of the construction industry in economic development of Turkmenistan* (Durdyey & Ismail 2012) or the publication on the Polish economy: *The view of construction companies' managers on the impact of economic, environmental and legal policies on investment process management* (Sobieraj *et al.* 2021). On the other hand, a broader view of the construction market was presented by Chinese researchers: Ye, Lu and Jiang in the article *Concentration in the international construction market* (Ye *et al.* 2009).

A literature analysis of the construction market showed that the topic is important and interesting, but there are not many research works that comprehensively deal with construction markets. This shows that research in this area is needed.

2. Forecasting Construction Cost Index: A Literature Review

Accurate forecasting of the Construction Cost Index (CCI) is essential for effective project planning, cost control, and risk management in the construction industry. Over the past five decades, various forecasting methodologies have been explored, evolving from traditional econometric models to advanced machine learning techniques. This section reviews the key studies on CCI forecasting, with a particular focus on applied methodologies and their effectiveness.

Early research in this field primarily relied on econometric and regression-based models to analyze historical cost trends. One of the pioneering studies was conducted by Williams (1994), who applied neural networks to predict changes in construction cost indices. Wang and Mei (1998) further explored forecasting models for Taiwan's CCI using time series techniques. In subsequent years, autoregressive (AR) and autoregressive integrated moving average (ARIMA) models gained prominence due to their ability to model cost index trends based on past values (Xu & Moon, 2013; Moon, Chi & Kim, 2018). A notable advancement came from Moon and Shin (2018), who introduced the vector error correction model (VECM), capturing long-term equilibrium relationships among cost-related variables. However, these statistical models were constrained by their linear nature, limiting their ability to capture complex, nonlinear patterns in cost fluctuations.

With the advancement of computational techniques, machine learning-based models have emerged as effective tools for CCI forecasting. Elfahham (2019) demonstrated that artificial neural networks (ANNs) could outperform traditional regression and time series methods. Wang and Ashuri (2017) explored machine learning algorithms to enhance CCI predictions, highlighting their adaptability to dynamic construction markets. To improve accuracy, researchers have increasingly integrated classical time series models with machine learning techniques. For example, Kim et al. (2022) combined ARIMA and ANN models to enhance prediction precision. Similarly, Cao and Ashuri (2020) applied a long short-term memory (LSTM) network to predict the volatility of highway construction costs, outperforming traditional forecasting models.

Recent studies have also explored stochastic and network-based forecasting techniques to address the dynamic nature of construction costs. Xu and Moon (2013) utilized a cointegrated vector autoregression (VAR) model for stochastic CCI forecasting, while Joukar and Nahmens (2016) employed a generalized autoregressive conditional heteroskedasticity (GARCH) model to analyze cost index volatility patterns. Additionally, network-based approaches have gained traction as an alternative forecasting method. Zhang et al. (2018) introduced a visibility graph-based forecasting model, while Mao and Xiao (2019) developed a complex network-based approach to improve predictive performance.

The latest research in CCI forecasting has increasingly focused on hybrid methodologies and deep learning techniques. Al Kailani et al. (2024) utilized fuzzy logic and machine learning to enhance prediction accuracy in Jordan's construction sector. Similarly, Altalhoni, Liu, and Abudayyeh (2024) conducted a comprehensive review of existing forecasting techniques, identifying key influential factors and emerging trends.

These studies indicate that artificial intelligence and nonlinear models will continue to play a pivotal role in improving predictive accuracy.

Given the existing literature, it is evident that forecasting methodologies have evolved significantly from traditional statistical models to advanced AI-driven approaches. However, nonlinear modeling techniques, such as splines, remain underexplored in CCI prediction. This study aims to fill this research gap by introducing a nonlinear splines-based model to forecast changes in the construction cost index. By leveraging the flexibility of splines in capturing complex, nonlinear relationships, this approach seeks to enhance forecasting accuracy and provide valuable insights for cost management in the construction industry. This literature review highlights the importance of innovative forecasting methods, reinforcing the relevance of the proposed nonlinear splines model in the broader research landscape.

3. The use of spline methods in research – examples of application

In the field of economics and management, various methods of data analysis are used in the decision-making process. These methods are useful for creating models on the basis of which events can be simulated with the possibility of creating different scenarios of the event and are helpful in the decision-making process. In economics, methods included in the widely understood econometric modelling, as well as the use of artificial intelligence tools, are generally used for this purpose. Commonly used methods include, for example, correlation, regression, econometric modelling, MA, VAR, or ARIMA methods, cf. (Findley et al. 2016; Nieto & Carmona-Benítez 2018; Osiewalski & Osiewalski 2013; Paci & Consonni 2020). In the group of artificial intelligence tools, methods based on neural networks or genetic algorithms are commonly used, cf. (Butler et al. 2021; Cheung et al. 2006; Lin et al. 2021). In the area of modelling economic phenomena, mathematical or simulation methods reserved for technical sciences, i.e. wavelet transform or analysis of the coherence of the wavelet, are relatively rarely used, cf. (Dash & Maitra 2019; Naccache 2011).

In the literature, it is also possible to find few studies showing the usefulness of such a tool as a spline. Studies using the methods of spline can be found in the work of Humphrey & Vale (2004), which demonstrates the usefulness of an elastic cost function when analyzing economies of scale and estimating the cost-effectiveness of bank mergers. The inflexibility of the translogarithmic cost function was pointed out, and the results obtained were compared with more flexible cost functions of the folded curve and Fourier curve types. Using these different approaches, an ex-ante effect on the costs resulting from mergers between 1987 and 1998 was projected using a panel study on a sample of 130 Norwegian banks. Mergers are predicted to result in an average reduction in costs. Predictions using the Fourier approach or compound curves are consistent with calculated actual changes in ex-post-merger costs. Other studies, described in (Monteiro et al. 2008), present a new approach to estimating the risk-neutral probability density function of the future prices of an underlying asset from the prices of options written on the asset. The estimation is carried out in the space of cubic spline functions, yielding appropriate smoothness. The resulting optimization problem, used to invert the data and determine the corresponding density function, is a convex quadratic or semi-definite programming problem, depending on the formulation. Both of these problems can be efficiently solved by numerical optimization software. Monteiro et al. (2008) tested their approach using data simulated from Black–Scholes option prices and using market data for options on the S&P 500 Index. Their results show the effectiveness of the proposed methodology for estimating the risk-neutral probability density function. The application of spline-based phase analysis to macroeconomic dynamics can be found in the work of Gadasina & Vyunenko (2022), in which they use spline-based phase analysis to study the dynamics of a time series of low-frequency data on the values of a certain economic indicator. The approach includes two stages. In the first stage, the original series is approximated by a smooth twice-differentiable function, which is obtained from natural cubic splines. Such splines have the smallest curvature over the observation interval compared to other possible functions that satisfy the choice criterion. In the second stage, a phase trajectory is constructed in the space, which corresponds to the original time series, and a phase shadow as a projection of the phase trajectory onto the plane. The approach is applied to the values of GDP indicators for the G7 countries. The interrelation between phase shadow loops and cycles of economic indicators evolution is shown. The study also discusses the features, limitations, and prospects for the use of spline-based phase analysis (Gadasina & Vyunenko 2022).

On the other hand, the effectiveness of steering processes of complex socio-economic systems using the method of spline analysis has been demonstrated by Ilyasov & Yakovenko (2021), and the utility of tensor spline approximation in economic dynamics with uncertainties was presented in the paper (Chu et al. 2011). Of course, one can also encounter applications of compound curves in fields other than economic science. Examples include studies describing spline functions as an alternative to estimating income-expenditure relationships for beef (Huang & Raunikar 1981), describing evaluating the water sector in Italy through a two-stage method using the conditional robust nonparametric frontier and multivariate adaptive regression splines (Vidoli 2011) or forecasting energy demand in transportation by using the multivariate adaptive regression splines (Sahraei *et al.* 2021).

4. Research methodology

4.1. Data and general outline of the research

The data used for the study comes from Eurostat[4] databases and include monthly construction costs (or producer prices) index for Poland. This index is related to new residential buildings, except residences for communities. The data covers the period from January 2000 to June 2022 and the unit of measure is a percentage change on previous period. Since investments in construction are long-term in nature, our model is going to be assessed in ten-month ahead forecast. Therefore, we divided the data into 260 months for training and the rest 10 months for testing purposes. Thus, the last 10 observations were compared to the forecast of the model. This approach somehow simulates the usefulness of the model in the future. Figure 1 depicts monthly construction costs index in the training set.

Having a wide range of time series modelling techniques, from the popular ARIMA approach to artificial intelligence tools, we chose the spline model, which is one of the non-linear and local regression methods. The variant we used, namely smoothing splines, does not require setting the hyper-parameters of the method. This would demand some background knowledge or the use of a validation set, which does not always lead to a stable solution. All of that make this method a convenient tool for practitioners.

Before we fitted the model, the data was pre-processed. At this stage, we dealt with outliers and transformed the data to stabilize the variance. After fitting the model, we performed a residual analysis and then assessed the accuracy of the forecast on the test set. The result was compared to ARIMA, which is one of the most widely used forecasting methods for univariate time series analysis, as well as to neural network variant of ARIMA (Figure 2).

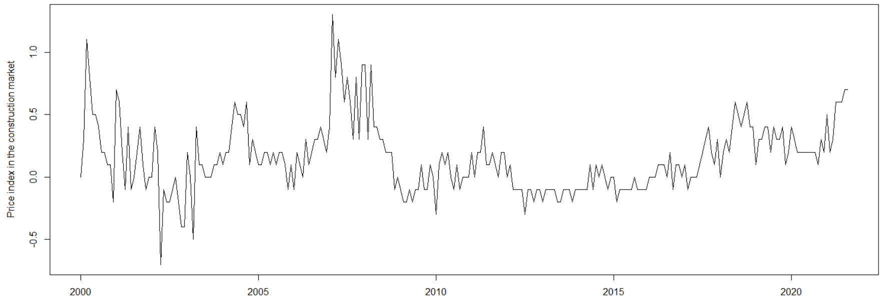


Figure 1. Price index in the construction market up to August 2021 (training set)

[4] https://ec.europa.eu/eurostat/data/database (04.01.2023).

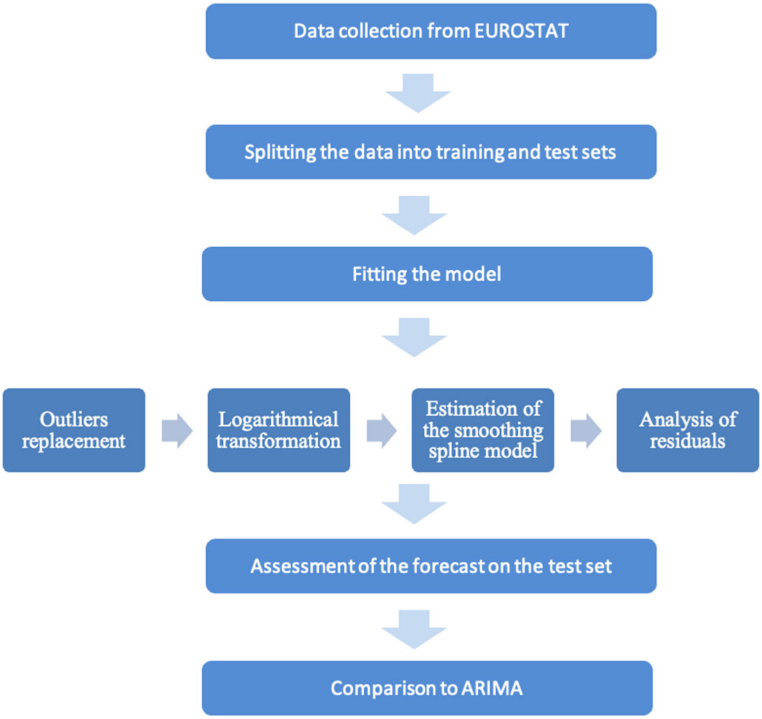

Data collection from EUROSTAT
Splitting the data into training and test sets
Fitting the model
Outliers replacement
Logarithmical transformation
Estimation of the smoothing spline model
Analysis of residuals
Assessment of the forecast on the test set
Comparison to ARIMA

Figure 2. Scheme of the research

All computations were carried out using R packages {tseries} and {forecast}. R is an open source environment freely available on the web site: www.r-project.org. This software is a collaborative project with many contributors and supported by world research community.

4.2. Nonlinear splines model

Nonlinear regression methods have a key meaning in econometric modelling. They found numerous applications for both data independent of time and time series. The popularity they have achieved is because they go beyond simple linear dependencies, which are present in many real domains. The nonlinear approach serves a possibility of extremely accurate data fitting but it is a well-known fact that these kinds of models can be overfitted and, as a result, they usually perform poorly on unseen data. In terms of time series' forecasting, we say that *ex-ante* prognoses are inaccurate. This is unacceptable when prediction is the goal, and that is the case in most economic applications. To deal with the problem of overfitting, the vast majority of regression methods introduce regularization into the model selection stage. Instead of minimizing the quadratic loss function, which consequently results in the best possible fit, the criterion contains a penalty for model complexity. The core idea behind this approach is the trade-off between the accurate fit and the number of model parameters. Simpler models that do not fit the training data perfectly usually perform better on unseen data, i.e. on the forecast horizon in the case of time series. However, an overly simplified model does not reflect a systematic factor in the data and is naturally unlikely to be accurately predicted. The common approach is to use information criteria like AIC, AICc, or BIC to compare candidate models and to choose the final one. In more complex regression methods, i.e. splines, the regularization criterion is embedded in the algorithm of finding the final function formula. This is the case in the regression splines method which we present in this paper.

Generally, regression models can be divided into global or local. In the first case, one curve is fitted to the data in the entire domain, whereas in the second, the domain is partitioned into regions and the regression functions are approximated separately in each of them. Due to the context of the time series we focus on regression with one independent variable, which is time.

The local approach to modelling the regression curve requires choosing K points $\{\xi_k\}$ that divide the domain of X into $K + 1$ contiguous intervals. The points $\{\xi_k\}$ are called knots and it is a set that ξ_1 and ξ_K are equal to the minimal and maximal observation of variable X respectively. In each interval, function f_k is fitted to the data separately. In general, this function can take any form, but a frequent choice is polynomial. Since the high-order polynomial not only tends to overfit but causes estimation problems due to a large number of parameters, a usual choice is degree 3.

The basic problem of a piecewise polynomial function is continuity, thus the constraints:

$$f_k(\xi_k) = f_{k+1}(\xi_k), \tag{1}$$

for each $k \in \{1, \dots, K\}$ are imposed on the knots. Additionally, imposing the equality of the first and second derivatives in the knots:

$$f'_k(\xi_k) = f'_{k+1}(\xi_k); \;\; f''_k(\xi_k) = f''_{k+1}(\xi_k), \tag{2}$$

one can obtain a smooth regression line. That sort of regression models is known as cubic splines. The model has a form of the linear combination of the basis functions:

$$f(X) = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3 + \sum_{k=1}^{K} \beta_k (X - \xi_k)_+^3, \tag{3}$$

where

$$(X - \xi)_+^3 = \begin{cases} (X - \xi)^3 & for \;\; X > \xi \\ 0 & otherwise \end{cases} \tag{4}$$

and its parameters can be estimated by the least square method. Having given the number of the knots, they are usually set uniformly in quantiles of variable X. As the confidence bands can be relatively wide in the boundary regions, an additional constraint can be imposed. The function is to be linear for $X < \xi_1$ and for $X > \xi_K$. This model is known as a natural cubic spline.

The problem of determining the number and position of knots is naturally solved in the smoothing splines. They simply use all realizations of variable *X* as the knots, and parameters of the model (3) are estimated by the minimization of the criterion:

$$\sum_{i=1}^{N}(y_i - g(x_i))^2 + \lambda \cdot \int_D (g''(t))^2 dt, \tag{5}$$

where *D* is a range of *t*, and λ is a regularization parameter. The first term is a quadratic loss function, which reflects how well the model fits the observed data. The second one is a penalty for a large variability in function *g*. Interestingly, Hastie *et al.* (2009) showed that the function that minimizes the criterion (5) is a natural cubic spline with knots in the observations of variable *X*. The absolute value of the second derivative is large when function g is very wiggly near *t*, and otherwise, for a straight line that is perfectly smooth, the second derivative is equal zero. The amount of the penalty is regulated by the λ parameter, which takes values from zero to infinity. When $\lambda = 0$, there is no penalty and the function *g* accurately approximates data points. On the other hand, when $\lambda \to \infty$, the regression curve obtains perfect smoothness, and thus it becomes a straight line. Having taken the intermediate values of λ, one has a model that is a compromise between exact fit and amount of smoothness. Setting this parameter has a crucial role in modelling and yielding accurate predictions (or prognosis in the case of time series). The common solution for performing this task is cross-validation. For small samples or one period ahead forecasting, the LOOCV version (leave-one-out cross-validation) is a natural choice:

$$CV(\hat{g}_\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i - \hat{g}_\lambda^{-i}(x_i)\right]^2. \tag{6}$$

In this case, the computations demand *N* models to be built, therefore in practice reformulation of the CV estimate of the mean square error is used:

$$CV(\hat{g}_\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - S_\lambda(i,i)}\right]^2, \tag{7}$$

where $S_\lambda(i,i)$ are diagonal elements of $\boldsymbol{S}_\lambda$, which is the projection matrix in linear fit $\hat{\boldsymbol{y}} = \boldsymbol{S}_\lambda \boldsymbol{y}$. The last formula allows computing LOO criterion with only one fit. The construction of $\boldsymbol{S}_\lambda$ matrix is somewhat technical. As mentioned before, the solution of (5) has a natural spline representation. Given *N* basis functions, the criterion (5) can be expressed in matrix form and transformed to projection form (for more details see

(Hastie *et al.* 2009)). Replacing the diagonal elements with their mean in (7) one can obtain an approximation that is known as generalized cross-validation:

$$GCV(\hat{g}_\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{y_i - \hat{g}_\lambda(x_i)}{1-\frac{trace(S_\lambda)}{N}}\right]^2. \tag{8}$$

The trace of matrix $\boldsymbol{S}_\lambda$ is called the effective degrees-of-freedom and it controls the complexity of the model. Note that the GCV criterion is widely used in multivariate adaptive regression splines (Friedman 2009).

5. Analysis of the results

In this section, we intend to model the costs index on the construction market. Our purpose is to build a model that could be useful in the longer term. Looking at Figure 1 one can observe that the variance in the period before 2009 is clearly higher than afterwards. Before modelling this time series we pre-processed the data. Firstly, the data was cleaned. For this purpose, we used the procedure implemented in {forecast} package for the R program, where time series is decomposed and then the trend (and optionally seasonal component) is removed. For the remainder series, outliers are identified as values greater than Q3+3Q or less than Q1-3Q, and then they are replaced using linear interpolation.[5] In this way observations from (03/2000; 04/2002; 02/2007; 04/2007) appeared to be outliers and their actual values (1.1, -0.7, 1.3, 1.1) were replaced by (0.62293325, -0.04283356, 0.61761271, 0.80996611) respectively. Then the data was logarithmically transformed according to the formula: $y = \ln(1+x)$. The addition of one in the argument of the logarithm was necessary due to the requirement of a domain that must be positive. Even after these actions, the time series remained non-stationary, which was confirmed by augmented Dickey-Fuller test (p-value=0.3281). Although one might expect that there are seasonal fluctuations in the construction market, this is not confirmed by the graph in Figure 3, where indexes are shown in particular months. In addition, we checked this by performing two seasonality tests available in the R package {forecast} (Hyndman & Khandakar 2008). Both tests ruled out the significance of seasonal fluctuations. The Osborn, Chui, Smith and Birchenhall (OCSB) test yielded a test statistic value of -9.9689, while the 5% critical value is -1.803. The Canova and Hansen test, on the other hand, did not yield a basis for rejecting the null hypothesis of no unit roots at seasonal frequencies. The p-values for individual months ranged from 0.1539 to 0.9644.

[5] For more details which are not supplied in {forecast} package see https://robjhyndman.com/hyndsight/tsoutliers/.

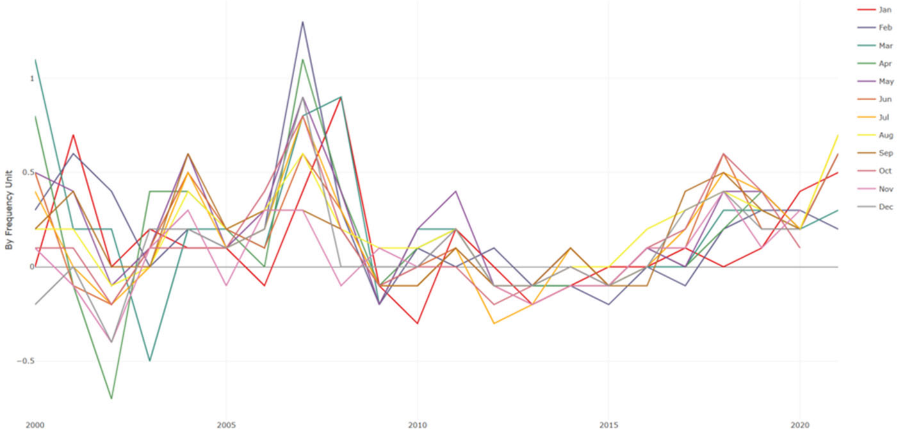


Figure 3. Construction costs index from seasonal perspective

All these preliminary studies were the motivation for choosing the local regression method to model this time series. The spline methodology automatically divides the time axis into periods (intervals) and fits the regression curves to them separately. Due to constraint (1), a continuous line is obtained over the entire time axis. The price index in the construction market will be approximated by a cubic smoothing spline model (3). In order to control the complexity of the model we use the GCV criterion (8), which geometrically results in a smooth regression line. The result is an increasing linear forecast over a 10-month horizon (Figure 4).

Assuming a significance level of 0.05, the independence of the residuals was verified. We used the Ljung-Box test (Ljung & Box 1978) with the number of lags equal to 10. The obtained p-value = 0.26 supports not rejecting the null hypothesis, according to which there is no autocorrelation between residuals. Although the p-value in the Shapiro-Wilk test for residuals is much less than 0.05, outliers appear to be the reason of this result. The isolated bars in the tails of the distribution are clearly visible on the histogram of the residuals (Figure 5). We identified them by assuming that outliers are observations meeting one of the conditions of $y_t < Q_1 - 3Q$ or $y_t > Q_3 + 3Q$, where $Q_i$are quartiles and Q is the quartile deviation. There were nine such observations from: 03/2000, 01/2001, 04/2001, 02/2002, 01/2003, 03/2003, 04/2003, 10/2004, 01/2010. It should be noted that all but one date from the very early period of the data under consideration. The most recent is from 01/2010, which is halfway through the time series. After removing these atypical residuals, we obtained p-value of 0.1467 in the Shapiro-Wilk test, thus there is no statistical evidence to reject the hypothesis of a normal distribution of residuals in this case.

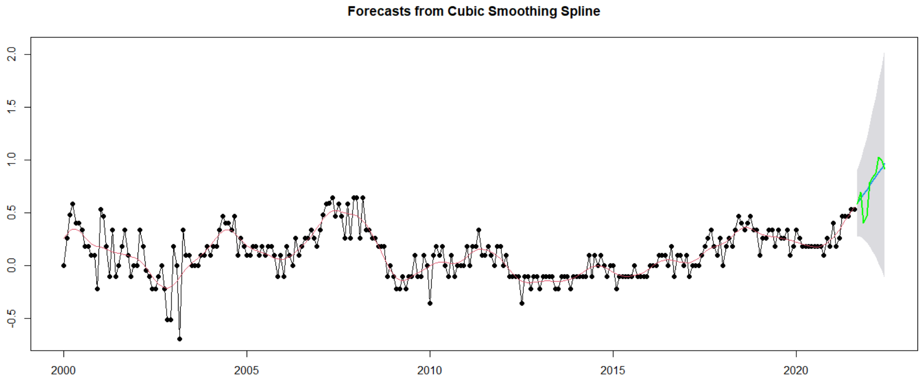


Figure 4. Fitted regression spline curve to the construction costs index (in red). Blue line depicts a 10-month ahead forecast whereas the green one shows observations from the test set

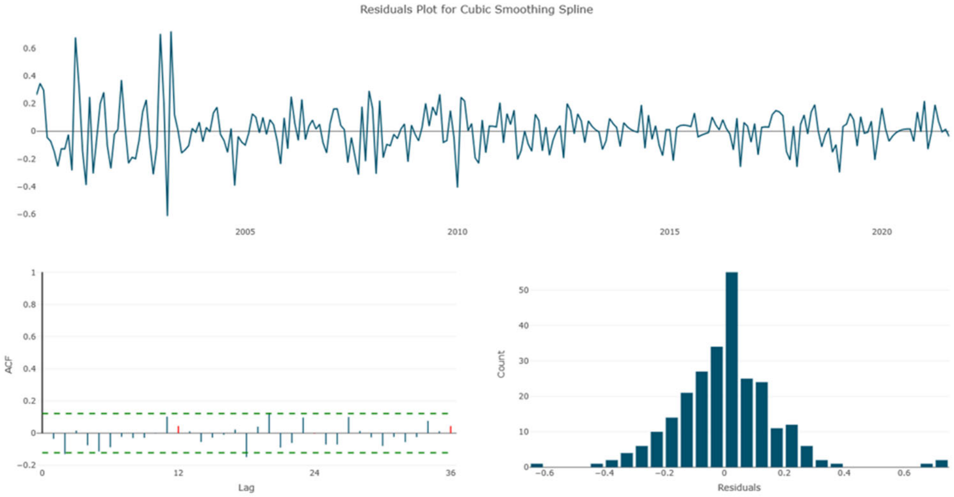


Figure 5. Residuals of the spline model – their autocorrelations and histogram

Table 1 presents the forecast in a 10-month horizon, together with confidence intervals. We obtained the test errors of the forecasts as follows: RMSE = 0.2534, MAE = 0.1943, MAPE = 23.77%. Although the errors are slightly high, note that the biggest differences between the actual values and the forecast are observed in the third and fourth steps of the forecast, which are November 2021 and December 2021 (Figure 6). They mainly influenced the overall values of the considered error measures. At that time, the Polish government introduced a second lockdown due to the COVID-19 pandemic, which could have caused disturbances in the economy, and in particular in the construction market. Note that the cubic smoothing spline model applied to the original data, which was neither cleaned nor transformed, turned out to be definitely inferior: RMSE = 0.3510, MAE = 0.3083, MAPE = 29.38%.

Table 1. The forecast 10 month ahead.

Month	Test observations	Point forecast	Lower limit .95	Upper limit .95
Sep 2021	0.8	0.8104521	0.31633653	1.490045
Oct 2021	1.0	0.8870110	0.31061334	1.716904
Nov 2021	0.5	0.9668073	0.28780969	2.003806
Dec 2021	0.6	1.0499780	0.25092565	2.359440
Jan 2022	1.2	1.1366658	0.20315095	2.794487
Feb 2022	1.3	1.2270194	0.14734652	3.322683
Mar 2022	1.4	1.3211937	0.08593149	3.961584
Apr 2022	1.8	1.4193505	0.02090157	4.733419
May 2022	1.7	1.5216579	-0.04610327	5.666087
Jun 2022	1.5	1.6282917	-0.11374306	6.794486

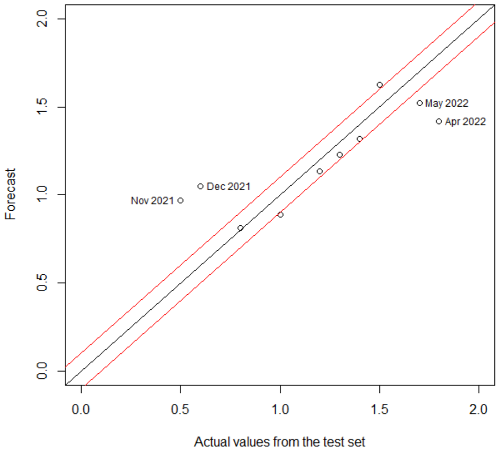


Figure 6. The differences between the actual values of the construction costs index in the test set and the spline model forecast. The red lines show +/- 0.1 outstanding from the exact forecast

One of the most popular method of time series analysis and forecasting is ARIMA, which stands for Auto-Regressive Integrated Moving Average (Box et al. 2008). Thus, it is justified to compare our results with this econometric approach. Instead of analyzing autocorrelation plots to determine the number of lags, degree of differencing, and order of the moving average model, we used the auto.arima function in R (Hyndman & Khandakar 2008). It enables searching of the parameter space guided by information criterion. Since the considered time series is of moderate size, we used an exhaustive search. The final model took the form ARIMA(4,0,0):

$$\begin{array}{l} \hat{Y} = \underset{(0.0632)}{0.1364} + \underset{(0.0625)}{0.3593} \cdot Y_{t-1} + \underset{(0.0641)}{0.1645} \cdot Y_{t-2} + \underset{(0.0641)}{0.2595} \cdot Y_{t-3} + \underset{(0.0634)}{0.0759} \cdot Y_{t-4} \end{array} \tag{9}$$

and forecasted decreasing trend 10 month ahead (Figure 7). The standard errors of the coefficients are given in brackets below the model (9). Note that the model is expressed for transformed index of costs (in fact it has a multiplicative form for original data). The errors of the forecast are dramatically high: RMSE= 0.8632, MAE= 0.7216, MAPE= 52.38. Since this model has no predictive abilities we omit the analysis of the residuals.

Then we compared our model with non-linear variant of ARIMA, where autocorrelation function in AR(p) part of the model was estimated with a use of feed-forward neural network with a single hidden layer (Lippi & Zaniolo 2012). As previously, the parameters of the model were exhaustively examined to minimize the information criterion. The model took the form NNet-ARIMA(3,1,2) and also predicted a downward trend in the next 10 months (Figure 8). Similar to classical ARIMA, all error measures considered are high: RMSE= 0.8579, MAE= 0.7138, MAPE= 51.57.

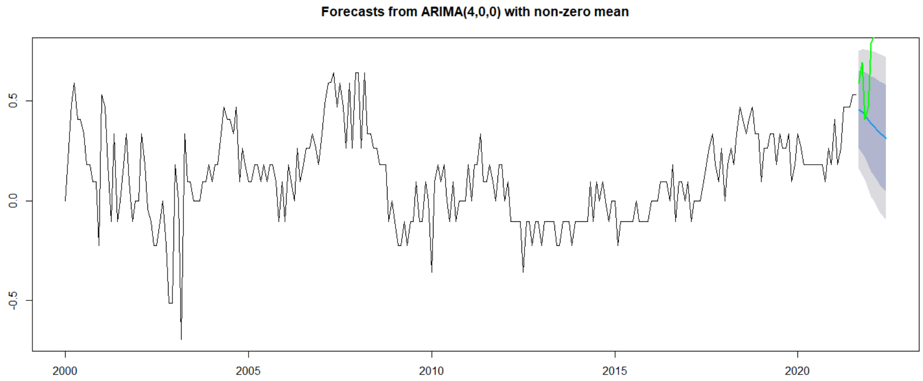


Figure 7. Decreasing forecast (blue line) pointed out by ARIMA model. Actual observations from the test set are in green

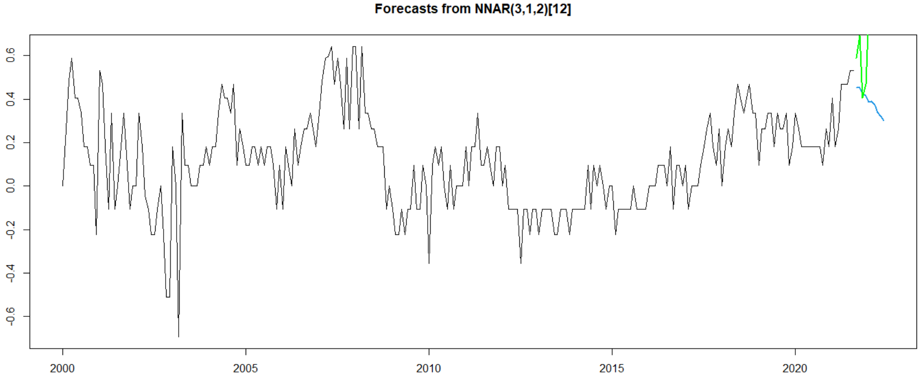


Figure 8. Decreasing forecast (blue line) pointed out by neural network version of ARIMA. Actual observations from the test set are in green

6. Conclusions

The construction costs index is an important component in the macroeconomic system. In the economy, it belongs to at least two of the three factors of production, namely capital and land. The value and dynamics of changes in construction production have a significant impact on the condition of the economic system, and in particular on the condition of the real estate market. The importance of built production in the economic system makes knowledge of its future values a key decision-making resource in modern management. Taking this into account, the authors of this research tested the usefulness of the nonlinear spline method for forecasting changes in the construction costs index in the long term.

Our work showed that nonparametric local regression is capable of being a valuable forecasting tool. In the case of the considered data, the commonly used ARIMA econometric model as well as its extension based on a neural network failed completely. Both ARIMA approaches indicated a decreasing forecasting when the test set data had in fact an increasing trend. The error measures on the test set for the spline model turned out to be not entirely satisfactory, but let us emphasize again that it could have been influenced by the lockdown decisions in Poland during the second wave of the pandemic. These error measures were overstated by the two-month observations. Nevertheless, the spline model detected a sharp increase in the costs index in the coming months, which provides valuable information in making investment decisions. It is noteworthy that smoothing splines can be recommended also due to a lack of tuning parameters, which is convenient for practitioners who are not advanced in data analysis.

References

Al Kailani, H., Sweis, G. J., Sammour, F., Maaitah, W. O., Sweis, R. J. and Alkailani, M., (2024). Predicting construction cost index using fuzzy logic and machine learning in Jordan. *Construction Innovation*. https://doi.org/10.1108/CI-08-2023-0182.

Altalhoni, A., Liu, H., Abudayyeh, O., (2024). Forecasting Construction Cost Indices: Methods, Trends, and Influential Factors. *Buildings*, 14(10), art. no. 3272.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C., (2008). Time Series Analysis: Forecasting and Control (4th ed.). *Wiley*.

Butler, S., Kokoszka, P., Miao, H. and Shang, H. L., (2021). Neural network prediction of crude oil futures using B-splines. *Energy Economics*, 94. https://doi.org/10.1016/j.eneco.2020.105080.

Cao, Y., Ashuri, B., (2020). Predicting the Volatility of Highway Construction Cost Index Using Long Short-Term Memory. *Journal of Management in Engineering*, 36(4). https://doi.org/10.1061/(ASCE)ME.1943-5479.0000784.

Cheung, S. O., Wong, P. S. P., Fung, A. S. Y. and Coffey, W. V., (2006). Predicting project performance through neural networks. *International Journal of Project Management*, 24(3), pp. 207–215. https://doi.org/10.1016/j.ijproman.2005.08.001.

Chu, M. T., Kuo, Ch and Lin, M. M., (2013). Tensor Spline Approximation in Economic Dynamics with Uncertainties. *Computational Economics,* 42, pp. 175–198. https://doi.org/10.1007/s10614-012-9331-1.

Dash, S. R., Maitra, D., (2019). The relationship between emerging and developed market sentiment: A wavelet-based time-frequency analysis. *Journal of Behavioral and Experimental Finance*, 22, pp. 135–150. https://doi.org/10.1016/j.jbef.2019.02.006.

Durdyey, S., Ismail, S., (2012). Role of the construction industry in economic development of Turkmenistan. *Energy Science and Research*, Vol. 29, No. 2, pp. 883–890.

ELFAHHAM, Y., (2019). Estimation and prediction of construction cost index using neural networks, time series, and regression. *Alexandria Engineering Journal*, 58(2), pp. 499–506.

Findley, D. F., Lytras, D. P. and Maravall, A., (2016). Illuminating ARIMA model-based seasonal adjustment with three fundamental seasonal models. *SERIEs*, 7, pp. 11–52. doi:10.1007/s13209-016-0139-4.

Friedman, J. H., (2009). Multivariate adaptive regression splines. *The Annals of Statistics*, Vol. 19, No. 1, pp. 1–141.

Gadasina, L., Vyunenko, L., (2022). Applying spline-based phase analysis to macroeconomic dynamics. *Dependence Modelling*, 10(1), pp. 207–214. https://doi.org/10.1515/demo-2022-0113.

Hastie, T., Tibshirani, R., Friedman, J., (2009). The Elements of Statistical Learning: Data Mining. *Inference and Prediction*, *Springer*, New York.

Huang, C., Raunikar, R., (1981). Spline Functions: An Alternative to Estimating Income-Expenditure Relationships for Beef. *Journal of Agricultural and Applied Economics*, 13(1), pp. 105–110. doi:10.1017/S0081305200024626.

Humphrey, D. B., Vale, B., (2004). Scale economies, bank mergers, and electronic payments: A spline function approach. *Journal of Banking and Finance*, 28(7), pp. 1671–1696. https://doi.org/10.1016/j.jbankfin.2003.05.003.

Hyndman, R. J., Khandakar, Y., (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 26(3).

Ilyasov, R. H., Yakovenko, V. S., (2021). Spline Analysis of Flow Correlation in Economic Systems, in The Challenge of Sustainability in Agricultural Systems, BOGOVIZ, A. V., eds., *Springer International Publishing, Springer*.

Joukar, A., Nahmens, I., (2016). Volatility forecast of construction cost index using general autoregressive conditional heteroskedastic method. *Journal of Construction Engineering and Management*, 142(1), art. no. 04015051.

Kim, S., Choi, C.-Y., Shahandashti, M. and Ryu, K. R., (2022). Improving Accuracy in Predicting City-Level Construction Cost Indices by Combining Linear ARIMA and Nonlinear ANNs. *Journal of Management in Engineering*, 38(2), art. no. 04021093.

Lin, R. F. Y., Ou, C., Tseng, K. K., Bowen, D., Yung, K. L. and Ip, W. H., (2021). The spatial neural network model with disruptive technology for property appraisal in real estate industry. *Technological Forecasting and Social Change*, 173. https://doi.org/10.1016/j.techfore.2021.121067.

Lippi, M., Zaniolo, F., (2012). A feedforward neural network model for time series forecasting with applications in finance and economics. *Artificial Intelligence in Finance and Economics*, pp. 1–20.

Ljung, G. M., Box, G. E., (1978). On a measure of lack of fit in time series models. *Biometrika*, Vol. 65, No. 2, pp. 297–303. https://doi.org/10.1093/biomet/65.2.297.

Lopes, J., (2011). Construction in the economy and its role in socio-economic development, in New Perspectives on Construction in Developing Countries, OFORI, G., eds., pp. 40–71, London, *Routledge*.

Mao, S., Xiao, F., (2019). A novel method for forecasting Construction Cost Index based on complex network. *Physica A: Statistical Mechanics and its Applications*, 527, art. no. 121306.

Monteiro, A. M., Tütüncü, R. H. and Vicente, L. N., (2008). Recovering risk-neutral probability density functions from options prices using cubic splines and ensuring nonnegativity. *European Journal of Operational Research*, 187(2), pp. 525–542. https://doi.org/10.1016/j.ejor.2007.02.041.

Moon, S., Chi, S. and Kim, D. Y., (2018). Predicting Construction Cost Index Using the Autoregressive Fractionally Integrated Moving Average Model. *Journal of Management in Engineerin*g, 34(2), art. no. 04017063.

Moon, T., Shin, D. H., (2018). Forecasting Model of Construction Cost Index Based on VECM with Search Query. *KSCE Journal of Civil Engineering*, 22(8), pp. 2726-2734.

Naccache, T., (2011). Oil price cycles and wavelets. *Energy Economics*, 33(2), pp. 338–352. https://doi.org/10.1016/j.eneco.2010.12.001.

Nieto, M. R., Carmona-Benítez, R. B., (2018). ARIMA + GARCH + Bootstrap forecasting method applied to the airline industry. *Journal of Air Transport Management*, 71(June), pp. 1–8. https://doi.org/10.1016/j.jairtraman.2018.05.007.

Osiewalski, K., Osiewalski, J., (2013). A Long-Run Relationship between Daily Prices on Two Markets: The Bayesian VAR(2)–MSF-SBEKK Model. *Central European Journal of Economic Modelling and Econometrics*, 5(1), 65–83. https://doi.org/10.24425/cejeme.2013.119253.

Paci, L., Consonni, G., (2020). Structural learning of contemporaneous dependencies in graphical VAR models. *Computational Statistics & Data Analysis*, 144, 106880. https://doi.org/10.1016/j.csda.2019.106880.

Pheng, L. S., Hou, L. S., (2019). The Economy and the Construction Industry, In Construction Quality and the Economy. Management in the Built Environment, PHENG, L.S., HOU, L.S., eds., pp. 21–54, Singapore, *Springer*.

Ruddock, L., Ruddock, S., (2008). The scope of the construction sector: Determining its value, in Economics for the Modern Built Environment. RUDDOCK, L., eds., pp. 99–113, London, *Routledge*.

Sahraei, M.A., Duman, H., Çodur, M. Y. and Eyduran, E., (2021). Prediction of transportation energy demand: Multivariate Adaptive Regression Splines, Energy. *Elsevier Ltd*, Vol. 224. doi:10.1016/j.energy.2021.120090.

Sobieraj, J., Metelski, D. and Nowak, P., (2021). The view of construction companies' managers on the impact of economic, environmental and legal policies on investment process management. *Archives of Civil Engineering*, Vol. LXVII No. 1, pp. 111–129.

Vidoli, F., (2011). Evaluating the water sector in Italy through a two-stage method using the conditional robust nonparametric frontier and multivariate adaptive regression splines. *European Journal of Operational Research*, Vol. 212, No. 3, pp. 583–595. doi: https://doi.org/10.1016/j.ejor.2011.02.003.

Wang, C.-H., Mei, Y.-H., (1998). Model for forecasting construction cost indices in Taiwan. *Construction Management and Economics*, 16(2), pp. 147–157.

Wang, J., Ashuri, B., (2017). Predicting ENR Construction Cost Index Using Machine-Learning Algorithms. *International Journal of Construction Education and Research*, 13(1), pp. 47–63.

Williams, T. P., (1994). Predicting changes in construction cost indexes using neural networks. *Journal of Construction Engineering and Management*, 120(2), pp. 306–320.

Xu, J.-W., Moon, S., (2013). Stochastic forecast of construction cost index using a cointegrated vector autoregression model. *Journal of Management in Engineering*, 29(1), pp. 10–18.

Ye, K., Lu, W., Jiang, W., (2009). Concentration in the international construction market. *Construction Management and Economics*, Vol. 27, No. 12, pp. 1197–1207.

Zhang, R., Ashuri, B., Shyr, Y., Deng, Y., (2018). Forecasting Construction Cost Index based on visibility graph: a network approach. *Physica A: Statistical Mechanics and its Applications*, 493, pp. 239-252.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 89–110, https://doi.org/10.59139/stattrans-2025-041
Received – 23.11.2024; accepted – 18.05.2025

The truncated Schröter recursive algorithm for the computation of aggregate claim amounts

Friday I. Agu[1]

Abstract

This study introduces and evaluates the truncated Schröter recursive algorithm for computing aggregate claim amounts in the insurance sector. The algorithm addresses the limitations in the existing methods by incorporating truncation at 1, which is crucial for an accurate modelling of insurance claims where the events leading to a claim are pivotal. Using the AutoCollision dataset, the study compares the truncated Schröter algorithm with the Panjer and Schröter recursion algorithms, focusing on computational efficiency and accuracy. Furthermore, the descriptive statistics revealed substantial variability and risk factors, such as higher claim severity for business-use vehicles and young drivers aged 17–20. The results demonstrate that the truncated Schröter algorithm substantially reduces the execution time while maintaining high accuracy, thus making it a superior tool for risk management and premium setting.

Key words: insurance claim amounts, aggregate claim distribution, recursive algorithm, insurance risk management, computational efficiency.

1. Introduction

In the insurance domain, company profits depend largely on the premiums collected from policyholders and the claim amounts paid to insured individuals. Unlike in other market sectors, such as manufacturing, determining the appropriate premium for an insurance portfolio is particularly challenging. This complexity arises from the need to account for future uncertainties and ensure sustained and adequate investment. To address this, insurance companies employ models designed to accurately compute aggregate claim amounts within a collective risk framework and estimate the probability that total claims will not exceed a specified threshold. The process begins with an estimation of expected costs to establish a baseline premium. This is then adjusted by adding margins that account for uncertainties, provide a profit buffer, and

[1] Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia & Department of Sensory Information Systems and Technologies, Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia.
E-mail: agu1@uniba.sk. ORCID: https://orcid.org/0000-0002-2367-4732.

reflect potential aggregate claims payable to policyholders (Yartey, 2020). Central to this approach is the distribution of aggregate claim amounts, which is derived from the convolution of claim frequencies and severities. This distribution plays a crucial role in pricing insurance portfolios as it informs the likelihood and magnitude of potential losses. Accurate estimation of aggregate claim amounts is therefore critical for insurance companies as it supports informed decisions about pricing competitiveness, risk margins, and capital allocation. However, a persistent challenge in actuarial mathematics lies in modeling this distribution when discrete, non-negative integer values represent the number of claims and the severity of claims. Accurately capturing this behavior is essential for reliable risk assessment and premium setting.

1.2. Literature review

Historically, before the advent of modern computing, actuaries relied primarily on estimation and approximation techniques that lacked a rigorous theoretical foundation for determining aggregate claim amounts. These methods were limited in accuracy and reliability, making data-driven decision-making in insurance challenging. A widely adopted approach for analyzing the distribution of aggregate claim amounts involves identifying suitable counting distributions defined over the non-negative integers and fitting them separately to the number of claims and claim severities. However, while claim frequencies are inherently discrete, claim severities are typically modeled as continuous random variables and are thus best represented by continuous distributions. Numerous studies, such as those by Hogg and Stuart (2009), Gray and Pitts (2012), Packová and Brebera (2015), Pacáková and Gogola (2013), Jindrová and Pacáková (2016), and Dzidzornu and Minkah (2021), have examined various methods for fitting distributions to insurance claim datasets. Despite their widespread use, these approaches can be unreliable as they often fail to accurately capture the convolution between the number of claims and claim severity, two central components of the aggregate claim distribution. This convolution forms the basis of the aggregate claims model and has been applied extensively in actuarial science to solve various insurance-related problems (Albrecher et al., 2017; Klugman et al., 2019; Mildenhall & Major, 2022). However, computing this convolution presents substantial challenges, primarily due to the absence of a closed-form expression and the associated computational complexity.

To address these issues, alternative computational strategies have been developed, such as the normal power approximation and fast Fourier transform (FFT) techniques (Beard et al., 1977; Cooley & Tukey, 1965; Heckman & Meyers, 1983; Mildenhall, 2024). Although these methods enhance theoretical understanding, they often become computationally intensive and less accurate when applied to large datasets with high claim frequencies and severities. These limitations have motivated the search for more efficient and robust approaches. One such approach is the recursive method, often referred to as the "exact method". Unlike convolution-based techniques, the recursive approach assumes that the number of claims and claim severity distributions are discrete, enabling the computation of aggregate claim amounts through recursive formulas. This method substantially reduces computational burden while maintaining accuracy, particularly in scenarios involving a large number of claims. A foundational contribution in this area was made by Panjer (1981), who introduced the Panjer recursive family of discrete distributions and the corresponding recursion formula for computing aggregate claim amounts. The Panjer recursive formula has spurred extensive research in actuarial science, with notable contributions from Sundt and Vernic (2009), Yartey (2020), Dickson (2016), and Ghinawan et al. (2021). More recently, Tzaninis and Bozikas (2024) extended the Panjer family of claim number distributions by treating the family's parameters as random variables, thereby deriving a more flexible compound distribution. Their formulation assumes that claim sizes are conditionally independent and identically distributed, as well as conditionally independent of the number of claims. In a related development, Fackler (2023) introduced a reparameterization of the Panjer family, enhancing its modeling flexibility.

Although the Panjer recursion effectively models aggregate claim amounts, its applicability is confined to a narrow class of counting distributions that have a fixed, positive probability at zero. To address this constraint, Schröter (1990) proposed the Schröter recursive formula, which accommodates a broader range of counting distributions and more accurately captures the dynamics of aggregate claims. However, this method relies on convolution operations, making it computationally demanding, especially when dealing with high claim frequencies and large claim amounts. Recent advances in computational modeling have substantially broadened the methodologies available for estimating aggregate claim amounts, supplementing, and in some cases outperforming, traditional actuarial approaches. For instance, Qiu (2019) compared classical reserving methods, such as the Chain Ladder and Bornhuetter-Ferguson techniques, with machine learning-based individual claims reserving. The study found that models like generalized linear models, artificial neural networks, random forests, and support vector machines delivered superior performance on simulated datasets rich in claim-level features. However, these advantages diminished when applied to smaller, real-world datasets. Likewise, Hofmann (2022) proposed fast Fourier transform (FFT)-based algorithms as a computationally efficient alternative to the Panjer recursion under arbitrary claim frequency distributions, incorporating exponential tilting to reduce wrap-around effects and better capture distribution tails. Additionally, Gamaleldin et al. (2025) introduced a hybrid CNN-LSTM model that captures both spatial and temporal patterns in insurance claims data, considerably improving volatility forecasting and enabling proactive risk management. While these studies underscore the growing influence of machine learning in enhancing the precision, scalability, and adaptability of aggregate claims modeling, they also highlight a key trade-off: improved predictive performance often comes at the cost of increased computational complexity and resource demands during implementation and model tuning.

The computation of aggregate claim amounts plays an increasingly pivotal role in risk management and the pricing of insurance coverage. Insurance companies are inherently motivated to minimize claim payouts while maximizing premium income, thereby strengthening their ability to manage future uncertainties and withstand catastrophic losses. Within this highly competitive landscape, insurers face the added challenge of dealing with the unpredictable nature of claim occurrences embedded in insurance contracts.

Despite the utility of the Schröter recursive formula, it does not fully capture the dynamics of claim amounts truncated at one. This practice holds significant practical relevance in real-world insurance settings. In many cases, insurers are primarily concerned with the number of events that generate claims, rather than the exact amounts. Once a claim is reported, the minimum observed claim amount is often truncated at one, effectively implying a zero probability for a claim amount of zero. This reflects typical policy structures that include deductibles, where insured individuals are responsible for losses below a certain threshold, and only the excess is reimbursed. Consequently, minor losses below the deductible are frequently unreported, making one the effective lower bound for observed claim amounts. This truncation has a substantial impact on the modeling of risk exposure, influencing both the accuracy of risk assessment and the determination of premium rates. In risk theory, truncated distributions are essential for modeling claim severities and inter-arrival times, providing insurers and actuaries with critical tools to better understand the frequency and magnitude of losses. As such, accurately modeling the number of claims truncated at one is vital for capturing the true nature of insurance liabilities. It requires careful consideration of the underlying distributions that govern both claim frequency and severity, ultimately supporting more precise pricing and effective risk management. To address this gap, the present study introduces and explores the truncated Schröter recursive formula—a mathematical framework designed to improve accuracy in the computation of aggregate claim amounts. The study further assesses the computational efficiency of the proposed algorithm by analyzing its runtime performance, offering insights into its practical applicability for large-scale insurance datasets.

2. The recursive formulas

2.1. The Panjer recursive formula

The Panjer (1981) recursive formula is defined as

$$P_k = \left(a + \frac{b}{k}\right) P_{k-1}, \; k = 1,2,3,\ldots \tag{1}$$

where a and $b$are parameters, P_k denotes the recurrent probability, $P_{k-1}$is the backward recurrent probability, and by definition, $P_k = 0$ for $k < 0$. The counting distributions that satisfied (1) were explored in Panjer (1981). Furthermore, Panjer (1981) obtained the corresponding recursion algorithm for (1) defined as:

$$g(s) = \frac{1}{1-af_0}\sum_{i=1}^{s}\left(a+\frac{bi}{s}\right)f_i g(s-i), \tag{2}$$

and by definition, $f_0 = P(X = 0) = 0$and$g(0) = p_0$, where$p_0$denotes the probability mass function of the counting distribution evaluated at zero, that is, the initial probability. For instance, if$p_n$is the Poisson distribution function from the recursive family defined in (1), then $p_n$evaluated at zero (p_0) and one (p_1) represents the initial probabilities of no claim and the probability of a claim, respectively.

2.2. The Schröter recursive formula

While the Panjer recursive formula addresses the challenges of the traditional convolution method, it is limited to a few distributions. Hence, Schröter (1990) generalized (1) and obtained the recursive formula expressed as:

$$P_k = \left(a+\frac{b}{k}\right)P_{k-1} + \frac{c}{k}P_{k-2},\ k = 1,2,3,\ldots, \tag{3}$$

where a, b, and c are parameters, P_{k-1} and P_{k-2} are recursive backward probabilities, and $P_k = 0$ for $k < 0$ (by definition). Note that for $c = 0$, (1) becomes a particular case of (3). Additionally, the counting distributions defined by (3) also contain the convolutions of the Poisson distribution and another distribution from (1) (see Schröter, 1990). Furthermore, Schröter (1990) obtained the corresponding recursion algorithm for (3) defined as:

$$g(s) = \frac{1}{1-af_0}\sum_{i=1}^{s}\left[\left(a+\frac{bi}{s}\right)f_i + \frac{ci}{2s}f_i^{2*}\right]g(s-i), \tag{4}$$

where f_i^{2*}has to be evaluated by the convolution formula $f_i^{2*} = \sum_{j=0}^{i} f_j f_{i-j}$and for $c = 0$, (4) becomes (2). The parameter estimation of (3) has been studied in Agu, Mačutek, and Szűcs (2023).

3. The truncated Schröter recursive formula

In this section, we present the truncated Schröter recursive formula. We defined the truncated Schröter recursive formula as:

$$P_k = \left(a+\frac{b}{k}\right)P_{k-1} + \frac{c}{k}P_{k-2},\ k = 2,3,4,\ldots, \tag{5}$$

where the parameters are defined as in (3) and note that (5) is truncated at 1.

First, let Kbe a discrete random variable taking non-negative integer values as defined in (5) and using the fact that the probability generating function is defined as:

$$G(s) = \sum_{k=0}^{\infty} s^k P_k,$$

where $s \in [0,1]$such that $G(s) \geq 0$and $P_k$is the recursive probability defined in (5) and $\sum P_k = 1$. Thus, the probability generating function corresponding to (5) is:

$$G(s) = e^{-\frac{c(s-1)}{a}} \left(\frac{1-a}{1-as}\right)^{\frac{a(a+b)+c}{a^2}}, \tag{6}$$

for $|as| \neq 1$.

The derived truncated probability mass function corresponding to (5) is given as:

$$q_n = \frac{e^{\frac{c}{a}}(1-a)^{\frac{a(a+b)+c}{a^2}} \sum_{i=0}^{n} \binom{\frac{a(a+b)+c}{a^2}+i-1}{i} \frac{\left(-\frac{c}{a}\right)^{n-i} a^i}{(n-i)!}}{1-e^{\frac{c}{a}}(1-a)^{\frac{a(a+b)+c}{a^2}}}, \; n = 1,2,\ldots, \; 0 < a < 1, \; b,c \in \mathbb{R}. \tag{7}$$

Let $r = \frac{a(a+b)+c}{a^2}$, $x = \frac{c}{a}$, and define the generating function for the negative binomial coefficient as:

$$\sum_{k=0}^{\infty} \binom{r+k-1}{k} z^k = (1-z)^{-r}, \; |z| < 1.$$

The goal is to express the finite sum in a form that leverages the generating function.

To relate $\sum_{i=0}^{n} \binom{r+i-1}{i} \frac{(-x)^{n-i} a^i}{(n-i)!}$ to the generating function for the negative binomial coefficient above, we differentiate $(1-z)^{-r}$with respect to $z$evaluated at $z = x - a$ $(0 < a < 1)$ to obtain terms that match the structure of our sum. We have that

$$\sum_{i=0}^{n} \binom{r+i-1}{i} \frac{(-x)^{n-i} a^i}{(n-i)!} = \frac{\Gamma(r+n)}{n!\,\Gamma(r)} \left(\frac{a}{a-c+a^2}\right)^{(r+n)}.$$

Hence, (7) can be expressed as

$$q_n = \frac{e^{\frac{c}{a}}(1-a)^r \frac{\Gamma(r+n)}{n!\Gamma(r)} \left(\frac{a}{a-c+a^2}\right)^{(r+n)}}{1 - e^{\frac{c}{a}}(1-a)^r}, \; n = 1,2,\ldots, \; 0 < a < 1, \; b \geq 0, c \in \mathbb{R}.$$

Also, the log-likelihood function corresponding to (7) can be simplified as:

$$\ell(a,b,c|n_1,\ldots,n_k)$$
$$= k\,log\left[e^{\frac{c}{a}}(1-a)^r\right] + \sum_{j=1}^{k} log\left[\frac{\Gamma(r+n)}{n!\,\Gamma(r)}\left(\frac{a}{a-c+a^2}\right)^{(r+n)}\right]$$
$$- k\,log\left[1 - e^{\frac{c}{a}}(1-a)^r\right].$$

3.1. The truncated Schröter algorithm

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed claim severities over the non-negative integers with probability density $f_k = p(X_i = k)$ for $i = 1,2,\ldots,n, k = 0,1,2,\ldots$, and $f^{k*} = P(X_1 + X_2 + \ldots + X_n = k)$ denotes the n-fold convolution of f_k. Additionally, let Nbe a discrete random variable representing the number of claims with a discrete probability mass function defined as $p_n = P(N = n)$, such that$X_i$are stochastically independent of N, and $S = \sum_{i=1}^{N} X_i$ is the aggregate claim. While the truncated Schröter algorithm is derived under the classical assumption that claim frequency and severity are stochastically independent, it is important to note that this assumption may not fully reflect the complexities of real-world insurance portfolios. In practice, claim frequency and severity may be influenced by common risk factors (e.g. policyholder behavior, geographic or economic conditions), potentially inducing dependence between them. Ignoring this dependence can lead to biased estimates of aggregate risk, particularly in portfolios characterized by frequent and large claims, although the independence assumption facilitates analytical derivation and computational feasibility. For all the severity distributions f^{k*}, we derived the recursive algorithm as:

$$g(s) = \sum_{k=2}^{\infty} P_k f^{k*}(s), \tag{8}$$

where P_k is defined in (5).

The Panjer recursion formula defined in (2) is based on the expression $f^{k*}(s) = \frac{k}{s}\sum_{i=1}^{s} i f_i f_{s-i}^{k-1}$, $s = k = 1,2,3...$, (see Schröter,1990; page 164). We can write this as: $f(s) = \frac{1}{s}\sum_{i=1}^{s} i f_i$.

Thus,

$$g(s) = \sum_{k=2}^{\infty} \left[\left(a + \frac{b}{k}\right) P_{k-1} + \frac{c}{k} P_{k-2}\right] f^{k*}(s). \tag{9}$$

We have that

$$g(s) = a \sum_{k=0}^{\infty} P_k \left(\sum_{i=0}^{s} f_i f_{s-i}^{k-1}\right) + \sum_{k=0}^{\infty} P_k \left(\sum_{i=1}^{s} \frac{bi}{s} f_i f_{s-i}^{k-1}\right) + \gamma,$$

where $\gamma = \sum_{k=0}^{\infty} P_k \left(\sum_{i=1}^{s} \frac{ci}{s} f_i f_{s-i}^{k-1}\right)$.

Note that $\sum_{k=0}^{\infty} P_k = 1$.

Therefore, it follows that

$$g(s) = \frac{1}{1-af_0} \sum_{i=1}^{s} \left[a + \frac{i}{s}(b+c)\right] f_i g(s-i), \tag{10}$$

for $s \neq 0$and a, b, and c are the parameters. Additionally, $f_0 = P(S = 0) = 0$ and $g(0) = p_0$ is the initial probability. If $c = 0$ in (10), we obtain (2), and if we define $f^{k*}(s)$ as $f^{k*}(s) = \frac{k}{ts}\sum_{i=1}^{s} i f_i^{t*} f_{s-i}^{(k-t)*}$, $i = 1,2,\dots,$ for $t \in \{1,2,\dots,k\}$ in (9), (4) becomes a special case of (10). To execute (10), we treat f_i as the claim frequencies per number of policies.

To ensure numerical stability and convergence of (10), the parameters a, b, and $c$were estimated via maximum likelihood of (7) using the *nlminb()* optimizer with box constraints: $0 < a < 1$, $b \geq 0$ and $c \in \mathbb{R}$. These constraints prevent instability in the recursion weights and guarantee the validity of the logarithmic expressions in the likelihood function.

Theoretically, unlike (4), the recursion algorithm defined in (10) eliminates the need for any form of convolution.

This study considers the Negative binomial distribution as the count distribution for the number of claims (see Section 4, Table 4).

The probability mass function for the Negative binomial distribution is defined as

$$h(S=s)=\binom{s+r-1}{s}(1-p)^s p^r,\ s=0,1,2\ldots,\ r>0, p\in[0,1].$$

We have that

$$h(S=0)=\binom{r-1}{0}p^r.$$

From (10), we define

$$g(0)=p_0=h(S=0)=p^r. \tag{11}$$

3.2. The numerical implementation procedure of the truncated Schröter algorithm

The implementation of the truncated Schröter recursive algorithm involves several computational stages designed to estimate the distribution of aggregate claim amounts.

The procedure can be summarized as follows:

i. Data Preparation: Obtain and clean claim count data (from real-world and simulations). Compute the empirical frequency distribution $f_i$and normalize it to ensure $\sum f_i=1$. Furthermore, the distribution of the data is determined (see Table 4). The parameters in equation (10) are estimated using the maximum likelihood estimation (MLE) method based on the observed data. The log-likelihood function is constructed from the truncated probability mass function of the claim counts. Parameter estimation is carried out using the nlminb() optimizer in R, which is well-suited for bounded, nonlinear optimization problems. This approach ensures numerical stability and facilitates the explicit enforcement of parameter constraints that are critical to the recursive structure of the model. A similar approach is applied by truncating the corresponding probability mass functions of the Panjer and Schröter families (see Panjer, 1981; Schröter, 1990).

ii. Initialization: Determine $g(0)$ as in (11) and initialize a numeric vector to store $g(s)$ for $s=1,2,\ldots,n$.

iii. Recursive computation: For each $s=2,\ldots,n$, compute $g(s)$ using (10).

iv. Performance evaluation: Evaluate the sum of $g(s)$ values and record execution time per iteration to assess computational efficiency.

v. Visualization: Utilize graphical tools (e.g. bar plots, execution time plots) to display the algorithm's output and benchmark it against the Panjer and standard Schröter methods.

4. Numerical evaluation

In this section, we examine the run-time computational efficiency of the introduced truncated Schröter algorithm using the Automobile UK Collision Claims (AutoCollision) data obtained from https://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html.

First, we began by exploring the descriptive statistics of the dataset, analyzing average claim severity and average claim counts across various age groups and vehicle categories to identify patterns and determine how frequently each group files claims. Particular attention was given to combinations of age groups and vehicle use categories associated with high claim severity and frequency, as these represent higher risk factors for insurers and may necessitate adjustments in insurance coverage strategies.

To model the claim count data, we fitted both the Negative Binomial and Generalized Poisson distributions, selected for their ability to handle overdispersion commonly observed in count data. The choice between these distributions was guided by model fit, using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to select the model with the lowest values. Furthermore, to implement the truncated algorithm defined in (10), we employed the truncated probability mass function introduced in (7) to obtain numerical estimates of the parameters $\hat{a}$, $\hat{b}$, and $\hat{c}$as $\hat{a} = 0.99070, \hat{b} = 1.29297$, and $\hat{c} = 0.29330$ using the maximum likelihood estimation method. The computation of aggregate claim amounts using (2), (4), and (10) was performed as defined in Section 4.1.

In this study, all statistical analyses and computations of recursion algorithm run times were performed using RStudio on a Lenovo PC equipped with an 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz processor and 8.00 GB of RAM.

Table 1. Descriptive Statistics

Min.	Max.	Mean	Variance	Kurtosis	Skewness
5.00	970.00	279.44	58374.38	4.08	1.25

The descriptive statistics offer a comprehensive summary of the dataset's distribution and central tendency. The minimum and maximum values define the data range, while the mean provides a central value around which the data are distributed. The high variance indicates substantial variability (overdispersion), and the positive skewness and kurtosis indicate a right-skewed distribution with the presence of outliers.

Table 2. Analysis of Average Claim Severity by Vehicle Use

Vehicle Use	Average Claim Severity	Claim Count
Business	395.21	1075
DriveLong	265.26	2710
DriveShort	231.74	3888
Pleasure	213.20	1269

Table 2 shows that vehicles used for business purposes exhibit the highest average claim severity. Although the claim count in this category is relatively low compared to others, each claim carries a substantial financial impact, indicating that business use presents a higher risk of costly claims.

Vehicles used for long-distance driving exhibit a moderate average claim severity, which is substantially lower than that of business use but higher than for short drives and pleasure use. The relatively high claim count indicates that long drives are associated with frequent incidents, though each claim tends to be less severe than those in the business category.

Short drives register the highest claim count but a lower average claim severity. This indicates that while short trips result in more frequent claims, the financial impact of each is comparatively minor. The high frequency highlights a notable number of incidents with less severe consequences per occurrence.

Pleasure use is associated with the lowest average claim severity and a relatively low claim count, indicating that leisure driving poses the least risk. It results in both fewer claims and lower financial losses, making it the lowest-risk category in terms of both frequency and severity in the UK Automobile Collision Claims dataset.

Table 3. Analysis of Average Claim Severity by Age

Age	Average Claim Severity
17–20	391.80
21–24	293.17
25–29	284.84
30–34	279.73
35–39	212.43
40–49	249.99
50–59	251.11
60+	247.68

Table 3 shows that drivers aged 17–20 have the highest average claim severity, indicating that accidents involving the youngest drivers tend to result in greater financial losses and represent a substantial risk to insurers. A notable decrease in average claim severity is observed among drivers aged 21–24, indicating a reduced but still relatively high financial risk as drivers gain minimal experience. The trend of decreasing claim severity continues in the 25–29 age group, reflecting a further decline in financial impact as drivers mature and gain experience. This downward trend persists in the 30–34 age group, with a slight reduction in average claim severity compared to the previous cohort. A substantial drop is observed in the 35–39 age group, indicating a much lower severity of claims and a correspondingly reduced financial risk. Interestingly, the 40–49 age group sees a modest increase in average claim severity compared to the 35–39 group, though it remains lower than that of drivers under 30, indicating a moderate financial risk. Claim severity levels for the 50–59 age group are comparable to those of the 40–49 cohort, pointing to a stable level of financial risk among middle-aged drivers. Finally, drivers aged 60 and over exhibit slightly lower average claim severity than the 50–59 group, indicating a consistent and moderate financial risk, marginally higher than that of the 35–39 group but lower than the younger cohorts.

Table 4. The fitting of Negative Binomial and Generalized Poisson distributions

	Negative Binomial	**Generalized Poisson**
Parameters Estimate	$\hat{p} = 0.00453, \hat{r} = 1.25042$	$\hat{\theta} = 12.78279, \hat{\lambda} = 0.95426$
N-Loglikelihood	-211.9633	-216.1883
AIC	427.9267	436.3767
BIC	430.8581	439.3081

As shown in Table 4, the Negative Binomial distribution yields a higher (i.e. less negative) log-likelihood and the lowest AIC and BIC values, clearly indicating a superior fit to the AutoCollision claim count data compared to the Generalized Poisson distribution. These results indicate that the Negative Binomial model is more appropriate for capturing the underlying data structure. Both AIC and BIC are essential for model selection, as they balance goodness-of-fit with model complexity, thereby mitigating the risk of overfitting, an especially important consideration in actuarial modeling. Beyond information criteria, residual diagnostics further validate this conclusion. A comparative analysis of the Negative Binomial (see **Fig. 5**) and Generalized Poisson models (see **Fig. 6**) reveals that the former produces Pearson and deviance residuals tightly clustered around zero, with minimal dispersion and no extreme outliers. The histogram of Pearson residuals is approximately symmetric and unimodal. In contrast, the Q–Q plot of deviance residuals aligns closely with the theoretical quantile line, indicating that the model assumptions are well met. In contrast, the Generalized Poisson model exhibits more dispersed residuals, noticeable outliers, a skewed residual histogram, and a Q–Q plot that substantially deviates from the reference line, indicating potential model misspecification. Taken together, these statistical and graphical diagnostics confirm that the Negative Binomial model provides a more accurate and reliable representation of the claim count data, establishing it as the preferred modeling choice for this analysis.

4.1. Computation of aggregate claim

In this section, we used the estimates of $\hat{a}$, $\hat{b}$, and $\hat{c}$for (2), (4), and (10) to compute aggregate claim amounts for each recursive algorithm and their computational run time using Claim count data from the AutoCollision data. Based on Table 5, the performance of the three recursion algorithms is analyzed in terms of computational run time and the computed aggregate claim amounts.

The truncated Schröter recursion algorithm demonstrates the fastest run time at 0.051093 seconds and yields the highest aggregate claim sum of 0.005483, indicating that it either captures more aspects of the claim data or incorporates a more comprehensive modeling approach (see **Fig. 1**). This result implies a probability of approximately 0.55% that the total claim amount will not exceed 32 units, and conversely, a 99.45% probability that it will exceed this threshold.

The Panjer recursion algorithm, with a slightly longer runtime of 0.060898 seconds, computes an aggregate claim sum of $ 0.004887. Although still efficient, this result may indicate a more conservative or less data-sensitive approach (see Fig. 2). The corresponding probability that the total claim amount does not exceed $32 is approximately 0.49%, implying a 99.51% chance of exceeding this amount. The standard Schröter recursion algorithm, which has the longest runtime at 0.173438 seconds, produces an aggregate claim sum of $ 0.004930. This outcome implies a balance between sensitivity and comprehensiveness; however, it comes with higher computational demands due to the convolution component involved in the algorithm (see Fig. 3). The probability that the total claim amount will not exceed 32 units is approximately 0.45%. In contrast, the probability that it will exceed 32 units is around 99.55%.

The general interpretation of these results is that the likelihood of the total claim amount being less than or equal to 32 units is very low, with probabilities ranging from approximately 0.45% to 0.55%.

Consequently, the probability that the total claim amount will exceed 32 units is extremely high, ranging from 99.45% to 99.55% for the AutoCollision dataset. These findings indicate that, across all recursion algorithms evaluated, it is almost certain that total claims will surpass 32 units, underscoring the high-risk nature of the claims being modeled.

These results provide valuable insights for effective risk management and premium setting in the insurance sector. The high probability of large aggregate claims indicates that insurers must prepare for substantial payouts. Understanding this risk landscape allows insurers to more accurately assess claim distributions and frequencies, leading to more informed pricing strategies that ensure financial sustainability. Insurers can utilize these insights to allocate adequate reserves for high-expectation claims, thereby reducing the risk of insolvency. Moreover, policy designs can incorporate deductibles, limits, and exclusions that align with the high likelihood of large claims, striking a balance between customer affordability and insurer profitability. These findings also support the development of targeted reinsurance strategies, allowing insurers to transfer a portion of high-risk exposures and minimize the financial impact of large claims.

Table 5. Aggregate claim for the truncated Schröter, Panjer, and Schröter algorithms

s	g(s) The Truncated Schröter	g(s) The Panjer	g(s) The Schröter
1	711000×10^{-3}	711000×10^{-3}	711000×10^{-3}
2	13488×10^{-5}	11953×10^{-5}	11958×10^{-5}
3	78790×10^{-6}	69718×10^{-6}	69763×10^{-6}
4	18245×10^{-6}	16054×10^{-6}	16132×10^{-6}
5	21313×10^{-5}	18881×10^{-5}	18908×10^{-5}
6	58042×10^{-5}	51406×10^{-5}	51450×10^{-5}
7	32214×10^{-5}	28454×10^{-5}	28493×10^{-5}
8	15979×10^{-5}	14070×10^{-5}	14113×10^{-5}
9	48176×10^{-5}	42617×10^{-5}	42727×10^{-5}
10	11819×10^{-4}	10454×10^{-4}	10474×10^{-4}
11	11254×10^{-4}	99323×10^{-5}	99514×10^{-5}
12	49170×10^{-5}	43135×10^{-5}	43306×10^{-5}
13	46463×10^{-5}	40787×10^{-5}	41095×10^{-5}
14	15944×10^{-4}	14065×10^{-4}	14119×10^{-4}
15	13782×10^{-4}	12090×10^{-4}	12149×10^{-4}
16	74915×10^{-5}	65005×10^{-5}	65502×10^{-5}
17	65619×10^{-5}	57003×10^{-5}	57681×10^{-5}
18	18092×10^{-4}	15882×10^{-4}	15999×10^{-4}
19	16315×10^{-4}	14191×10^{-4}	14324×10^{-4}
20	95516×10^{-5}	81572×10^{-5}	82706×10^{-5}
21	11549×10^{-4}	99748×10^{-5}	10108×10^{-4}
22	36393×10^{-4}	31961×10^{-4}	32173×10^{-4}
23	30416×10^{-4}	26473×10^{-4}	26719×10^{-4}
24	17428×10^{-4}	14879×10^{-4}	15094×10^{-4}
25	15023×10^{-4}	12823×10^{-4}	13062×10^{-4}
26	35570×10^{-4}	30989×10^{-4}	31358×10^{-4}
27	27630×10^{-4}	23653×10^{-4}	24074×10^{-4}
28	17519×10^{-4}	14558×10^{-4}	14920×10^{-4}
29	18952×10^{-4}	15950×10^{-4}	16353×10^{-4}

Table 5. Aggregate claim for the truncated Schröter, Panjer, and Schröter algorithms (cont.)

s	g(s) The Truncated Schröter	g(s) The Panjer	g(s) The Schröter
30	30447×10^{-4}	26076×10^{-4}	26676×10^{-4}
31	27511×10^{-4}	23012×10^{-4}	23706×10^{-4}
32	22584×10^{-4}	18434×10^{-4}	19048×10^{-4}
Sum of Probabilities	0.005483	0.004887	0.004535
Execution time in seconds(s)	0.051093	0.060898	0.173438

The observed differences in computational run times and aggregate claim sums across algorithms are attributable to the inherent complexity and structural differences of the recursion methods. The truncated Schröter algorithm, with its three-parameter structure, strikes an efficient balance between model complexity and computational speed, yielding both fast run times and higher aggregate claims. The Panjer recursion algorithm, while simpler with only two parameters, offers efficient computation but may not capture as many underlying data features. In contrast, the Schröter recursion algorithm, which incorporates an additional convolution term, requires more computation time but provides a nuanced perspective on aggregate claim modeling.

4.2. Simulation study

Here, we generate random claim amounts data from the Negative binomial distribution by setting $r = 100$ and $p = 0.05$. We varied the sample size to examine the aggregate claim computational efficiency and run time of the truncated Schröter, Panjer, and Schröter recursion algorithms. Initially, we generated 5000 random numbers from the Negative binomial distribution and fit (7) to the data to obtain the estimate of the parameters $\hat{a}$, $\hat{b}$, and $\hat{c}$as $\hat{a} = 0.9905$, $\hat{b} = 18.3096$, $\hat{c} = -1.2840$, and compute$g(0) = 0.00117$ to implement the algorithms. Tables 6, 7, and 8 present the sample sizes, aggregate claim amounts, and the execution time in seconds for each recursion algorithm.

Fig. 4 illustrates the execution times of the truncated Schröter recursion, Panjer recursion, and Schröter recursion algorithms for varying values of n, highlighting significant differences in computational efficiency as the sample size increases. The truncated Schröter recursion algorithm consistently demonstrates the lowest execution times across all values of n, starting at 0.0006919 s for n = 20 and increasing to 1.5046701 s for n = 5000 (see **Fig. 4**).

Table 6. Efficiency of the truncated Schröter algorithm on simulated claim data

Recursion Algorithm	Sample (n)	sum of g(s)	Execution time (s)
The truncated Schröter algorithm			
	20	2.7080348	0.0006919
	50	1.5211150	0.0036724
	100	1.1730166	0.0188396
	150	1.1747598	0.0335643
	200	1.1296751	0.0586591
	300	0.9208012	0.1158113
	600	0.8504156	0.3374069
	1500	0.6127028	0.8211629
	2000	0.5397614	1.0198436
	3000	0.4756758	1.2040498
	4000	0.4180231	1.4369745
	5000	0.3979199	1.5046701

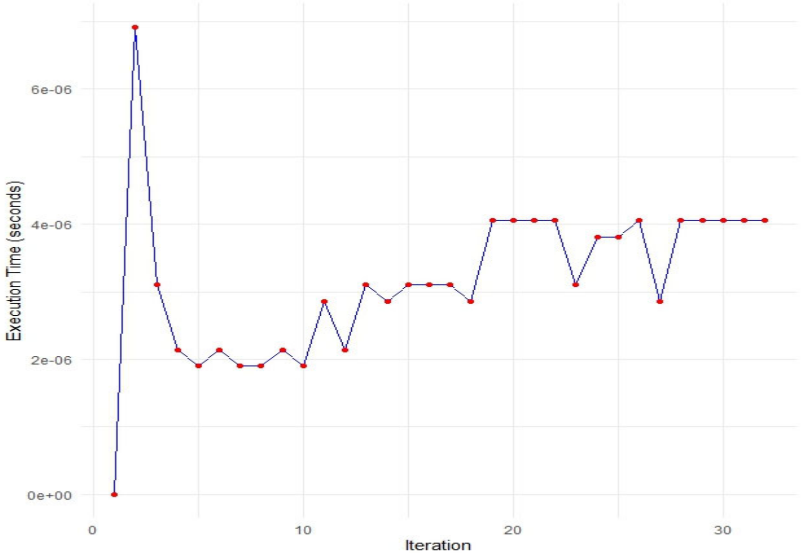


Figure 1. The execution time plot of the truncated Schröter recursion algorithm for each iteration

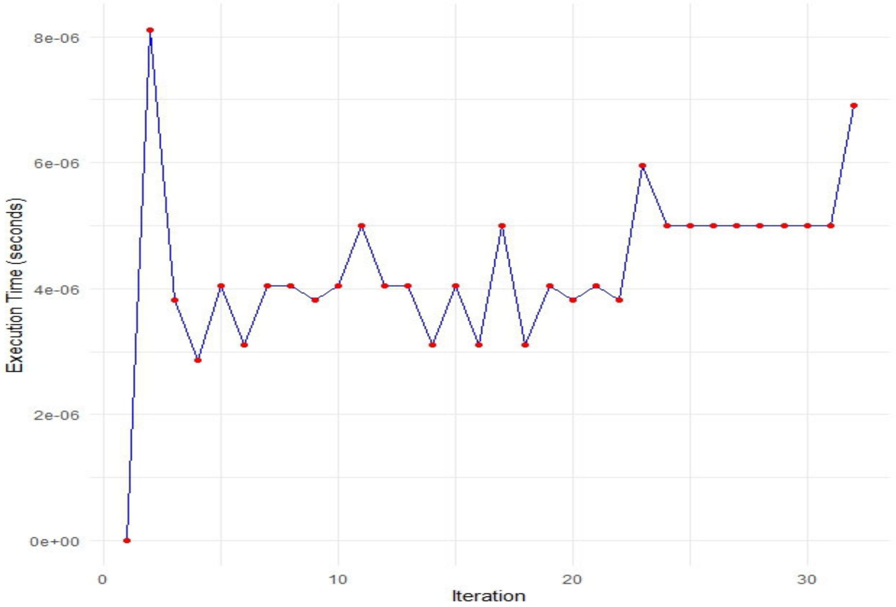


Figure 2. The execution time plot of the Panjer recursion algorithm for each iteration

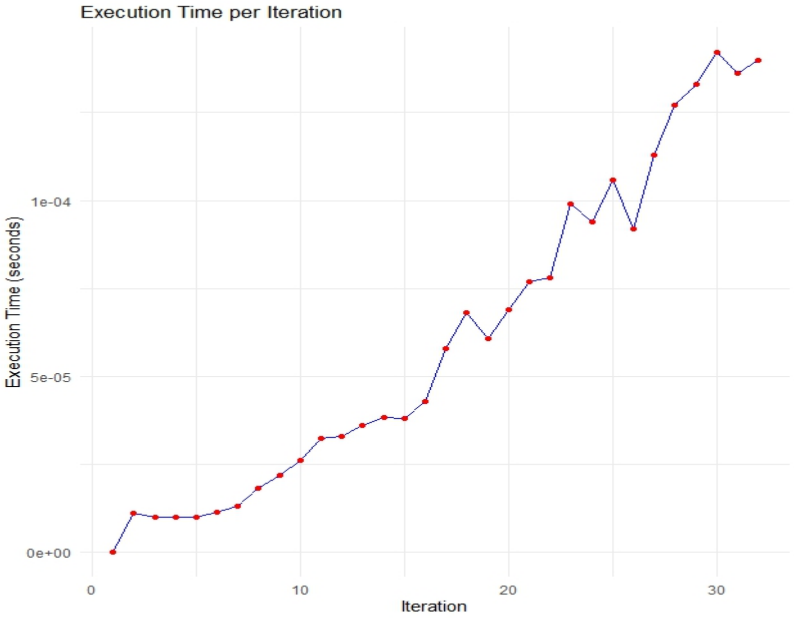


Figure 3. The execution time plot of the Schröter recursion algorithm for each iteration

Table 7. Efficiency of the Panjer algorithm on simulated claim data

Recursion Algorithm	Sample (n)	sum of g(s)	Execution time (s)
The Panjer algorithm			
	20	0.0075441	0.0014127
	50	0.0074026	0.0047266
	100	0.0074002	0.0212255
	150	0.0074619	0.0378468
	200	0.0074672	0.0692441
	300	0.0072761	0.1391730
	600	0.0073219	0.4021811
	1500	0.0073130	1.0212069
	2000	0.0072711	1.0606146
	3000	0.0072833	1.3638492
	4000	0.0072209	1.6406126
	5000	0.0072351	1.7650454

Table 8. Efficiency of the Schröter algorithm on simulated claim data

Recursion Algorithm	Sample (n)	sum of g(s)	Execution time (s)
The Schröter algorithm			
	20	0.0062785	0.0028598
	50	0.0063757	0.0195415
	100	0.0064331	0.0765483
	150	0.0064909	0.1735694
	200	0.0065073	0.2585254
	300	0.0063946	0.6078202
	600	0.0064442	1.8534706
	1500	0.0064868	4.6577935
	2000	0.0064729	5.4729755
	3000	0.0065035	6.8906786
	4000	0.0064749	8.0277340
	5000	0.00649490	8.7507973

This performance indicates that the algorithm is highly efficient and scalable, capable of handling larger datasets with minimal computational burden. The Panjer recursion algorithm also exhibits increasing execution times with larger n, beginning at 0.0014127 s for $n = 20$ and rising to 1.7650454 s for $n = 5000$. While reasonably efficient, it demonstrates less scalability compared to the truncated Schröter algorithm (see **Fig. 4**). In contrast, the Schröter recursion algorithm, which includes a convolution component,f_i^{2*}, shows substantially higher execution times, starting at 0.0028598 s for $n = 20$ and escalating sharply to 8.7507973 s for $n = 5000$ (see **Fig. 4**). This steep increase reflects poor scalability and reduced efficiency, particularly for large sample sizes, making it the least optimal option among the three algorithms evaluated. Overall, the truncated Schröter recursion algorithm emerges as the most efficient and scalable, followed by the Panjer recursion algorithm. The Schröter recursion algorithm, while potentially offering greater modeling flexibility, is substantially less efficient due to its computational complexity.

To assess the consistency of execution times, each sample size was tested across five independent runs. The variation in computational times was negligible, indicating that the execution times were stable and reproducible. However, it is worth noting that minor fluctuations may still be influenced by the operational state of the computing system during execution.

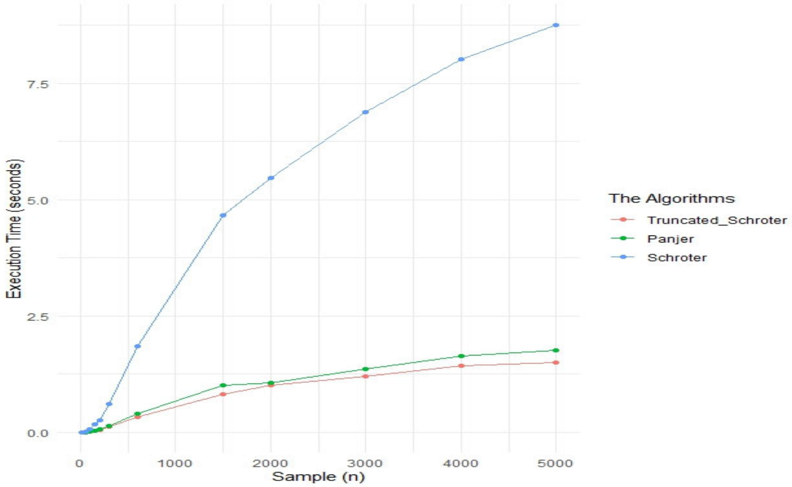


Figure 4. Visual representation of the execution time of the truncated Schröter recursion, Panjer recursion, and Schröter recursion algorithms

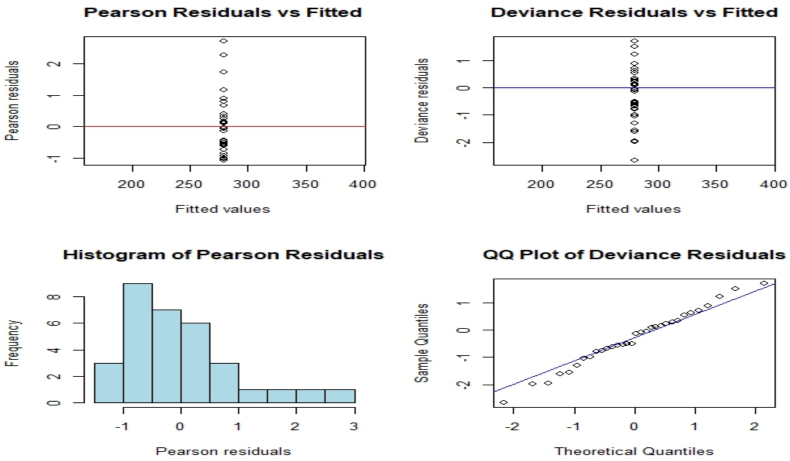


Figure 5. Graphical residual analysis of the fitted Negative Binomial distribution to the claim data

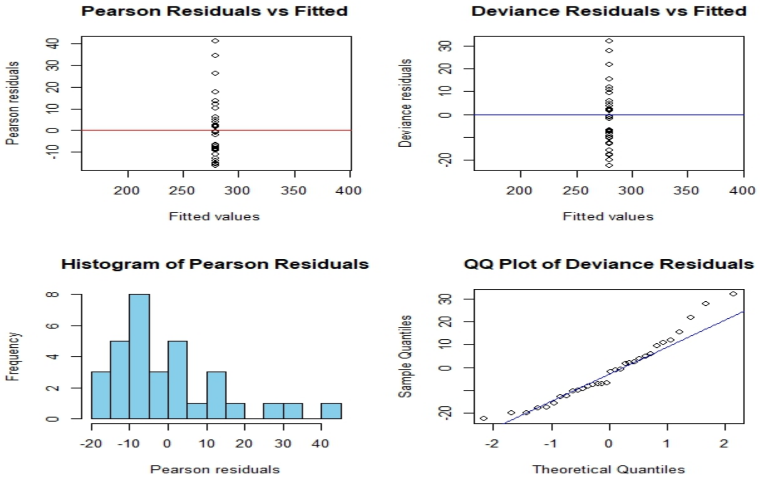


Figure 5. Graphical residual analysis of the fitted Generalized Poisson distribution to the claim data

5. Conclusion

This study investigated the computation of aggregate claim amounts using various recursive algorithms, with a particular focus on the newly introduced truncated Schröter recursion algorithm. The primary objective was to enhance both the accuracy and computational efficiency of aggregate claim estimation, an essential component of effective risk management and premium setting in the insurance industry.

The truncated Schröter recursion algorithm demonstrated superior performance in numerical evaluations and comparative analysis. When applied to the AutoCollision dataset, it consistently delivered the fastest execution times and the highest aggregate claim sums, indicating both computational efficiency and modeling comprehensiveness. For modeling claim count data, the Negative Binomial distribution was favored over the Generalized Poisson distribution due to its ability to accommodate overdispersion, as supported by AIC and BIC selection criteria.

Simulation studies further validated the performance of the truncated Schröter algorithm across varying sample sizes, consistently outperforming the Panjer and standard Schröter recursion algorithms in terms of execution time and scalability, while effectively capturing data variability for more refined analysis. The findings also underscored the critical importance of selecting appropriate counting distributions and recursion methods when modeling aggregate claim amounts.

In conclusion, the truncated Schröter recursion algorithm emerges as a robust and reliable tool for calculating aggregate claim amounts, offering a strong balance between computational speed and modeling accuracy. Its adoption has the potential to improve risk assessment substantially and premium pricing strategies, ultimately benefiting both insurers and policyholders. Future research could explore enhancements to this algorithm, such as incorporating machine learning techniques to optimize parameter estimation based on evolving claim patterns dynamically. Moreover, applying the algorithm to other insurance domains beyond automobile claims could further validate its generalizability and inform domain-specific refinements.

Funding

This study was supported by the Slovak Academy of Sciences Doktogrant: APP0515 (Friday I. Agu).

Acknowledgement

The author expresses his sincere gratitude to his esteemed supervisor, Dr. Jan Macutek, for his invaluable guidance, support, and encouragement throughout this study. His unwavering belief in my abilities and insightful mentorship have shaped my academic and professional development. I am deeply grateful for his continued support and inspiration.

Conflict of interest

There is no conflict of interest for this study.

References

Agu, F. I., Mačutek, J. and Szűcs, G., (2023). A Simple Estimation of Parameters for Discrete Distributions from the Schröter Family. *Statistika: Statistics & Economy Journal*, 103(2).

Albrecher, H., Beirlant, J. and Teugels, J. L., (2017). Reinsurance: actuarial and statistical aspects. *John Wiley & Sons.*

Beard, R. E., Pentikäinen, T. and Pesonen, E., (1977). Risk theory (2nd ed.). *Chapman and Hall.*

Cooley, J. W., Tukey, J. W., (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*, 19(90), pp. 297–301.

Dickson, D. C. (2016). Insurance risk and ruin. *Cambridge University Press.*

Dzidzornu, S. B., Minkah, R., (2021). Assessing the Performance of the Discrete Generalised Pareto Distribution in Modelling Non-life Insurance Claims. *Journal of Probability and Statistics*, 2021(1), 5518583.

Fackler, M., (2023). Panjer class revisited: one formula for the distributions of the Panjer (a, b, n) class. *Annals of Actuarial Science*, 17(1), pp. 145–169.

Gamaleldin, W., Attayyib, O., Alnfiai, M. M., Alotaibi, F. A. and Ming, R., (2025). A hybrid model based on CNN-LSTM for assessing the risk of increasing claims in insurance companies. *PeerJ Computer Science*, 11, e2830.

Ghinawan, F., Nurrohmah, S. and Fithriani, I., (2021). Recursive and moment-based approximation of aggregate loss distribution. In *Journal of Physics*: *Conference Series*, Vol. 1725, No. 1, p. 012101. *IOP Publishing*.

Gray, R. J., Pitts, S. M., (2012). Risk modeling in general insurance: From principles to practice. *Cambridge University Press*.

Heckman, P. E., Meyers, G. G., (1983). The calculation of aggregate loss distributions from claim severity and claim count distributions. *In Proceedings of the Casualty Actuarial Society* , Vol. 70, No. 133–134, pp. 49–66. Casualty Actuarial Society.

Hofmann, L., (2022). Approximation Methods for the Total Claim Amount in Collective Risk Modeling/submitted by Hofmann Louisa.

Hogg, R. V., Klugman, S. A., (2009). Loss distributions. *John Wiley & Sons*.

Jindrová, P., Pacáková, V., (2016). Modeling of extreme losses in natural disasters. *International Journal of Mathematical Models and Methods in Applied Sciences*, Vol. 10, issue 2016.

Klugman, S. A., Panjer, H. H. and Willmot, G. E., (2012). Loss models: from data to decisions, Vol. 715. *John Wiley & Sons*.

Mildenhall, S., (2024). Aggregate: fast, accurate, and flexible approximation of compound probability distributions. *Annals of Actuarial Science*, pp. 1–40.

Mildenhall, S. J., Major, J. A., (2022). Pricing insurance risk: Theory and practice. *John Wiley & Sons*.

Pacáková, V., Gogola, J., (2013). Pareto Distribution in Insurance and Reinsurance. In Conference proceedings from 9th International Scientific Conference Financial Management of Firms and Financial Institutions. *VŠB Ostrava*, pp. 298–306.

Packová, V., Brebera, D., (2015). Loss distributions in insurance risk management. *Recent advances in economics and business administration*, pp. 17–22.

Panjer, H. H., (1981). Recursive evaluation of a family of compound distributions. ASTIN Bulletin: *The Journal of the IAA*, 12(1), pp. 22–26.

Qiu, D., (2019). Individual claims reserving: Using machine learning methods (*Doctoral dissertation, Concordia University*).

Schröter, K. J., (1990). On a family of counting distributions and recursions for related compound distributions. *Scandinavian Actuarial Journal*, 1990(2–3), pp. 161–175.

Sundt, B., Vernic, R., (2009). Recursions for convolutions and compound distributions with insurance applications. *Springer Science & Business Media*.

Tzaninis, S. M., Bozikas, A., (2024). Extensions of Panjer's recursion for mixed compound distributions. *arXiv preprint arXiv*:2406.17726.

Yartey, E., (2020). The (a, b, r) class of discrete distributions with applications (*Doctoral dissertation, Laurentian University of Sudbury*).

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 111–125, https://doi.org/10.59139/stattrans-2025-042
Received – 16.05.2024; accepted – 28.01.2025

Inverse Power Lomax Poisson distribution: properties and applications in modelling negatively-skewed reliability data

Adebisi A. Ogunde[1], **Emmanuel F. Nymphas**[2]

Abstract

In this paper, we propose a new, four-parameter distribution with increasing, decreasing, bathtub-shaped and a unimodal failure rate, called the Inverse Power Lomax Poisson (IPLP) distribution. The new distribution combines Inverse Power Lomax (IPL) and Poisson distributions. We derive several properties of the new distribution: its probability density function, its reliability and failure rate functions, the quantiles, the stress-strength parameter, complete and incomplete moments, the moment generating function, the probability weighted moment, Rènyi and q-entropies, and order statistics. The study presents the estimation of the model's parameters based on the maximum likelihood method. The applications of the new distribution are presented using two real data sets, showing its flexibility and potential in modelling lifetime data.

Key words: probability weighted moments, incomplete moments, quantile function, Renyi entropy.

1. Introduction

The Inverse Power Lomax (IPL) distribution, introduced and developed by Hassan and Abd-Allah (2019), as a reciprocal of the Power Lomax distribution, contains distributions with bathtub-shaped and unimodal failure rates, as well as a broader class of monotone failure rates. The IPL model provides a tractable and close-form solution to many problems in reliability studies. However, it does not give a reasonably good parametric fit in some real-life applications most especially when the data is extremely skewed, Hassan and Abd-Allar (2019). However, several works have been done to develop new families of probability distributions that extend standard probability distributions while at the same time making them more flexible and tractable. Abdul-

[1] Department of Statistics, University of Ibadan, Ibadan, Oyo State, Nigeria. E-mail: debiz95@yahoo.com. ORCID: https://orcid.org/0000-0001-8708-8612.

[2] Corresponding author. Department of Physics, University of Ibadan, Ibadan, Oyo State. Nigeria. E-mail: efnda@yahoo.co.uk.

Moniem and Abdel-Hameed (2012) studied the exponentiated Lomax distribution. Lemonte and Cordeiro (2013) studied the properties of beta Lomax, Kumaraswamy Lomax and McDonald developed the Lomax distributions. Cordeiro et al. (2013) introduced the gamma-Lomax model. The Weibull Lomax was proposed and studied by Tahir et al. (2015). The Gumbel-Lomax distribution was investigated by Tahir et al. (2016). The type II Topp Leone power Lomax distribution was studied by Al-Marzouki et al. (2020), Haq et al. (2020) studied the Marshal-Olkin Power Lomax distribution. The sine Power Lomax and the sine Inverse Power Lomax distribution was studied by Nagarjuma and Chesneau (2021, 2022). The Kumaraswamy generalised Inverse Lomax and the type II Topp-Leone Inverse Power Lomax distributions were proposed and studied by Ogunde et al. (2023, 2024). They developed the new model using the Kumaraswamy and type II Topp-Leone generators, respectively.

A random variable X follows the IPL distribution if its cumulative distribution function (CDF) takes the form

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

The corresponding probability density function (PDF) is

$$g(x;\alpha,\rho,\lambda)=\frac{\alpha\rho}{\lambda}x^{-\alpha-1}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho-1},\ x>0;\ \alpha,\rho,\lambda>0 \tag{2}$$

The survival and hazard rate functions of the IPL distribution are, respectively,

$$S(x;\alpha,\rho,\lambda)=1-G(x;\alpha,\rho,\lambda)=1-\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}, \tag{3}$$

and

$$h(x;\alpha,\rho,\lambda)=\frac{g(x;\dot{\alpha},\rho,\lambda)}{S(x;\alpha,\rho,\lambda)}=\frac{\alpha\rho x^{-\alpha-1}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho-1}}{\lambda\left\{1-\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}\right\}}. \tag{4}$$

Where α and ρ are positive shape parameters and λ is a scale parameter. In the literature, several authors proposed a new distribution to model lifetime data by combining some discrete distribution together with other known continuous distributions. Roman et al. (2012) proposed a long-term exponential geometric distribution. Recently, compounding distributions for Lomax with discrete one has been presented by some authors. For instance; the Lomax Poison distribution was proposed by Abd-Elfattah et al. (2013). Ramos et al. (2013) studied the exponentiated Lomax Poisson distribution, Al-Zahrani and Sagor (2014) developed and studied the Lomax-Logarithm distribution. Al-Zahrani (2015) and Hassan and Abd-Alla (2017) developed the extended Poisson Lomax and the exponentiated Lomax distribution, respectively. Hassan and Nassr (2018) investigated the properties of the Power Lomax Poisson distribution. Nargajuma et al. (2022) proposed and studied the Nadarajah–Haghighi Lomax distribution, among many others.

In this study, we propose and study a new four-parameter distribution, named the Inverse Power Lomax Poisson ($IPLP$) distribution, which contains the Inverse Power

Lomax (IPL), the Inverse Lomax Poisson (ILP), and the Inverse Lomax distributions as the sub-models. The chief motivation for introducing the $IPLP$ distribution is that the distribution, due to its flexibility, can accommodate different forms of the shape of the hazard function. The distribution also provides a reasonable parametric fit to skewed data that cannot be properly fitted by other distributions and is a suitable model in other areas such as insurance, seismography, medicine, actuarial science, demography, reliability, and survival studies.

The paper is organized as follows. In Section 2, we developed the $IPLP$ distribution and derived its density, survival and hazard rate, cumulative, and reversed hazard rate, and the quantile functions. Some of the properties of the $IPLP$ distribution are given in Section 3, which includes moments, moment generating functions, incomplete moments, Renyi and q entropies, probability weighted moments, and order statistics. Estimation and real data application was demonstrated in Section 4. In Section 5, we concluded.

2. The Inverse Power Lomax Poisson distribution

Suppose that the random variable X has the IPL distribution, where its cdf and pdf are given in (1) and (2). Given N, let $X_1, \ldots, X_n$ be independent and identify distributed random variables from IPL distribution. Suppose N is distributed according to zero truncated Poisson distribution with PDF

$$T(N = n) = \frac{e^{\zeta}\zeta^{n}}{n!(1-e^{-\zeta})}, \qquad n = 1,2,\ldots,\ \zeta > 0 \tag{5}$$

Let $T = max(X_1, \ldots, X_N)$, then the CDF of $T/N = n$ is given by

$$F_{T/N=n}(t) = \left(1 + \frac{t^{-\alpha}}{\lambda}\right)^{-\rho n}, \tag{6}$$

which is the Exponentiated Inverse Power Lomax distribution with parameters ρn, α and λ. The $IPLP$ distribution, represented by $IPLP(\alpha,\ \lambda,\ \rho,\ \zeta)$, is defined by the marginal CDF of T, i.e.

$$F(x;\alpha,\rho,\lambda,\zeta) = \frac{1-e^{-\zeta\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}}}{-e-^{\zeta}+1}, \tag{7}$$

This newly developed distribution contains the Inverse Lomax and the Inverse Lomax Poisson distribution. The pdf of the $IPLP$ distribution is given by

$$f(x;\alpha,\rho,\lambda,\zeta) = \frac{\alpha\rho\zeta x^{-\alpha-1}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho-1}\left\{e^{-\zeta\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}}\right\}}{\lambda(-e^{-\zeta}+1)}, \tag{8}$$

where α, ρ, ζ are positive shape parameters and λ is a positive scale parameter. The reliability ($R(x)$) and hazard rate (($h(x)$) functions, reversed hazard and cumulative hazard functions of the $IPLP$ distribution are, respectively, given by

$$R(x;\alpha,\rho,\lambda,\zeta) = 1 - \frac{1-e^{-\zeta\left(1+\frac{t^{-\alpha}}{\lambda}\right)^{-\rho}}}{-e^{-\zeta}+1} = \frac{e^{-\zeta\left(1+\frac{t^{-\alpha}}{\lambda}\right)^{-\rho}} - e^{-\zeta}}{-e^{-\zeta}+1}, \tag{9}$$

$$h(x;\alpha,\rho,\lambda,\zeta)=\frac{\alpha\rho\zeta x^{-\alpha-1}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho-1}e^{-\zeta\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}}}{\lambda e^{-\zeta}\left\{e^{-\zeta\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}}-1\right\}}, \tag{10}$$

$$\varphi(x;\alpha,\rho,\lambda,\zeta)=\frac{\alpha\rho\zeta x^{-\alpha-1}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho-1}e^{-\zeta\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}}}{\lambda\left\{1-e^{-\zeta\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}}\right\}}, \tag{11}$$

and

$$H=log\left(1-e^{-\zeta\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}}\right)-log\left(-e^{-\zeta}+1\right). \tag{12}$$

The plots of distribution, density and hazard rate functions of the *IPLP* distribution for different values of $(\alpha,\rho,\zeta,\lambda)$ are given in Figures 1, 2 and 3, respectively.

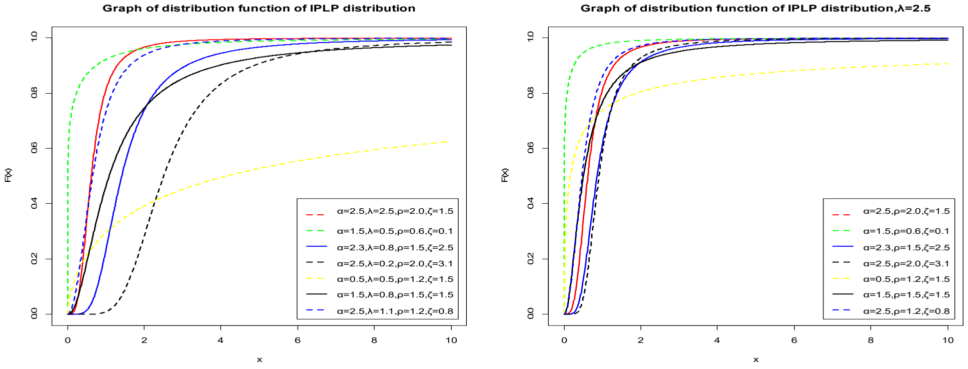


Figure 1. Graph of distribution function of *IPLP* distribution

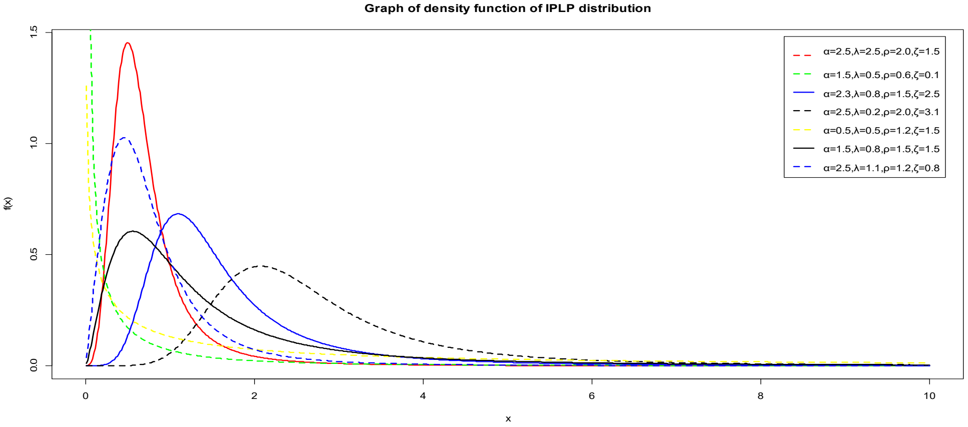


Figure 2. Graph of density function of *IPLP* distribution

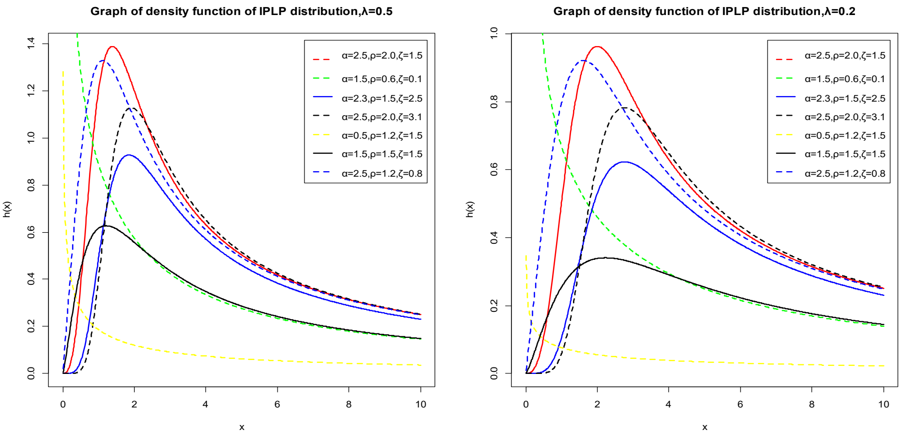


Figure 3. Graph of hazard function of $IPLP$ distribution

From Figure 3, it can be observed that the hazard rate function of the $IPLP$ model exhibits decreasing, increasing, reversed bathtub, and reversed J-shape curves. This indicates that the $IPLP$ distribution can be used effectively to model skewed data exhibiting various shapes of the hazard function.

2.1. Quantiles of the $IPLP$ distribution

The quantile function can be used in the study of some important features and characteristics of a distribution which includes dispersion, skewness and kurtosis. Also, the quantiles of a distribution can be employed in data generation from a distribution. The k^{th} quantile of the $IPLP$ distribution is given by

$$x_k = \left\{\lambda\left[\left(-\frac{1}{\zeta} ln[1+(1-k)(-e^{-\zeta}+1)]\right)^{-1/\rho} - 1\right]\right\}^{-1/\alpha}, \tag{13}$$

which is used for data generation from the $IPLP$ distribution. The median (middle quartile) and the upper quartiles of the $IPLP$ distribution can be obtained by taking $k = 0.5$ and 0.75 respectively.

2.2. Mixture representation of $IPLP$ model

Using the binomial series expansion given by

$$e^z = \sum_{p=0}^{\infty} \frac{z^p}{p!}, \tag{14}$$

the mixture representation of $IPLP$ model is given by

$$f(x;\alpha,\rho,\lambda,\zeta) = \frac{\alpha\rho\zeta}{\lambda} \sum_{m=0}^{\infty} \frac{1}{m!(e^{\zeta}-1)} x^{-\alpha-1} \left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-[\rho(m+1)+1]}. \tag{15}$$

The expression given in (15) can be described as the Exponentiated Inverse Power distribution with scale parameter λ and shape parameters α and $\rho(m+1)$.

3. Statistical Properties of $IPLP$ distribution

The following properties of the IPLP model are investigated.

3.1. Moments of $IPLP$ distribution

The r^{th} moments of the $IPLP$ distribution can be expressed as

$$\mu_r' = E(X^r) = \int_{-\infty}^{\infty} x^r f(x;\alpha,\rho,\lambda,\zeta)dx \tag{16}$$

Using (15) in (16) we get

$$\mu_r' = \frac{\alpha\rho\zeta}{\lambda}\sum_{m=0}^{\infty}\frac{1}{m!(e^{\zeta}-1)}\int_{-\infty}^{\infty} x^{r-\alpha-1}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-[\rho(m+1)+1]} dx \tag{17}$$

After some algebraic manipulation, we have

$$\mu_r' = \frac{\rho\zeta}{\lambda^{r/\alpha}}\sum_{m=0}^{\infty}\frac{(-1)^m}{m!(e^{\zeta}-1)}B[(1-{}^{r}/_{\alpha}),({}^{r}/_{\alpha}+\zeta(m+1)] \tag{18}$$

The moment generating function of X, $M_x(t)$, is given by

$$M_x(t) = \int_{-\infty}^{\infty} e^{tX} f(x;\alpha,\rho,\lambda,\zeta)dx = \sum_{r=0}^{\infty}\frac{t^r}{r!}E(X^r). \tag{19}$$

Using the expression given in (18) for the r^{th} moments of the $IPLP$ distribution, we have

$$M_x(t) = \frac{\rho\zeta}{\lambda^{r/\alpha}}\sum_{m=0}^{\infty}\frac{t^r}{m!r!(e^{\zeta}-1)}B[(1-{}^{r}/_{\alpha}),({}^{r}/_{\alpha}+\zeta(m+1)], \tag{20}$$

where $B(a,b) = \frac{\Gamma a\Gamma b}{\Gamma(a+b)}$. From the above expression in (20), setting $r = 1,2,3,4,5,$ and 6, respectively, we obtain the first six moments about the origin of $IPLP$ distribution.

The n^{th} central moment of X, of $IPLP$ model, say μ_n, is given as

$$\mu_n = E(x-\mu)^n = \sum_{p=0}^{\infty}(-1)^p\binom{n}{p}\mu_r'^{p}\mu_{n-p}'.$$

The cumulant ($ϗ_n$) of X can be obtained as

$$ϗ_n = \mu_n' - \sum_{r=0}^{n-1}\binom{n-1}{r-1}ϗ_r\mu_{n-r}'.$$

Table 1 presents the first six moments, standard deviation (σ), coefficient of variation (CV), skewness (S_k), and kurtosis (k_u) for various values of the parameters of $IPLP$ distribution. It could be observed that as the values of the parameters increase the values of the lower moment decrease and increase for higher moments. The same is observed for skewness and kurtosis except for higher values of the parameters. This further demonstrates the flexibility of the $IPLP$ model in handling data of different degree of skewness and kurtosis.

Table 1. First six moments, σ, CV, S_k, and k_u for $IPLP$ distribution

Specification	$\alpha = 7.0, \lambda = 5.5$				
$Moment$	$\rho = 0.5,$ $\zeta = 0.5$	$\rho = 1.2,$ $\zeta = 1.5$	$\rho = 2.5,$ $\zeta = 4.0$	$\rho = 3.5,$ $\zeta = 4.5$	$\rho = 6.5,$ $\zeta = 6.5$
μ_1'	0.6525	0.8148	0.9165	0.9015	0.8988
μ_2'	0.4776	0.7011	0.8666	0.8320	0.8161
μ_3'	0.3899	0.6448	0.8538	0.7917	0.7498
μ_4'	0.3600	0.6491	0.8938	0.7877	0.6989
μ_5'	0.3944	0.7546	1.0413	0.8491	0.6661
μ_6'	0.6062	1.2026	1.5792	1.1308	0.6703
σ	0.2274	0.1929	0.1632	0.1389	0.0909
CV	0.3485	0.2367	0.1781	0.1541	0.1011
S_k	0.8988	1.8008	2.4746	2.5583	1.9237
k_u	6.9213	13.0353	20.6319	22.5148	15.7185

3.2. Incomplete moment of IPLP distribution

The r^{th} incomplete moments of the $IPLP$ distribution is defined by

$$\varphi_r(t) = \alpha\rho\zeta \int_{-\infty}^{t} x^{r-\alpha-1} \frac{\left(1+\frac{x-\alpha}{\lambda}\right)^{-\rho-1}\left\{e^{-\zeta\left(1+\frac{x-\alpha}{\lambda}\right)^{-\rho}}\right\}}{\lambda(-e^{-\zeta}+1)} dx \tag{21}$$

Using (14), we can write the expression given in (21) as

$$\varphi_r(t) = \frac{\alpha\rho\zeta}{\lambda} \sum_{m=0}^{\infty} \frac{1}{m!(e^{\zeta}-1)} \int_{-\infty}^{t} x^{r-\alpha-1} \left(1+\frac{x-\alpha}{\lambda}\right)^{-[\rho(m+1)+1]} dx \tag{22}$$

After some algebraic manipulation, we have

$$\varphi_r(t) = \frac{\rho\zeta}{\lambda^{r/\alpha}} \sum_{m=0}^{\infty} \frac{1}{m!(e^{\zeta}-1)} B\left[(1 - {}^{r}/_{\alpha}), ({}^{r}/_{\alpha} + \zeta(m+1); \frac{t-\alpha}{\lambda}\right]. \tag{23}$$

3.3. Rènyi and q-entropies of $IPLP$ distribution

Suppose X is a random variable with continuous cumulative distribution function $F(x)$ and probability density function $f(x)$. Then the fundamental uncertainty measure for distribution F (named the entropy of F) is defined as $I(x) = E[-\log(f(X))]$. Statistical entropy is a probabilistic measure of uncertainty, also a measure of a reduction in that uncertainty. Numerous entropy and information indices are considered in the literature, among them the Rényi and q entropy. The Rènyi entropy of a random variable X can be used to obtain the measures of uncertainty and variation of a system and it is defined ($\partial > 0$ and $\partial \neq 1$) as:

$$I_R(\partial) = \frac{1}{1-\partial} log[M(\partial)], \tag{24}$$

where

$$M(\partial) = \int_{-\infty}^{\infty} f^{\partial}(x) dx,$$

Using

$$M(\partial)=\frac{\alpha^{\partial}\rho^{\partial}\zeta^{\partial}}{\lambda^{\partial}(-e^{-\zeta}+1)^{\partial}}\int_{-\infty}^{\infty}x^{-\partial(\alpha+1)}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\partial(\rho+1)}\left\{e^{-\zeta\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho}}\right\}^{\partial}dx, \tag{25}$$

After some algebraic manipulation we have

$$M(\partial)=\frac{\alpha^{\partial}\rho^{\partial}}{\lambda^{\partial}(-e^{-\zeta}+1)^{\partial}}\sum_{i=1}^{\infty}\frac{(-1)^{i}\zeta^{\partial+i}\partial^{i}}{i!}\int_{-\infty}^{\infty}x^{-\partial(\alpha+1)}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-[\partial(\rho+1)+\partial i]}dx, \tag{26}$$

Further simplification gives

$$M(\partial)=W^{i}B\left[\frac{\partial(\alpha+1)-1}{\alpha},\frac{2\alpha+1-\partial(\alpha+1)-\alpha\{\rho i-\partial(\rho+1)\}}{\alpha}\right] \tag{27}$$

where

$$W^{i}=\frac{\alpha^{\partial-1}\rho^{\partial}\lambda^{\frac{\partial-1}{\alpha}}}{(-e^{-\zeta}+1)^{\partial}}\sum_{i=1}^{\infty}\frac{(-1)^{i}\zeta^{\partial+i}\partial^{i}}{i!}.$$

Finally, we obtain an expression for the Renyi entropy of IPLP distribution as

$$I_R(\partial)=\frac{1}{1-\partial}log\left\{W^{i}B\left[\frac{\partial(\alpha+1)-1}{\alpha},\frac{2\alpha+1-\partial(\alpha+1)-\alpha\{\rho i-\partial(\rho+1)\}}{\alpha}\right]\right\}, \tag{28}$$

The q-entropy, Z_q, is defined by

$$Z_q=\frac{1}{q-1}log[1-(q-1)M(\partial)]$$

Using $M(\partial)$, we have

$$Z_q=\frac{1}{q-1}log\left[1-(q-1)W^{i}B\left[\frac{\partial(\alpha+1)-1}{\alpha},\frac{2\alpha+1-\partial(\alpha+1)-\alpha\{\rho i-\partial(\rho+1)\}}{\alpha}\right]\right]$$

3.4. Probability Weighted Moments (PWMs)

Probability weighted moments (PWMs) are defined as the expectations of certain functions of a random variable. They are only considered when the ordinary moments of the random variable exist. The PWMs method can generally be employed in estimating the parameters of a distribution whose inverse form cannot be expressed explicitly. In this paper we obtained PWMs of the $IPLP$ distribution since they can be used for estimating the $IPLP$ parameters. For a random variable with the pdf $f(.)$ and cdf $F(.)$, the PWMs function can be obtained as follows:

$$\Gamma_{p,r}=E[X^{p}F(X)^{r}]=\int_{-\infty}^{\infty}x^{p}\big(F(x)\big)^{r}f(x)dx \tag{29}$$

Putting (7) and (8) in (29), followed by algebraic manipulation, we have

$$\Gamma_{p,r}=\frac{\alpha\rho\zeta}{\lambda(-e^{-\zeta}+1)^{1+r}}\sum_{i,j=0}^{\infty}\frac{(-1)^{r+j}}{j!}\binom{r}{i}(1+r)^{j}\int_{-\infty}^{\infty}x^{\rho-\alpha-1}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-[\rho(i+j)+1]} \tag{30}$$

$$=\frac{\rho\zeta\lambda^{-p/\alpha}}{(-e^{-\zeta}+1)^{1+r}}\sum_{i,j=0}^{\infty}\frac{(-1)^{r+j}}{j!}\binom{r}{i}(1+r)^{j}\zeta^{j}B\left[(1-p/\alpha),\frac{p+\alpha[\rho(1+j)-1]}{\alpha}+1\right]$$

3.5. Order statistics

In real life experiment order statistics plays a very crucial and informative role in understanding the concepts of system reliability. Randomly selecting samples from $IPLP$ distribution and arranging them in increasing/decreasing other of magnitude, i.e. $(T_{1:n} < T_{2:n} < \cdots < T_{n:n})$, constitute an ordered sample which can be investigated as order statistics.

3.5.1 Derivation of the j^{th} order statistics

Consider $X_{(j:n)}$ denoting the j^{th} ordered sample from the $IPLP$ distribution given in (8). Then the Probability density for the j^{th} order statistics is

$$f_j\left(x_{(j)}, \Psi\right) = \frac{1}{B(j,n+j+1)}\left\{G\left(\left(x_{(j)}, \Psi\right)\right\}^{n-1} g\left(x_{(j)}, \Psi\right)\left\{1 - G\left(\left(x_{(j)}, \Psi\right)\right\}^{n-j} \tag{31}$$

where $\Psi = (\alpha, \rho, \zeta, \lambda)$. Further simplification using Taylor series expansion gives

$$f_j\left(x_{(j)}, \Psi\right) = \frac{1}{B(j,n+j+1)}\sum_{i=0}^{\infty}(-1)^i\binom{n-j}{i}\left\{G\left(\left(x_{(j)}, \Psi\right)\right\}^{n+i-1} g\left(x_{(j)}, \Psi\right) \tag{32}$$

Inserting (7) and (8) in (32) followed by further simplification using Taylor series, we have

$$f_j\left(x_{(j)}, \Psi\right) = \frac{\alpha\rho\zeta}{\lambda B(j, n + j + 1)}\sum_{i=0}^{n-j}\sum_{j=l=0}^{\infty}(-1)^{i+k+l}\binom{n-j}{i}\binom{i}{k}(k+1)^l\zeta^{l+1}$$
$$\times\frac{x^{-\alpha-1}\left(1+\frac{x^{-\alpha}}{\lambda}\right)^{-\rho(l+1)}}{(-e^{-\zeta}+1)^{i+1}i!} \tag{33}$$

4. Estimation

Let $\underline{x} = x_1, x_2, \dots, x_n$ represent a random sample of the $IPLP$ distribution with unknown parameter vector $\Psi = (\alpha, \lambda, \rho, \zeta)$. The log likelihood $l = l(\underline{x}, \Psi)$ for Ψ is

$$l(\underline{x}, \Psi) = log\left(\frac{\alpha\rho\zeta}{\lambda(-e^{-\zeta} + 1)}\right) - (\alpha + 1)\sum_{i=1}^{n} log(x_i) + (\rho + 1)\sum_{i=1}^{n} log\left(1 + \frac{x^{-\alpha}}{\lambda}\right)$$
$$\times -\zeta\sum_{i=1}^{n}\left(1 + \frac{x^{-\alpha}}{\lambda}\right)^{-\rho} \tag{34}$$

The score function $U(\Psi) = \left(\frac{\partial l}{\partial\alpha}, \frac{\partial l}{\partial\rho}, \frac{\partial l}{\partial\zeta}, \frac{\partial l}{\partial\lambda}\right)^T$ has components

$$\frac{\partial l}{\partial\alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} log(x_i)(\rho + 1)\sum_{i=1}^{n}\frac{x^{-\alpha}logx}{\lambda\left(1+\frac{x^{-\alpha}}{\lambda}\right)} + \zeta\sum_{i=1}^{n}\rho\left(1 + \frac{x^{-\alpha}}{\lambda}\right)^{-\rho} x^{-\alpha}logx \tag{35}$$

$$\frac{\partial l}{\partial\rho} = \frac{n}{\rho} + \sum_{i=1}^{n} log\left(1 + \frac{x^{-\alpha}}{\lambda}\right) + \zeta\sum_{i=1}^{n}\left(1 + \frac{x^{-\alpha}}{\lambda}\right)^{-\rho} log\left(1 + \frac{x^{-\alpha}}{\lambda}\right)^{-\rho} \tag{36}$$

$$\frac{\partial l}{\partial\zeta} = \frac{n}{\zeta} - \frac{n}{(-e^{-\zeta}+1)} - \sum_{i=1}^{n}\left(1 + \frac{x^{-\alpha}}{\lambda}\right)^{-\rho} \tag{37}$$

$$\frac{\partial l}{\partial\lambda} = -\frac{n}{\lambda} - (\rho + 1)\sum_{i=1}^{n}\frac{x^{-\alpha}}{\lambda^2\left(1+\frac{x^{-\alpha}}{\lambda}\right)} + \zeta\sum_{i=1}^{n}\frac{x^{-\alpha}}{\lambda^2\left(1+\frac{x^{-\alpha}}{\lambda}\right)} \tag{38}$$

The maximum likelihood estimate (MLE) $\widehat{\Psi}$ of Ψ is calculated numerically from the nonlinear equations $U(\Psi) = 0$. We use Adequacy Model in R to obtain $\widehat{\Psi}$.

4.1. Real data applications

In this section, we analyze two real data sets to demonstrate the flexibility and applicability of the proposed $IPLP$ model. The first data set, representing strengths of 1.5 cm glass fibers, was previously studied by Smith and Naylor (1986). The second data set contain 40 times to failure of turbocharger of one type of engine and was previously studied by Al Sobhi (2022). The two data sets are carefully selected because they are negatively skewed and are either over- or under-dispersed. The $IPLP$ model is compared to the one of the following competitive models: Inverse Lomax Poisson (ILP), Inverse Power Lomax (IPL), and Inverse Lomax models. In order to have a fair model comparison, we also use the following measures of goodness-of-fit criteria: Cramér Von-Mises ($CVMS$), Anderson-Darling (ADS), Kolmogorov-Smirnov (KSM), as well as those based on the log-likelihood: minus estimated -2*log-likelihood (-2$\hat{l}$), Akaike information criterion ($AICr$), consistent Akaike information criterion ($CAICr$). The model with the minimum values for $CVMS$, ADS, KSM, $AICr$, and $CAICr$ is considered to provide the best reasonable fits for the proposed data. Table 2 shows the exploratory data analysis for the two data sets which indicates that data I consist of 63 observations, negatively skewed, over-dispersed, with excess kurtosis of 0.92 that is leptokurtic. Data set II consist of 40 observations, under-dispersed, with excess kurtosis of -0.56 that is mesokurtic. Tables 3 and 5 gives the estimate of the parameters of the distributions considered.

Table 2. Exploratory data analysis of the two data sets

Specification	n	$Min.$	q_1	$Median$	q_3	$mean$	$Max.$	$Var.$	$Kurt.$	$Skew.$
$Data\ I$	63	0.55	1.38	1.59	1.69	1.51	2.24	0.11	3.92	-0.90
$Data\ II$	40	1.60	4.95	6.40	7.83	6.17	9.0	3.93	2.44	-0.55

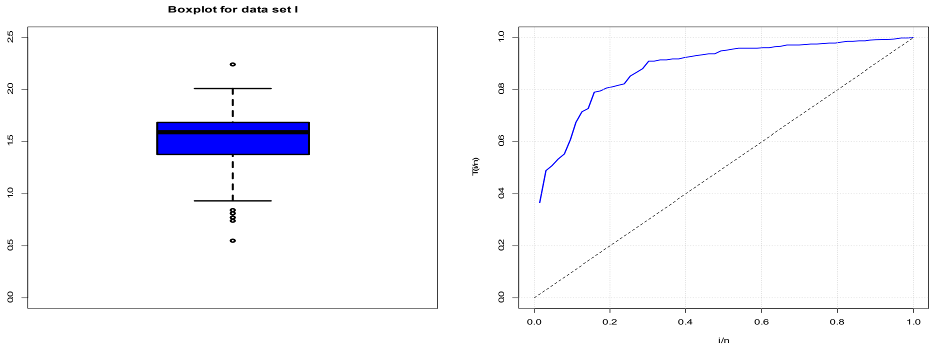


Figure 4. Box plot and the Total time on Test (TTT) plot for data set I

Figure 4 indicate that data set I is negatively skewed exhibiting an increasing failure rate.

Table 3. MLEs, their standard error (in parenthesis), confidence interval (curly) bracket for data set I

Model	α	λ	ρ	ζ

$IPLP$	12.3939(0.7202)
	{10.9823,13.8055}
		0.0025(0.0007)
		{0.0011,0.0039}
			−2.7082(1.329)
			{−5.3130, −1033}
				0.2541(0.1010)
				{0.0561,0.4521}

ILP	−
	(−)
		0.3638(0.1939)
		{−0.0162,0.7438}
			−2.2028(2.0980)
			{−6.3149,1.9093}
				0.6170(0.3486)
				{−0.0663,1.3003}

IPL	11.5729(0.5958)
	{10.4051,12,7407}
		0.0021(0.0004)
		{0.0013,0.0029}
			0.4419(0.0615)
			{0.3214,0.5624}
				−
				(−)

IL	−
	(−)
		10.7375(5.8660)
		{−0.7599,22.2349}
			15.5166(8.3418)
			{−0.8333,31.8665}
				−
				(−)

Table 4. Measures of goodness-of-fit value for data set I

Model	$-2l$	$AICr$	$BICr$	$HQICr$	$CAICr$	KSN	ADS	$CVMS$	PV
$IPLP$	25.46	33.46	42.029	36.83	34.15	0.1188	0.8358	0.1521	0.3358
ILP	57.42	63.42	69.851	65.95	63.83	0.2364	3.2897	0.6001	0.0017
IPL	30.38	36.39	42.816	38.83	36.79	0.1643	1.3963	0.2534	0.0666
IL	182.48	186.48	190.76	188.16	186.68	0.4889	4.5411	0.8360	1.7e-13

From Table 4 it can be observed that the new developed inverse Power Lomax Poisson model has better fit than other three notable competitive models because it possessed the smallest value of the $AICr$, $CAICr$, $BICr$, $HQICr$, KSM, ADS and $CVMS$ as well as largest PV value in modeling the glass fiber data.

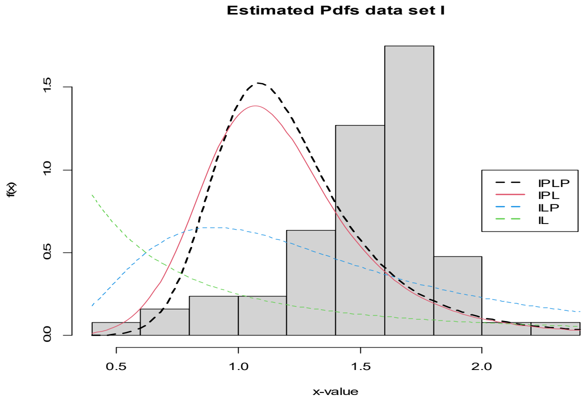


Figure 5. Graph of the fitted density for data set I

Figure 5 clearly indicates that $IPLP$ model provides a better fit than all other models considered in the study.

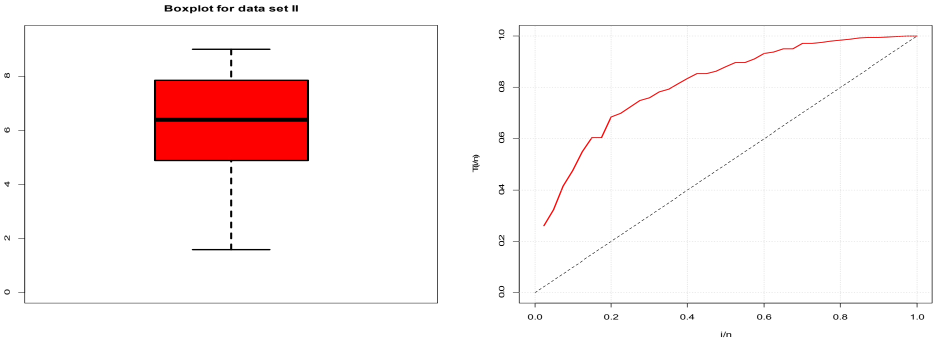


Figure 6. Box plot and the Total time on Test (TTT) plot for data set II

Figure 6 indicates that data set II is negatively skewed without any form of outlier exhibiting an increasing failure rate.

Table 5. MLEs, their standard error (in parenthesis), confidence interval (curly) bracket for data set II

Model	α	λ	ρ	ζ

$IPLP$	1.8844(0.2453)
	{1.6391,2.3652}
		0.0077(0.0040)
		{−0.0001,0.0117}
			29.9580(20.5404)
			{−10.3011,50.4984}
				2.2498(0.3983)
				{1.4691,2.6481}

ILP	−
	(−)
		10.4091(1.5213)
		{7.4274,11.9304}
			−9.8683(0.3210)
			{−10.4975, −9.2391}
				5.6043(1.2456)
				{3.1629,8.0457}

IPL	3.5615(0.2042)
	{3.1613,3.5615}
		0.0025(0.0006)
		{0.0013,0.0031}
			−
			(−)
				−
				(−)

IL	−
	(−)
		7.4952(8.9801)
		{−10.1058,25.0962}
			39.5313(6.7337)
			{26.3333,52.7294}
				−
				(−)

Table 6. Measures of goodness-of-fit value for data set II

Model	$-2l$	$AICr$	$BICr$	$CAICr$	$HQICr$	KSM	ADS	$CVMS$	PV
$IPLP$	170.08	178.08	184.83	179.22	180.52	0.1038	0.7661	0.1069	0.7817
ILP	231.62	237.62	242.69	238.29	239.56	0.4346	2.0806	0.3355	5.5e-07
IPL	182.12	188.12	193.19	188.79	189.96	0.1756	1.5482	0.2406	0.1698
IL	288.66	232.67	236.04	232.99	233.89	0.4411	2.2521	0.3673	3.5e-07

From Table 6 it can be observed that the new developed Inverse Power Lomax Poisson model has better fit than other three notable competitive models because it possesses the smallest value of $AICr$, $CAICr$, $BICr$, $HQICr$, KSM, ADS and $CVMS$ as well as the largest PV value in modeling the turbocharger data.

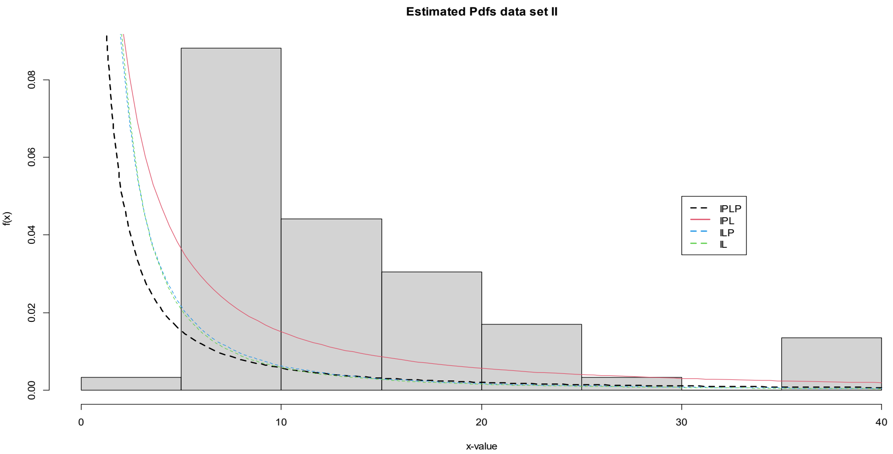


Figure 7. Graph of the fitted density for data set II

Figure 7 clearly indicates that $IPLP$ model provides a better fit than all other models considered in the study.

Table 7. LR test for the two data sets

Model	Hypothesis	LR	$P - value$
	Data set I		
$IPLP\ vs\ ILP$	$H_0: \alpha = 1\ vs. H_1: H_0\ is\ false$	31.96	< 0.001
$IPLP\ vs\ IPL$	$H_0: \zeta = 1\ vs. H_1: H_0\ is\ false$	4.92	< 0.00
$IPLP\ vs\ \ IL$	$H_0: \alpha = \zeta = 1\ vs. H_1: H_0\ is\ false$	157.02	< 0.001
	Data set II		
$IPLP\ vs\ \ ILP$	$H_0: \alpha = 1\ vs. H_1: H_0\ is\ false$	61.54	< 0.001
$IPLP\ vs\ IPL$	$H_0: \zeta = 1\ vs. H_1: H_0\ is\ false$	12.04	< 0.001
$IPLP\ vs\ \ IL$	$H_0: \alpha = \zeta = 1\ vs. H_1: H_0\ is\ false$	58.58	< 0.001

It can be observed from Table 7 that in each of the cases considered we accept the alternative hypothesis which is enough evidence that the $IPLP$ model has a better fit than all other models considered for the two data sets and can effectively be used for fitting the data.

5. Concluding remarks

We have developed and studied the $IPLP$ distribution along with its properties such as moments, incomplete moments, weighted moments, moment generating functions, Rènyi and q entropies, Bonferroni and Lorenz curves, reliability studies, stress-strength reliability and multi component stress-strength reliability model.

Maximum likelihood estimates are computed. Goodness-of-fit shows that the *IPLP* distribution is a better fit. Applications of the *IPLP* model to glass fiber and turbocharger data are presented to demonstrate its greater significance and better flexibility. We have shown that the *IPLP* distribution empirically provides reasonable fit for both the glass fiber and turbocharger data as supported by the graph of fitted densities and the likelihood ratio test statistics. In view of the shapes of the density and failure rate function, it can be concluded that the proposed model is a suitable candidate model in reliability analysis, data modeling, and other related fields. For future study, bivariate extension of the Inverse Power Lomax Poisson model can be considered.

Acknowledgement

We thank all the anonymous reviewers who have contributed to the final production of this article.

References

Abd-Elfattah, A. M, Hassan, A. S. and Hussein, A. M., (2013). On the Lomax-Poisson Distribution. Proceeding of the 48th The Annual Conference On Statistics, Computer Sciences & Operations Research, Institute of Statistical Studies Research. *Cairo University*, pp. 25–39.

Abdul-Moniem, I. B., Abdel-Hameed, H. F., (2012). A lifetime distribution with decreasing failure rate. *International Journal of Mathematical Education*, 33(5), pp. 1–7.

Al-Marzouki, S., Jamal, F., Chesneau, C. and Elgarhy, M., (2020). *Type II Topp Leone Power Lomax Distribution with Applications. Mathematics 2020*, 8, 4, doi: 10.3390/math8010004.

Al Sobhi, M. M., (2022). The extended Weibull distribution with its properties, estimation and modeling skewed data. *Journal of King Saud University – Science*, 34(2), pp. 1–15.

Al-Zahrani, B., (2015). An extended Poisson-Lomax distribution. *Advances in Mathematics: Scientific Journal*, 4(2), pp. 79–89.

Al-Zahrani, B., Sagor, H., (2014). Statistical analysis of the Lomax-Logarithmic distribution. *Journal of Statistical Computation and Simulation*, 85, pp. 1883–1901.

Cordeiro, G. M., Ortega. E. and Popović, B., (2013). The gamma-Lomax distribution. *Journal of Statistical Computation and Simulation*, 85(2), pp. 305–319.

Haq, M. A., Hamedani, G. G., Elgarhy, M. and Ramos, P. L., (2020). Marshall–Olkin power Lomax Distribution: Properties and estimation based on complete and censored samples. *Int. J. Stat. Probab.*, 9(1), p. 48.

Hassan, A. S.; Abd-Allah, M., (2019). On the Inverse Power Lomax Distribution. *Ann. Data Sci.*, 6, pp. 259–278, doi:10.1007/s40745-018-0183-y.

Hassan, A. S., Nassr, S. G., (2018). Power Lomax Poisson distribution: Properties and Estimation. *Journal of Data Science*, 18, pp. 105–128.

Lemonte, A. J., Cordeiro G. M., (2013). An extended Lomax distribution. *Statistics*, 47(4), pp. 800–816.

Nagarjuma, V. B. V., Vardhan, R. V. and Chesnau, C., (2022). Nadarajah Haghighi-Lomax distribution and is Applications. *Math. Comput. Appl.,* 27 (30), pp. 1–13.

Ramos, M. W. A., Marinho, P. R. D., da Silva R. V. and Cordeiro, G. M., (2013). The exponentiated Lomax Poisson distribution with an application to lifetime data. *Advances and Applications in Statistics*, 34(2), pp. 107–135.

Roman, M., Louzada, F., Cancho, V. G. and Leite, J. G., (2012). A new long-term survival distribution for cancer data. *Journal of Data Science*, 10(2), pp. 241–258.

Simth, R. L., Naylor, J. C., (1987). A comparison of maximum likelihood and Bayesian estimators for three-parameter Weibull distribution. *Applied Statistics*, 36, pp. 358–369.

Tahir, M. H., Cordeiro, G. M., Mansoor, M. and Zubair, M., (2015). The Weibull-Lomax distribution: Properties and applications. *Hacettepe Journal of Mathematics and Statistics*, 44(2), pp. 461–480.

Tahir, M. H., Hussain, M. A., Cordeiro, G. M., Hamedani, G. G., Mansoor, M. and Zubair, M., (2016). The Gumbel-Lomax distribution: Properties and applications. *Journal of Statistical Theory and Applications*, 15(1), pp. 61–79.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 127–150, https://doi.org/10.59139/stattrans-2025-043
Received – 07.05.2024; accepted – 03.04.2025

Exploring variation in data on income inequality across databases and measures in post-socialist countries

Monika Wesołowska[1]

Abstract

Despite the growing interest in income inequality, cross-country evidence often shows variation between measures and databases, which complicates research and policy evaluation. The objective of the article is to compere the consistency of data on income inequality in post-socialist countries from Central and Eastern Europe and Central Asia for the commonly used measures on the basis of leading databases in this area. Other such analyses typically focus on individual measures, databases or specific countries, which prompted the idea to fill the research gap for a targeted country group. The formulated hypotheses were to test the consistency of the following: development trends, the rankings of countries from the most to the least equal in terms of income, and the values for different measures indicated by databases.
The study reveals high correlations in income inequality trends over the long term, particularly among the EU subgroups. Certain consistency was observed in the context of identifying countries with extreme income equality or inequality, and in the rankings between different measures from the same database. However, there was no full consistency, especially in non-EU countries, which highlights the impact of the methodological differences.
This article contributes to the existing body of research on income inequality by providing a broad analysis of the consistency and variability of the related data across different measures and databases, with a particular focus on post-socialist countries. It points to the importance of careful data selection when analyzing income inequality in the indicated group of countries, as individual differences between measures, databases and countries tend to affect the final results of the research.

Key words: income inequality, post-socialist countries, statistical analysis.

1. Introduction

Measuring income inequality is a comprehensive task that involves complex and multifaceted decisions. The process begins with selecting data collection methods and defining how to process and interpret the values. Key decisions also include aspects such as replacing missing responses, choosing the right measure, or interpreting

[1] Poznań University of Economics and Business, Poznań, Poland. E-mail: monika.wesolowska@ue.poznan.pl. ORCID: https://orcid.org/0000-0002-1438-1962.

findings in a social and economic context. In addition, inequality is measured at specific points in time, not providing a complete picture of its evolution. Even the most precise measures capture inequality only within a population at a given time, without reflecting changes in individual income or wealth over time (Pascola, Rucha, 2017).

The aim of the article is to examine the variation and consistency in changes in income inequality over time in post-socialist countries of Central and Eastern Europe and Central Asia, using leading databases and common measures that differ in their data collection methods, directly affecting the obtained results. Since the transformation from the socialist system in 1989, there were significant changes affecting income distribution in all studied countries, but with varying intensity among them (Milanovic, 1998; Brzezinski, Salach, 2022). A group of post-socialist countries began their economic transition with low levels of inequality and relatively small disparities between them. Today, however, the variation in income inequality within this group is substantial, ranging from low to high polarization.

The transformations of the past three decades pose greater challenges for measuring income polarization than in countries without such systemic shifts. However, studies on variation and homogeneity of income inequality mostly focus on Western European countries rather than on the post-socialist group. The study presented in this article focuses on a broad comparison of development trends, levels and rankings of income inequality using data for the Gini coefficient, income shares of individual deciles, Atkinson index, and Palma ratio from eight databases such as World Inequality Database, Standardized World Income Inequality Database, Luxembourg Income Study, OECD, World Income Inequality Database, World Development Indicators, Eurostat, and Global Consumption and Income Project The results may contribute to international analyses of income polarization within this group.

Section 2 discusses methodological issues related to inequality measurement and differences between databases and measures, followed by an analysis of inequality evolution since the 1990s. Section 5 examines trend stability, ranking consistency, and value variation, while the final section presents conclusions.

2. Methodological issues of measuring inequality

To provide a detailed introduction to the problem under study, this section presents the differences between the measures, databases, and data collection methods, as well as an overview of empirical studies focused on these three dimensions. Each of these components can affect the final values of income distribution equality differently, leading to a more or less accurate representation of reality. The literature reports cross-country variation in inequality levels, differences in data smoothing across sources, and partial consistency in long-term trends or country rankings.

2.1. Differences depending on methods of data collection

The choice of data collection method significantly affects inequality estimates and their reflection of reality. Unfortunately, various methodological problems are associated with different methods. The literature distinguishes three main approaches: survey-based, fiscal, and mixed methods.

Standardized questionnaires are a common quantitative research method, valued for structure and cross-respondent comparability. However, this method faces challenges of nonresponse and underreporting, which may distort estimates, particularly for high-income households (Vermeulen, 2016). Refusals to participate in surveys can further skew the representation of the surveyed population, although this approach ensures frequent data collection. Despite these limitations, surveys excel in representing the incomes of lower-income individuals or households but may not fully capture the impact of high earners on overall income inequality dynamics (Larrimore, Burkhauser, Armour, 2018). Moreover, respondent errors in reporting income introduce inaccuracies, potentially blurring the true income distribution, especially if not uniformly distributed across respondents. Therefore, while survey-based methods offer valuable insights, their integration with other data sources and rigorous statistical techniques is essential for a comprehensive understanding of income inequality dynamics.

Top incomes, although representing a small part of the population, contribute significantly to total income and tax revenues, making them crucial for inequality indicators (Alvaredo, 2011; Atkinson, Piketty, Saez, 2011; Blanchet et al., 2018). Tax data serve as the primary source for capturing this income level, free from survey participation biases. However, this method faces drawbacks, such as limited comparability over time due to legislative changes and across countries due to tax-system differences (Atkinson, Piketty, Saez, 2011). Additional challenges include tax evasion, underreporting, and omission of income sources such as transfers, informal earnings, or agriculture (Bukowski, Novokmet, 2017). These factors risk overestimating income inequality and underrepresenting lower earners in the analysis.

The strengths and weaknesses of survey-based and tax-based approaches complement each other. Surveys capture poorer households well but tend to underestimate inequality. Conversely, tax data accurately depict top incomes but may exaggerate inequality levels. Mixed methods, such as the UK's 'SPI adjustment' (Larrimore, Burkhauser, Armour, 2018) and the WID approach combining survey data for lower incomes (below the 0.90th percentile) with tax data for the top ones (above the 0.99th percentile) (Alvaredo et al., 2016), aim to reconcile these disparities. These methods reduce under- and overestimation, resulting in a more accurate picture of reality. Such approaches contribute to providing comprehensive insights into income distribution dynamics across different income percentiles.

In summary, the selection of data collection methods strongly impacts measurements of inequality and introduces methodological hurdles. Surveys provide structured data but suffer from downward bias, particularly for higher incomes. Fiscal data captures high earners but overlooks certain sources and lacks international uniformity. Mixed methods aim to counter these shortcomings by merging survey and tax data. Implementation challenges include limited access to fiscal data and methodological complexity. Collaborative endeavors are vital to refining methodologies and maximizing data utility for a comprehensive analysis of inequality. All this can lead to differences of several p.p. between data based on various data collection methods at specific points in time and even show different development trends over time.

2.2. Differences depending on selected measures of income inequality

The perception of income inequality depends not only on data collection methods but also on the choice of inequality measures. The choice of a particular measure of income inequality can significantly change the perceived level of it, and even, through methodological differences, indicate differential development trends, even though they are often really similar between measures (The Equality Trust, 2011). One of the reasons for the varying empirical results is being sensitive to different parts of the income distribution (De Maio, 2007).

The Gini coefficient (Farris, 2010) is a widely used measure of income inequality, calculated as the average income difference between all pairs in a population divided by twice the mean income. It ranges from 0 (perfect equality) to 1 (perfect inequality). While intuitive and easy to visualize, it is most sensitive to changes in the middle of the distribution and cannot be decomposed analytically (Solt, 2020). Moreover, identical Gini values may correspond to different income distributions, and the measure ignores demographic shifts or income mobility, which has raised methodological concerns (Piketty, 2014; Corak, 2013).

The Atkinson index (Atkinson, 1970) provides a broader perspective by incorporating social preferences for equality through a welfare function. Its value depends on the inequality-aversion parameter ε (Dubois, 2016; Latty, 2015), which weights disparities at different income levels. Unlike the Gini coefficient, it is sensitive to changes across the entire income distribution (De Maio, 2007). However, the index's subjectivity complicates cross-study comparisons when different ε parameters are applied. Despite being decomposable (Bellu and Liberati, 2006), it remains less commonly used than the Gini coefficient.

Income shares across quartiles, deciles, or percentiles provide valuable insights into distributional dynamics (Jędrzejczak, Pekasiewicz, 2018). While aggregate measures offer a general picture, examining individual distribution segments, especially pre- and post-transfer data, allows deeper analysis of who benefits from policy or economic change (Eurostat, 2020; Voitchovsky, 2005; Sitthiyot, Holasut, 2020). On this basis, positional indicators such as the Palma ratio compare the income share of the richest decile with that of the poorest 40% (Cobham, Schlögl, Sumner, 2016), focusing on the distribution tails, assuming stability in middle deciles (Cobham, Summer, 2013).

Income inequality analysis involves various measures that present slightly different calculations of the level of inequality in the income distribution, with individual sets of both the advantages and disadvantages. While the Gini index is popular and simple to interpret, it can take exactly the same values with widely varying income distributions skewing the final picture of inequality. The Atkinson Index provides a unique perspective based on social preferences, but this can be a problem when drawing conclusions. Data on average income and social group shares help to understand the dynamics of specific segments of society, but do not provide a clear, straightforward answer about the level of inequality in society as a whole. In contrast, using the Palma ratio, targeting the extremes of the distribution can better respond to changes in key areas of inequality, but it ignores 50% of income distribution. Therefore, a full understanding of income inequality can require a combination of different measures to capture the multifaceted nature of this complex phenomenon.

2.3. Differences between databases

The third factor affecting inequality estimates is the choice of database and its underlying methodology**.** Many databases are secondary sources but differ in interpolation, data types, and the treatment of missing observations or zero incomes. All these aspects can affect the outcomes, even if the original, primary-source dataset was identical among few databases. The most important databases include Luxembourg Income Study, Eurostat, Global Consumption and Income Project, Standardized World Income Inequality Database, World Income Inequality Database, OECD – Income Distribution Database, World Development Indicators, and World Inequality Database.

Luxembourg Income Study (LIS) provides harmonised, survey-based microdata collected by national statistical agencies under common protocols. These data are fully based on surveys and are kept only for individual years without interpolation, unlike most of the databases, where the coverage covers even several dozen years for individual countries. Its great advantage is the full methodological consistency between the countries surveyed, from the level of the questionnaire, which is compiled to make it easy to understand for the recipients, to the way it is harmonized. However, LIS remains vulnerable to top-income undercoverage and non-response (Ravallion, 2015).

Eurostat provides primary survey data from the European Union Statistics on Income and Living Conditions (EU-SILC) survey, which collects data on income, poverty, social exclusion, and living conditions of households and individuals across the EU. Data collection is outsourced to statistical offices in individual member states, which tune into the methodology adopted by Eurostat. EU-SILC is not inequality-specific; it primarily reports income (including quintiles) and poverty indicators. Eurostat also provides the so-called "experimental" data calculated as part of the Income, Consumption and Wealth (ICW) statistics, which are computed through the statistical matching of three data sources: the EU Statistics on Income and Living Conditions (EU-SILC), the Household Budget Survey (HBS) and the Household Finance and Consumption Survey (HFCS). Another database based on primary household survey data obtained from government statistical agencies and World Bank country departments is World Development Indicator (WDI), but for high-income economies data are incorporated mostly from the LIS database. Regardless of the source, the data used are subject to a uniform estimation method.

In terms of secondary source databases, the dataset by Lahoti et al. (2016) - the Global Consumption and Income Project (GCIP) - is another example based mostly on survey data, though not exclusively. Its survey component compiles data from multiple sources, mainly other databases focused on international comparisons, but also from national statistical offices and academic studies on individual countries, creating a large and diverse dataset built on a homogeneous methodology. The aim is comprehensive, integrated coverage that mitigates source-specific errors, though survey-method limitations remain. SWIID (Solt, 2009) also combines multiple sources and includes more government-provided and fiscally-based inputs. However, the main difference between the two databases lies in their approach to data standardization. Both use econometric estimations, but SWIID follows the LIS methodology as its gold standard (Solt, 2020), whereas GCIP applies its own quintile-specific consumption-income ratio method and additionally interpolates missing data. The World Income Inequality Database (WIID), compiled by WIDER, aggregates income data from numerous sources and studies into a single accessible dataset but does not standardize them, which distinguishes it from the two previous databases (UNU-WIDER, 2022). Similarly, the OECD's Income Distribution Database (IDD) combines multiple sources, mainly surveys, with occasional tax data. However, results may differ due to its correction procedure, which adjusts all household income components by the square root of household size (OECD, 2017; OECD, 2023).

World Inequality Database (WID) takes a completely different approach than previous databases. When it is possible for some fiscal data to appear in other databases, they are not subject to special treatment, at most averaged with the rest of the data. WID, on the other hand, assumes that survey data accurately reflect the income of the lowest part of society but underestimate the highest income, quite the opposite of fiscal data, therefore, it combines these data in an appropriate way to obtain more balanced results (Alvaredo, Atkinson, Chancel, Piketty, Saez, Zucman, 2016).

Among the selected databases, the majority relies on survey data, often combined with additional data from national offices. However, substantial differences between them can significantly impact the obtained results. An exception is WID, which innovatively employs fiscal and survey data. Notably absent are purely fiscal data sources, due to variations in their availability among countries and possible discrepancies between providers of such data.

2.4. Empirical studies on the variation of income inequality depending on the method of measurement, measure and database

Based on the presented data collection methods, inequality measures, and databases, clear differences emerge that can lead to variation in estimated inequality levels. Empirical studies show that estimates from administrative tax data often differ markedly from survey-based results, both in levels and trends over time. Moreover, methodological factors, including the choice of measure or database, can produce similar discrepancies due to differences in harmonisation and data interpolation.

Studies comparing survey and tax-adjusted data show differences of several percentage points (p.p.) in Gini coefficients, while maintaining consistent long-term trends. Jenkins (2017) indicates that the Gini coefficient for gross individual income in the UK estimated from tax data rose by 7–8%, whereas survey data showed a 5% decline over the same period. According to Bartels and Metzing (2019), the income shares of the top 1% in Germany were higher in the tax data than in the surveys by 3–6 p.p., but the estimates of the income share of the top 10–5% and top 5–1% are of similar magnitude in both data sources. Their research also indicates the relevance of the choice of data pooling method, as their integrated approach indicated slightly lower levels of income inequality than the decomposition method (Alvaredo, 2011). Similar gaps were found elsewhere: about 6 p.p. in the US (Burkhauser et al., 2012), 12–14 p.p. in Russia (Novokmet et al., 2018), and 14% in Spain (Ayala, Perez, and Prieto-Alaiz, 2021). In Poland (1994–2015), Brzezinski, Myck and Najsztub (2022) found that adjusting survey data with fiscal data on top incomes increased Gini values by 14–26% (4–8 p.p.) relative to unadjusted estimates. The authors also show that the adjustment changed the development trend of Ginis, with a sharp change in the level of inequality of income distribution, which was invisible in survey-only data.

In terms of databases, Bartels and Metzing (2019) comparing nine countries using EU-SILC and WID, found that differences are minor for some countries but reach up to 9% for others. However, differences can also arise with similar methods. Similar conclusions are also indicated by analyses focused on other databases such as LIS, OECD, EU-SILC, and WDI (Galbraith et al., 2016). Jenkins (2015), based on two secondary databases, SWIID and WIID, shows that differing data implementation can distort inequality estimates, with SWIID tending to over-smooth results. Ferrerira, Lustig and Teles (2015), comparing eight databases (including LIS, WIID, SWIID, IDD, among others), found a high degree of consistency in long-term trends across most countries. However, for specific country–year observations, methodological differences cause substantial discrepancies, sometimes leading to divergent conclusions depending on the chosen dataset. These discrepancies concern not only inequality levels but sometimes even the direction of year-to-year changes. In a similar comparison, Galbraith et al. (2016) also presented the overall consistency with the occurrence of large differences in specific countries, and added indications of significant deviations from other databases compared to data from the WDI.

Moreover, in terms of the measures, according to Trapeznikova (2019), research shows general agreement on trends and rankings of countries in terms of levels of income inequality, although the author notes the importance of including measures sensitive to changes in marginal income in order to arrive at more precise conclusions. However, Goda (2016) argues that due to methodological issues, the choice of a particular measure can indicate divergent development trends, even if they are often similar between the measures.

The literature has examined how the above aspects affect income distribution equality. Researchers note fluctuations between countries, varying degrees of data smoothing depending on the source, and inconsistent long-term trends or rankings across measures. However, most studies focus on individual countries or groups of countries from Western Europe or the US. Post-socialist countries have experienced some of the largest observed changes in income inequality in recent decades, with significant changes occurring both in individual countries and the group as a whole. However, research in this area has mainly focused on a few examples like Poland, Russia, or the Czech Republic, leaving a gap in studies analyzing the group as a whole. This article aims to address that gap.

Based on the analyzed studies, three research hypotheses were formulated. First, the analysis will examine whether post-socialist countries exhibit long-term trend consistency across different measures and databases. Second, the study will assess the consistency of country rankings within the group at specific points in time. While measures and databases may show similar inequality trends for a given country, rankings can still differ in identifying which countries are most or least equal at a given time. The differential ranking may be the result of significant differences in the measurement of values, for this reason, consistency of the measurement of inequality levels over the entire study period will also be tested. The research hypotheses are:

H1: A cohesive pattern is evident in the evolution of income inequality trends among post-socialist countries, irrespective of the measures and databases used;

H2: The classification of post-socialist countries based on income inequality reveals a consistent homogenization within the same measures from different database;

H3: Values for the same inequality measures exhibit consistent stability and limited variation across different databases in post-socialist countries.

3. Data selection and analytical strategy

This analysis examines the consistency of income inequality levels and trends over 30 years across selected measures and databases. The study includes eight major sources: World Inequality Database, Standardized World Income Inequality Database, Luxembourg Income Study, OECD - Income Distribution Database, World Income Inequality Database, World Development Indicators, Eurostat, and Global Consumption and Income Project. These databases, both primary and secondary, differ in data collection, processing, and implementation. The selected inequality measures - Gini coefficient, income shares by decile, Atkinson index, and Palma ratio - capture both overall inequality and distributional segments. More details on these databases and measures were discussed in the previous section. Table 1 outlines the selected measures from each source, with variations due to data availability.

Table 1. Summary of variable and database selection

Database	Measure
World Inequality Database	Gini coefficient
	Income shares of individual deciles
	Palma ratio
Global Consumption and Income Project	Gini coefficient
	Atkinson index
	Palma ratio
	Income shares of individual deciles
Standardized World Income Inequality Database	Gini coefficient
Luxembourg Income Study	Gini coefficient
	Atkinson index
OECD - Income Distribution Database	Gini coefficient
	Palma ratio
World Income Inequality Database	Gini coefficient
	Palma ratio
	Atkinson index
	Income shares of individual deciles
World Development Indicators	Gini coefficient
Eurostat	Gini coefficient (EU SILC)
	Gini coefficient (EU SILC - experimental)

Source: own compilation.

To confirm each of the three hypotheses, the following analyses were conducted. For the first hypothesis (H1), which posits cohesive patterns in inequality trends, Pearson correlation coefficients were calculated to assess their consistency over time. These correlations were examined both between databases for the same measure (cross-source consistency) and between measures within the same database (internal consistency). To confirm H2, which concerns the consistency of country classifications based on income inequality, country rankings were created for each measure and compared across datasets. The rankings, ordered from most equal to most unequal, were generated separately for each measure and analyzed across selected years to assess whether countries maintained similar rankings within a given year. For the third hypothesis (H3), concerning the stability and limited variation of inequality, we statistically analyzed variation in inequality levels across measures. Detailed results are presented in Section 5.

4. Data on income inequality in post-socialist countries

Income inequality trends across countries, based on the previously discussed measures and databases, indicate an initial rise in inequality until around 1995 or 2000, varying by country. Future EU members generally experienced a milder and shorter polarization phase than other post-socialist nations, followed by a period of relative stability. The time-series evidence confirms a broad consistency among different databases and inequality measures regarding long-term dynamics, which aligns with the findings of Ferrerira, Lustig, and Teles (2015). Nonetheless, individual observations reveal some notable discrepancies.

Starting with the Gini coefficient, the only measure analyzed in the study with data available from the Standardized World Income Inequality Database, the trend analysis confirmed the issue identified by Jenkins (2015) regarding excessive smoothing of data from this database over time. Consistent with previous studies (Vermeulen, 2016; Alvaredo, 2011; Atkinson et al., 2011; Blanchet et al., 2018), the World Inequality Database reports the highest Gini values across nearly all countries. While long-term trends align, year-on-year changes differ between databases, as noted by Ferrerira, Lustig, and Teles (2015). At the country level, the average year-to-year difference between databases was 11.83 points (on a scale of 0-100). The Czech Republic and Hungary showed the highest consistency, with only minor deviations, whereas Azerbaijan exhibited the largest discrepancy - exceeding 30 points in a single year - suggesting significant income polarization.

The data on the Palma ratio confirm conclusions similar to those drawn from the Gini coefficient. Once again, the World Inequality Database exhibits the widest spread in values, with several significant outliers. This issue is particularly evident in Lithuania, where between 2018 and 2019, the Palma ratio surged from 3.23 to an implausible 16.32—an error also reflected in the pre-tax data, indicating an almost 24-fold increase in one year. Due to this anomaly, observations for Lithuania had to be partially excluded from subsequent analyses to prevent data distortion. A similar problem with outlier observations appeared in the Global Consumption and Income Project, affecting Armenia and Kyrgyzstan, as well as in the World Income Inequality Database for Kyrgyzstan and Russia, though in these cases, the maximum change was 10.7 points.

The Atkinson index differs due to varying parameter ε settings across databases: 0.5 in the World Income Inequality Database, 1.5 in the Luxembourg Income Study, and unspecified in the Global Consumption and Income Project. According to Latty (2015) the difference significantly affects the level of the obtained WIID data points out values from 1.87 to 27.34, the LIS value does not exceed 0.2, and in the case of GCIP it is a range of 0.9-0.93. Hence, the data are comparable only in terms of rankings and correlations, not absolute levels. Notably, WIID and GCIP provide data spanning a much longer period than LIS and cover all countries, revealing recurring development patterns consistent with previous measures, as well as similar cases of outlier observations.

The data on income shares by decile particularly highlighted the earlier differences between the databases. For the total income of the poorer half of the population, the databases again converge on the direction of change and the overall development trend. Consistency is particularly evident between the World Inequality Database and the Global Consumption and Income Project, where income shares stabilized after a period of significant declines. However, the World Income Inequality Database indicates an additional partial increase in shares during the later period. Similar patterns are observed for the top 10% of income. All databases in this case show an increase in shares, with the largest changes occurring until around 1995, though the magnitude of these changes varies significantly between countries. This example also illustrates why WID shows the highest inequality among the measures, as its post-2000 values exceeds the highest values indicated by WIID, reflecting a much greater enrichment of the wealthy and impoverishment of the poorer population after 1989. This difference is particularly significant when examining the average inequality values over time, which are shown in Figure 1. WIID is the only database that shows an equalization of the incomes of the bottom 50% with the top 10%, followed by an increase in the incomes of the poorer 50%. These values also highlight the significant impact that the inclusion of non-survey data has on the final inequality measures. The WIID, based solely on survey data, indicated that on average during the period studied, the incomes of the richest 10% accounted for 0.95% of the incomes of the bottom 50%. The GCIP, which includes national accounts data, indicates a relationship of 122%, while the WID's inclusion of fiscal data results in a nearly doubled difference, at 199%. Given the significant methodological differences among the three databases, these discrepancies confirm De Maio's (2007) findings on differences in empirical results when targeting different parts of the income distribution, particularly in obtaining more precise data on the top decile. Moreover, the database containing fiscal data indicates greater fluctuations and more dynamic changes.

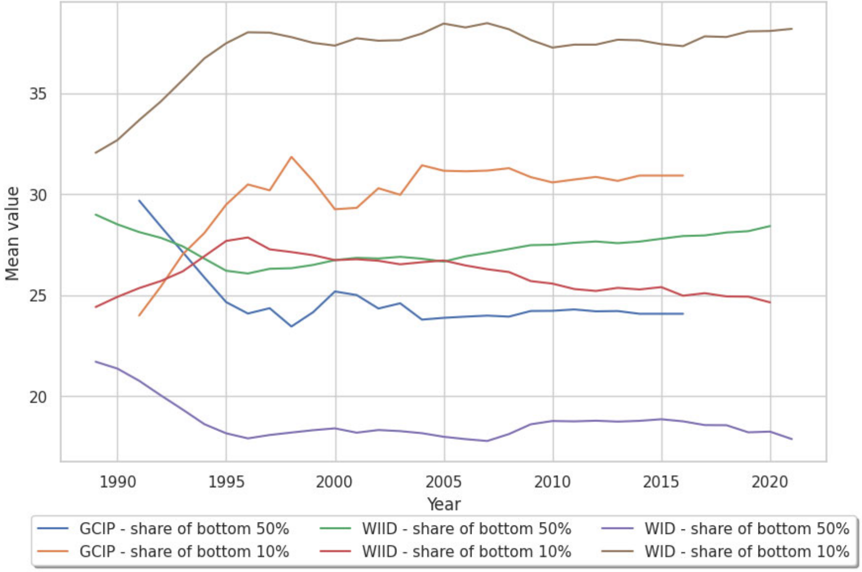


Figure 1. Average income shares of the 50% bottom earners and 10% top earners

Source: own compilation based on: Global Consumption and Income Project, World Income Inequality Database, World Inequality Database.

The data review confirms overall consistency in long-term trends but highlights inconsistencies in year-on-year changes, varying across countries. Outliers appear in each measure, with notable differences between EU and Central Asian countries. The analysis also reinforces concerns about data oversmoothing and the impact of survey-only vs. mixed data sources. This study further assesses how these inconsistencies influence income inequality research.

5. Findings

To properly conduct the study, the analysis was divided into three subsections aligned with the research hypotheses. The analysis begins with inequality trends previously observed in other country groups but not empirically verified for post-socialist countries. The focus then shifts to country rankings, moving from the level of individual countries to the entire group over time. The third stage involves value differentiation, which may not occur even if countries are ranked similarly and their development trends are perfectly correlated. Since differences in the data between EU and non-EU countries became apparent, the dataset was divided into subgroups for part of the analysis. In addition, because data availability for these subgroups is uneven, this division will provide more accurate results. Similarly, the issue of outlier observations and the inability to fully analyze variations in Atkinson index values due to differing ε parameters were addressed.

5.1. Consistency between the trends of income inequalities

The actual occurrence of consistency in income inequality trends was checked by analyzing correlation coefficients from available databases. This includes correlations within data for a single measure and between different measures, as methodological differences may result in more consistent data for specific metrics. The analysis covered the entire group of countries studied and two subgroups defined by membership in the European Union. Trend consistency analysis is particularly vulnerable to result distortion if there are unequal outcomes for any subgroup, as significantly higher (or lower) correlations in one group could skew the overall study results.

At the group level, correlation analysis of the Gini coefficient shows mostly high correspondences. The only low result (35%) was observed between Global Consumption and Income Project (GCIP) and World Development Indicators (WDI). World Inequality Database (WID) consistently exhibits lower correlations (0.63-0.83) with other datasets. This is expected as WID is the only database incorporating fiscal data which can inflate inequality estimates and capture fluctuations not visible in survey-based sources. All correlations are positive, confirming a consistent long-term trend, despite previously noted year-on-year inconsistencies. At the subgroup level, high correlations predominate among EU countries, where values exceed 0.9, reaching 0.98 in some cases, except for WID. Non-EU countries show much weaker or non-existent correlations, particularly in SWIID, WID, WIID, and GCIP, which provide the most data for this group.

Other measures show similar patterns. Palma ratios exhibit medium to high correlations across the entire group and EU countries, particularly after excluding Lithuania's 2019 outlier, which distorted results. Again, WID shows slightly lower correlations, but values remain high for EU countries (68–85%), compared to non-EU countries (37%–46%).

The correlation analysis for the Atkinson index leads to similar conclusions, with a correlation of 0.48 between GCIP and WIID (LIS provides single observations) for non-CEE countries. The last two measures examined—the income of the top 10% and bottom 50%—are also consistent with the above observations, showing a negative correlation as expected, since an increase in one measure should correspond to a decrease in the other. The lowest correlations were observed for WID, with -65% for EU countries and -22% for non-EU subgroups compared to WIID and GCIP. This highlights the impact of fiscal data on inequality estimates and the challenges of measuring inequality in post-Soviet countries. Methodological differences between databases, particularly the inclusion of fiscal data, result in higher recorded inequality levels (see Section 2) and distort comparisons with survey-based sources, which face their own methodological biases. Therefore, lower correlations reflect methodological rather than accuracy differences between fiscal and survey-based data.

Correlations occur not only within a single measure but also between different measures. Despite concerns that focusing on different parts of the income distribution could alter trends (Goda, 2016; De Maio, 2007), correlations remained high and positive between the Gini coefficient, Palma ratio, Atkinson index, and top 10% income shares. Likewise, high negative correlations were observed between bottom 50% income shares and the other indicators, as expected. Importantly, strong correlations were found not only within the same database (up to 99%, indicating internal consistency) but also across different databases measuring different inequality metrics. This suggests that regardless of the inequality measure chosen, the data indicate the same trend, minimizing the impact of methodological differences. For EU countries, the choice of income inequality measure does not affect the overall trend, as the data consistently reflect the same direction of change. In contrast, for non-EU countries, selecting the appropriate database and methodology is more crucial than the specific measure when analyzing long-term national trends.

In conclusion, for EU countries, high or very high correlations exist regardless of the measure or data source, with a consistent trend direction. This indicates stable relationships between inequality trends and long-term data consistency. The slightly lower correlations for WID, due to its inclusion of fiscal data, underscore the value of incorporating such data to capture inequality trends overlooked by survey-based sources, as observed in Poland (Bukowski & Novokmet, 2019) and Russia (Neef, 2020). In contrast, post-Soviet, non-EU countries show significant inconsistencies, with some datasets lacking any measurable correlation. The absence of fiscal data in these countries further complicates analysis, making dataset and measure selection crucial, as choices can substantially impact results. Finally, the apparent consistency of income inequality trends across measures and databases was largely driven by EU countries, masking methodological disparities and data availability issues. The inclusion of fiscal data in WID, contrasted with SWIID's excessive smoothing, further distorts inequality estimates, creating a misleading picture of actual trends. Ultimately, the first hypothesis is not confirmed for all post-socialist countries, shifting the analysis from long-term national trends to short-term group-level dynamics.

5.2. Consistency between the trends of income inequalities

The long-term consistency of development trends, particularly in EU countries, suggests persistent dependencies across individual nations. However, a dominant direction of change over time does not guarantee actual consistency, nor does it reflect relative changes between countries. To further investigate these findings, an individual-year analysis was conducted.

Following Trapeznikova (2019), who noted a consensus on country rankings by inequality, rankings for post-socialist countries were compiled across all datasets, sources, and measures. Rankings were prepared for key years: 1993, the earliest year with reliable data (earlier years, like 1991, posed analytical challenges); 2000, marking the stabilization of major inequality shifts; 2004, aligned with EU accession; and 2010, 2015, and 2020, representing later trends. Observing rankings at multi-year intervals allowed verification of trend consistency across different countries. For each dataset, the country with the lowest inequality received a rank of 1, while the country with the highest inequality was ranked 25. If fewer than 25 countries were available, rankings were adjusted accordingly. To ensure comparability, bottom 50% income shares were ranked inversely—higher values (indicating greater equality) received a rank of 1, aligning them with other measures where lower values indicate more equality.

Due to conceptual differences between measures, hypothesis H2 examines ranking homogeneity within the same measure rather than across different ones. Among the Gini coefficient, Palma ratio, Atkinson index, and income shares, no full consistency exists in rankings of post-socialist countries, and greater data availability often increases ranking discrepancies. Over time, ranking consistency does not show a clear improvement. However, when comparing complete or nearly complete rankings, noticeable differences emerge between measures. In the selected years, the Atkinson index exhibited the highest ranking stability (43% of observations had a maximum deviation of ±2 places), while the Palma ratio showed the greatest discrepancies (only 21% of observations within the same range). For the Gini coefficient and income shares, consistency within a ±2-place range was 27% and 26%, respectively.

The analysis also highlights differences between EU and non-EU countries. Splitting rankings into smaller subgroups reveals significant differences in consistency. Among EU countries, the Atkinson index rankings remained exactly the same in over 30% of cases, with a ±2-place deviation occurring in 91% of observations. Slightly lower consistency was found in income shares of the bottom 50% (45% and 74%), Palma ratio (39% and 70%), and top 10% income shares (38% and 67%). The Gini coefficient, which had the largest dataset, showed the lowest consistency—33% of rankings were identical, while 48% had only minor deviations. Among non-EU countries, the highest consistency was also observed for the Atkinson index (44% and 61%), while other measures ranged between 19% and 26%, with ranking differences reaching over 10 places in some years.

Subgroup analysis showed increasing ranking consistency over time for EU countries, a pattern not observed in non-EU nations. This improvement became apparent between 2004 and 2010, continued in 2015, and by 2020, rankings for Palma ratio and bottom 50% income shares remained within a ±2-place range for all countries.

Due to conceptual differences, comparing inequality measures is challenging. However, the Gini coefficient and Palma ratio allow for partial comparison, and databases that provide full rankings for both (WID and WIID) show very high consistency.

At the broadest level, rankings of the most and least income-equal countries show high consistency across measures and sources, remaining stable over time. Countries with low inequality in the 1990s—Czech Republic, Slovakia, Slovenia, Hungary, and Belarus—have maintained their positions, although Belarus's ranking may be influenced by unreliable data. Conversely, the most unequal countries are post-Soviet, non-EU states, including Armenia, Azerbaijan, Turkmenistan, and Georgia. This pattern holds even in incomplete rankings, where Georgia, for example, has appeared in positions "8" or "18", depending on dataset coverage.

In conclusion, full homogeneity in rankings of the same inequality measures across sources cannot be confirmed, although the Atkinson index rankings show the highest consistency. However, marginal rankings remain stable over time, particularly in EU countries, where rank variation is lower. Additionally, comparable measures show strong internal consistency when sourced from the same database. Given that databases differ significantly in methodology, choosing the right data source is more impactful than selecting a specific inequality measure, as it has a greater effect on final rankings.

5.3. Variation in income inequality values

The previous analysis assessed consistency in individual country trends and group-wide rankings over time. In both cases, some degree of inconsistency emerged, which can be linked to data variance. High variance can distort inequality trends and rankings, particularly when it affects multiple countries. To quantify this variation, statistical methods including analysis of variance, coefficient of variation, and data range were applied. Table 2 presents average values of these measures for the Gini coefficient, Palma ratio, bottom 50% income share, and top 10% income share, based on all available sources. The Atkinson index was excluded due to variability in the ε parameter, which prevents meaningful cross-source comparisons. While trends and rankings could still be analyzed, value variance was assessed separately for each measure due to their different scales.

Table 2. Summary of average values of variation and range of data on income inequality

country	Gini coefficient			Palma ratio			Bottom 50%			Top 10%		
	s^2	CV	max-min	s^2	CV	max-min	s^2	CV	max-min	s^2	CV	max-min
CZ	3.26	0.16	4.18	1.24	0.37	0.80	0.01	0.18	0.07	0.01	0.35	0.13
HU	10.06	0.09	7.56	0.17	0.24	0.85	0.00	0.12	0.07	0.01	0.24	0.11
SI	14.35	0.13	8.96	0.22	0.33	1.02	0.00	0.15	0.08	0.01	0.30	0.14
MK	14.91	0.08	8.33	0.41	0.25	1.30	0.00	0.12	0.08	0.01	0.28	0.11
BG	19.55	0.17	10.38	1.35	0.36	1.56	0.01	0.17	0.10	0.01	0.34	0.12
LV	21.98	0.11	10.54	0.63	0.34	1.65	0.00	0.14	0.09	0.01	0.32	0.13
SK	22.89	0.13	11.39	5.00	0.35	0.89	0.00	0.16	0.07	0.01	0.31	0.14
TM	25.18	0.13	5.89	4.87	0.35	4.01	0.00	0.16	0.18	0.01	0.31	0.13
GE	30.09	0.11	10.68	2.29	0.34	2.68	0.01	0.15	0.13	0.00	0.30	0.09
RU	31.90	0.10	9.48	3.03	0.46	2.90	0.01	0.21	0.16	0.01	0.34	0.13
LT	35.69	0.12	12.04	76.9	0.37	6.22	0.00	0.13	0.09	0.01	0.32	0.12
EE	36.02	0.14	13.71	1.32	0.42	2.36	0.01	0.18	0.13	0.01	0.35	0.14
TJ	36.94	0.10	9.49	1.48	0.39	2.25	0.01	0.19	0.14	0.01	0.33	0.13
BY	37.58	0.16	11.75	0.60	0.36	1.36	0.00	0.18	0.12	0.01	0.31	0.14
UA	37.89	0.13	12.02	4.71	0.35	1.45	0.01	0.16	0.12	0.01	0.31	0.13
AL	38.73	0.12	12.38	0.53	0.26	1.50	0.00	0.13	0.09	0.00	0.30	0.11
HR	40.01	0.16	14.00	0.47	0.33	1.50	0.00	0.15	0.09	0.01	0.32	0.14
UZ	41.74	0.10	9.07	4.85	0.51	3.77	0.01	0.22	0.18	0.01	0.08	0.12
PL	43.17	0.17	16.44	0.79	0.39	1.85	0.01	0.19	0.12	0.01	0.31	0.13
AM	43.24	0.14	13.26	1.99	0.35	2.44	0.01	0.16	0.13	0.00	0.33	0.12
RO	44.34	0.16	16.35	1.17	0.39	2.06	0.01	0.20	0.13	0.01	0.31	0.12
MD	46.55	0.15	13.84	0.74	0.28	1.79	0.00	0.13	0.09	0.01	0.32	0.13
KG	59.55	0.16	14.49	1.80	0.34	2.16	0.01	0.17	0.13	0.00	0.31	0.11
KZ	62.75	0.17	14.88	1.59	0.39	2.32	0.01	0.19	0.14	0.01	0.34	0.14
AZ	187.2	0.29	25.77	1.96	0.47	2.43	0.01	0.24	0.17	0.01	0.34	0.17

Where: s^2- variance, CV - coefficient of variation.

Source: own compilation.

Compared to the previous two aspects of data consistency tested, value consistency showed the greatest variation. For EU countries (excluding Poland), the range of income shares (bottom 50% and top 10%) does not exceed 20 p.p., with the average variation between the maximum and minimum value of the bottom 50% income share being less than 10 p.p. and 11-15 p.p. for the top decile. In non-EU countries, average variation exceeds 10 p.p., and in some cases, data sources are less consistent in estimating bottom 50% shares than top 10% shares. Across the entire group, standard deviation remains below 10%, but the coefficient of variation highlights issues with top-income measurement. While for lower-income groups it remains below 20% (except for Azerbaijan—24%, Uzbekistan—22%, and Russia—21%), at the top decile, even in countries with previously high consistency (e.g., Czech Republic), it exceeds 30%. This explains trend correlation differences found earlier in the analysis.

The Gini coefficient results showed different patterns. At the national level, the average difference between sources each year was 11.83 (on a 0-100 scale), meaning a country could be placed in different inequality groups in the same year. Czech Republic (4.18) and Hungary (7.58) showed the highest consistency, while Azerbaijan (25.77, and a maximum exceeding 30) exhibited the greatest variation. Poland (16.44) was the second least consistent, with twice the difference reported by Brzezinski, Myck, and Najsztub (2022). Regarding the coefficient of variation, most countries remained around 15%, except Azerbaijan (26%). Notably, both Czech Republic and Kyrgyzstan had the same variation level (16%), though their absolute data ranges differed significantly—3.3 vs. 60, respectively. This suggests a large spread in values despite a relatively stable ratio to the mean.

The Palma ratio again reveals substantial differences between subgroups. In EU countries (excluding Lithuania), the average variation between maximum and minimum values remained below 2, with Czech Republic (0.8) and Poland (1.85) showing the lowest fluctuation. In contrast, Central Asian countries exhibited significantly higher variation, with Uzbekistan (3.77) and Turkmenistan (4.01) indicating an average difference of nearly 4 in the income shares of the richest two deciles vs. the bottom 40%.

The value consistency analysis found no widespread stability in measures, disproving hypothesis 3. While EU countries exhibited lower variation, differences between subgroups persisted. In some cases, specific measures showed relatively stable values with low dispersion, but this stability was country-specific rather than measure-specific. The observed variation aligns with Bartels and Metzing (2019), who found that some countries exhibit minimal fluctuations, while others show significant discrepancies.

6. Conclusions

This article examines the consistency of income inequality data in post-socialist countries, from Central and Eastern Europe and Central Asia, across common measures and databases, considering the impact of methodological differences. Such analyses were mainly carried out for single measures, databases, or selected countries, and this article aimed to fill the research gap for a selected group of countries. The formulated hypotheses tested the consistency of development trends, the stability of country rankings, and the constancy of values across different measures. The analysis covered both long-term trends for individual countries and comparisons across time periods. Data were sourced from leading inequality databases—including WID, SWIID, WIID, OECD-IDD, GCIP, WDI, and Eurostat—and examined using the Gini coefficient, Palma ratio, Atkinson index, and income shares of the bottom 50% and top 10%.

The analysis of hypothesis H1 confirmed high consistency in long-term income inequality trends for EU countries where different measures showed aligned trajectories. In contrast, non-EU countries exhibited lower consistency, with occasional contradictory trends (Goda, 2016; De Maio, 2007). While databases focused solely on inequality and broader economic datasets produced similar trends, methodological differences remained relevant, particularly in WID data, which diverged due to its fiscal-data focus. For EU countries, correlations often exceeded 90%, confirming stable trends regardless of measure or source. However, in non-EU countries, database choice played a much greater role in determining trends.

Methodological choices significantly affected data consistency, particularly SWIID's oversmoothing and WID's fiscal-data inclusion, which lowered correlations in some cases (De Maio, 2007; Alvaredo et al., 2016). This conclusion particularly applies to post-Soviet countries that are not part of the EU, where the richest individuals may have disproportionately more income than in most CEE countries, and where access to high-income data may be more difficult, resulting in significantly lower data consistency.

Hypothesis H2 tested the stability of country rankings over time. Results showed that EU countries exhibited greater ranking consistency, particularly after 2004, likely due to improved data quality from EU statistical integration. In contrast, non-EU countries showed significant variation, with rankings shifting by up to 10 places for some measures.

Despite inconsistencies, some ranking patterns remained stable. Czech Republic, Slovakia, Slovenia, and Hungary consistently ranked among the least unequal, while Georgia and Turkmenistan were among the most unequal. Within each database, rankings remained largely consistent, typically shifting by only one place. This suggests that for non-EU countries, the choice of database has a stronger impact than the choice of measure.

Hypothesis H3 examined value consistency across databases, revealing that full consistency was not found. Variation was more limited in EU countries, while non-EU countries showed greater inconsistencies, particularly for measures like the Palma ratio.

In conclusion, the article indicates a high level of consistency in income inequality trends over the long term and highlights strong correlations between different data sources for the same measures. However, they are inflated by the high consistency of data for EU countries, which is why only for this subgroup it would be possible to truly confirm the existence of consistent trends. The ranking of countries is most consistent in the context of extreme equality or inequality and between measures from the same database, while the occurrence of full consistency in the values of individual measures practically does not occur, which is the result of inconsistency at the level of the values of given measures, even if the level of their variance is moderate.

The key finding is that data selection is crucial when studying income inequality, requiring awareness of methodological challenges across measures, sources, and countries. The analysis revealed that SWIID's oversmoothing and WID's use of fiscal data led to significant data divergences, sometimes even producing contradictory year-on-year trends. The extent of these issues depends on the country and research focus. For EU countries, data from a single database tend to show consistent development trends, regardless of the measure. However, in international comparisons, where country differences play a larger role, the choice of data source becomes more critical—although its influence weakens over time. For instance, selecting a Palma ratio dataset for 2000 requires greater caution than for 2020 due to historical inconsistencies. For non-EU post-socialist countries, low correlations between datasets, significant discrepancies between measures, and unstable rankings highlight the need for careful selection of both the measure and data source. In such cases, any choice can lead to vastly different results, making methodological justification essential. This applies both to analyses for single countries and especially to broad international comparisons.

References

Alvaredo. F., (2011). A note on the relationship between top income shares and the Gini coefficient. *Economics Letters,* Vol. 110(3). https://doi.org/10.1016/j.econlet.2010.10.008.

Alvaredo, F., Atkinson, A., Chancel, L., Piketty, T., Saez, E. and Zucman, G., (2016). Distributional National Accounts (DINA) Guidelines: Concepts and Methods used in WID.world. *Wid.world working paper,* No. 2016(2).

Atkinson, A. B., (1970). On the measurement of inequality. *Journal of Economic Theory 2*(3). In Atkinson, A.B., and Piketty T. (eds), Top Incomes Over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries. https://doi.org/10.1093/oso/9780199286881.003.0002.

Atkinson, A. B., (2007). Measuring top incomes: methodological issues.

Atkinson, A. B., Brandolini, A., (2010). On analyzing the world distribution of income. *World Bank Economic Review,* 24(1).

Atkinson, A.B., Piketty, T. and Saez, E., (2011). Top Incomes in the Long Run of History. *Journal of Economic Literature*, 49(1). DOI: 10.3386/w154.

Ayala, L., Pérez, A. and Prieto-Alaiz, M., (2022). The impact of different data sources on the level and structure of income inequality. *SERIEs,* 13. https://doi. org/ 10.1007/s13209-021-00258-0.

Bartles, Ch., Metzing, M., (2019). An integrated approach for a top-corrected income distribution. *Journal of Economic Inequality,* 17. https://doi.org/10.1007/s10888-018-9394-x.

Bellù, L. G., Liberati, P., (2006). Policy Impacts on Inequality: Inequality and Axioms for its Measurement. *EASYPol,* 054.

Blanchet, T., Flores, I. and Morgan, M., (2018). The Weight of the Rich: Improving Surveys with Tax Data. *WID.world Work. Pap. Ser.,* 2018/12

Brzeziński, M., Salach, K., (2022). Determinants of inequality in transition countries. *IZA World of Labor,* 2022(496). https://doi.org/10.15185/izawol.496.

Bukowski, P., Novokmet, F., (2017). Inequality in Poland: Estimating the whole distribution by g- percentile, pp. 1983–2015. *LIS Working papers,* 731, *LIS Cross-National Data Center in Luxembourg.*

Corak, M., (2013). Income Inequality, Equality of Opportunity, and Intergenerational Mobility. *Journal of Economic Perspectives,* 27(3). https://doi.org/10.1257/jep.27. 3.79.

Cobham, A., Schlögl, L. and Sumner, A., (2016). Inequality and the Tails: the Palma Proposition and Ratio. *Global Policy,* 7(1) https://doi.org/10.1111/1758-5899.12320.

Cobhan, A., Sumner, A., (2013). Is it all about the tails? The Palma measure of income inequality. *Center for Global Development Working Paper,* 2013(308).

Dubois, M., (2016). A note on the normative content of the Atkinson inequality aversion parameter. *Post-Print hal-01837118, HAL.*

De Maio, F. G., (2007). Income inequality measures. *Journal of Epidemiology & Community Health,* 61(10). https://doi.org/10.1136/jech.2006.052969.

Eurostat, (2020) Flash estimates of income inequalities and poverty indicators for 2019 (FE 2019) Experimental results. *Eurostat.*

Farris, F. A., (2010). The Gini index and measures of inequality. *The American Mathematical Monthly,* 117(10).

Ferreira, F. H. G., Lusting, N. and Teles, D., (2015). Appraising Cross-National Income Inequality Databases: An Introduction. *Journal of Economic Inequality,* 13(4). https://doi.org/10.1007/s10888-015-9316-0.

Goda, T., (2016). Global trends in relative and absolute income inequality. *Ecos de Economía,* 20(42). DOI: 10.17230/ecos.2015.42.3.

Galbraith, J. K., Choi, J., Halbach, B., Malinowska, A. and Zhang, W., (2016). A Comparison of Major World Inequality Data Sets: LIS, OECD, EU-SILC, WDI, and EHII. *Income Inequality Around the World,* Vol. 44. https://doi.org/10.1108/S0147-912120160000044008.

Jabkowski, P., (2009). Miary nierówności społecznych – podstawy metodologiczne. W: Podemski, K. (red.), Spór o społeczne znaczenie społecznych nierówności. *Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu.*

Neef, T., (2020). What's New About Income Inequality in Russia (1980-2019)? Trends in Comparison to Eastern Europe. *World Inequality Lab – Issue Brief 2020/05.*

Jenkins, S. P., (2015). World income inequality databases: an assessment of WIID and SWIID. *Journal of Economic Inequality,* 2015(13). https://doi.org/10.1007/s10888-015-9305-3.

Jenkins, S. P., (2017). Pareto Models, Top Incomes and Recent Trends in UK Income Inequality. *Economica,* Vol. 84(334). https://doi.org/10.1111/ecca.12217.

Jędrzejczak, A., Pekasiewicz, D., (2018) Properties of Selected Inequality Measures Based on Quantiles and Their Application to the Analysis of Income Distribution in Poland by Macroregion. *Argumenta Oeconomica Cracoviensia,* 18. https://doi.org/10.15678/AOC.2018.1803.

Larrimore, J., Burkhauser, R. V. and Armour, P., (2018). Accounting for income inequality in survey and administrative data: Evidence from the US Current Population Survey. *Journal of Economic and Social Measurement,* 43(1–2).

Latty, K., (2015). A five parameter Atkinson like index featuring relative income effects, with a seven-parameter extension for nonlinear (prioritarian) social welfare functions. *Notes on income inequality.*

OECD, (2023). Income distribution. *OECD Social and Welfare Statistics* (database). https://doi.org/10.1787/data-00654-en, https://www.oecd.org/els/soc/IDD-ToR.pdf.

Milanovic, B., (1999). Explaining the Increase in Inequality During the Transition. *Economics of Transition,* Vol. 7(2). https://doi.org/10.1111/1468-0351.00016.

Novokmet, F., Piketty, T. and Zucman, G., (2018). From Soviets to oligarchs: inequality and property in Russia 1905-2016. *The Journal of Economic Inequality,* 16. https://doi.org/10.1007/s10888-018-9383-0.

Pkietty T., (2014). Capital in the Twenty-First Century. Harvard University Press.

Pascola, R., Rocha, H., (2017). Inequality measures for wealth distribution: Population vs individuals perspective. *Physica A Statistical Mechanics and its Applications,* 492. https://doi.org/10.1016/j.physa.2017.11.059.

Ravallion, M., (2015). The Luxembourg Income Study. *The Journal of Economic Inequality,* 13(4). https://doi.org/10.1007/s10888-015-9298-y.

Sen, A., Foster, J., (1997). On Economic Inequality. *Oxford University Press.*

Sitthiyot, T., Holasut, K., (2020). A simple method for measuring inequality. *Humanities & Social Sciences Communication,* 6(112). https://doi.org/10.1057/ s41599-020-0484-6.

Solt, F., (2009). Standardizing the World Income Inequality Database. *Social Science Quarterly, Southwestern Social Science Association,* 90(2).

Solt, F., (2020). Measuring Income Inequality Across Countries and Over Time: The Standardized World Income Inequality Database. *Social Science Quarterly,* 101(3).

The Equality Trust, (2011). Income inequality: *Trends and Measures. Equality Trust Research Digest. 2.* Retrieved from: https://equalitytrust.org.uk/sites/default/files/research-digest-trends-measures-final.pdf.

Trapeznikova, I., (2019). Measuring income inequality. *IZA World of Labor* (462). https://doi.org/10.15185/izawol.462.

UNU-WIDER, (2022). World Income Inequality Database (WIID). https://doi.org/10.35188/UNU-WIDER/WIID-300622.

Vermeulen, P., (2016). Estimating the top tail of the wealth distribution. *American Economic Review*, 106(5).

Voitchovsky, S., (2005). Does the Profile of Income Inequality Matter for Economic Growth? *Journal of Economic Growth,* 10(3). https://doi.org/10.1007/s10887-005-3535-3.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 151–166, https://doi.org/10.59139/stattrans-2025-044
Received – 09.05.2025; accepted – 04.06.2025

On a new goodness-of-fit test for multivariate normality with fixed parameters based on the David-Hellwig test idea

Grzegorz Kończak[1]

Abstract

The article presents a proposal for a goodness-of-fit test for multivariate normality. The idea of the test is based on the empty cells test, which is well known in the literature. In the empty cells test, the area of the random variable's variability is divided into m disjoint cells. Assuming the truth of hypothesis H_0, which proclaims the multivariate normality of the distribution with given parameters, disjoint cells are arranged in such a way that random values with equal probabilities are in each cell. Based on the n-element sample, the number of empty cells, i.e. the cells without any elements from the sample, is determined. Crucial to the proposed procedure is the division of the multidimensional area of variation into disjoint cells. The advantage of this test is that it can be used for relatively small samples. In the article, a simulation comparison of the proposed test's properties and the Kolmogorov-Smirnov test's multivariate version is carried out.

Key words: multivariate normality, inferential statistics, empty cells test, Monte Carlo study.

1. Introduction

Evaluating normality is essential for various statistical methods that rely on the assumption of data being normally distributed. Normality assessments aim to ascertain whether a data assemblage deviates markedly from the Gaussian distribution (Hernandez, 2021). A variety of normality tests are employed in practice, with the most prevalently utilized being the Shapiro-Wilk test (Shapiro and Wilk, 1965), the Kolmogorov-Smirnov test (Kołmogorov, 1933; Smirnov, 1939), the Lilliefors test (Lilliefors, 1967), the Anderson-Darling test (Anderson and Darling, 1952), and the Jarque-Bera test (Jarque and Bera, 1980). The normality tests mentioned above represent merely a small component of such methodologies. Hernandez (2021) delineates a comparative analysis of 55 normality assessments. This plethora of normality assessments arises from the fact that deviations from normality can exhibit a wide variety.

[1] Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Katowice, Poland. E-mail: grzegorz.konczak@ue.katowice.pl. ORCID: https://orcid.org/0000-0002-4696-8215.

The evaluation of multivariate normality is considerably more complex. Multivariate normality constitutes a vital presumption in numerous multivariate statistical analyses, ensuring the integrity of methodologies such as MANOVA and PCA. Many statistical tests of multivariate normality have been documented in the literature. Traditional assessments for multivariate normality, such as those formulated by Mardia (1975), concentrate on evaluating skewness and kurtosis within multivariate datasets. These assessments have been foundational in discerning deviations from normality. Mardia (1975) elaborates on various methodologies for assessing multivariate normality, underscoring significant advancements such as techniques based on Mahalanobis angles and distances. Malkovich and Afifi (1973) discuss the generalization of univariate skewness and kurtosis statistics, in conjunction with the W statistic by Shapiro and Wilk, to evaluate multivariate normality using Roy's union-intersection principle. Kankainen, Taskinen, and Oja (2007) introduced assessments for multivariate normality that extend classical univariate measures to a multivariate context. These assessments are based on the Mahalanobis distance between multivariate location vector estimates and the distance between scatter matrix estimates. The authors developed an asymptotic theory to provide approximate null distributions and assess the asymptotic efficiencies of these assessments. These evaluations are particularly advantageous in practical applications where data may not strictly conform to normality assumptions, offering a reliable alternative to classical methodologies.

Liang and Yang (2022) proposed a novel assessment that amalgamates necessary-only characterization and statistical representative points. They also present an illustrative example, accentuating the assessment's supplementary function alongside existing assessments in the literature. Mudholkar, McDermo, and Srivastava (2024) presented a multivariate adaptation of the Lin and Mudholkar (1980) z-assessment for evaluating univariate normality, offering an accessible methodology. The research contributes to the understanding of multivariate normality evaluation and provides empirical refinements for the Zp assessment. Székely and Rizzo (2013) examined the theory and applications of energy statistics, showing their efficacy in inference and multivariate analysis. They discuss energy distance, a statistical distance that characterizes the equivalence of distributions of random vectors, drawing an analogy to Newton's gravitational potential energy. Mardia (2024) discussed the statistical properties of multivariate distributions, mainly focusing on multivariate skewness and kurtosis measures. It presented alternative forms of these measures and their applications in testing multivariate normality. The author derived exact moments and new approximations for the distributions of certain bilinear forms under multivariate normality. Additionally, the paper examines the impact of non-normality on the size of normality theory tests for covariance matrices.

1.1. Types of departures from the multivariate normality

Deviations from multivariate normality can take various forms. There are some typical departures from multivariate normality. Below, we present several typical examples of departures from multivariate normality distribution (Domański, 2009; Domański, 2011; Joenssen and Vogel, 2014; Ebner and Henze, 2020; Mardia, 2024):

- Uneven distribution of data in multidimensional space
  When data comes from a mixture of normal distributions, clusters with distinct locations and scales may occur. For example, a mixture of two normal distributions may generate bimodal data, i.e. with two distinct maxima. Quantile-quantile plots (*qq-plot*), probability plots, perspective plots, and contour plots can reveal such uneven distributions.
- Deviations from linearity on diagnostic plots
  On quantile-quantile and probability plots, deviations from linearity suggest departures from normality. For instance, a curved shape on a *qq-plot* may show skewness or kurtosis in the distribution.
- Non-zero values of multivariate measures of skewness and kurtosis
  Multivariate measures of skewness b_{1p} and kurtosis b_{2p} measure deviations from symmetry and the shape of the normal distribution. Non-zero values of these measures suggest departures from multivariate normality.
- Presence of outliers
  Outliers can significantly affect the results of multivariate normality tests. For example, on *qq-plot*, outliers often manifest as points far from the regression line. Removing outliers may lead to changes in test results and better fit to the normal distribution.
- Elliptical distributions
  Elliptical distributions form a broader class of distributions to which the normal distribution belongs. Multivariate normality tests may not be sensitive to deviations from normality within this class.
- Non-normal marginal distributions
  Therefore, it is essential to use multivariate normality tests, not just assessing the normality of particular variables.

Outliers can significantly affect the results of multivariate normality tests, and their removal can lead to a better fit for the normal distribution. It should be noted that even if the multivariate distribution is not normal, individual marginal variables may show a normal distribution.

There are many ways in which data can deviate from multivariate normality, including the uneven distribution of data in multivariate space, deviations from linearity in diagnostic charts, outliers, and elliptical distributions.

1.2. Types of tests for multivariate normality

Numerous methodologies for testing multivariate normality represent advancements in univariate normality testing. The majority of existing multivariate probability tests can be classified into four distinct categories (Domański and Pruska, 2000; Cramer and Howitt, 2004; Domański, 2009):

1. Procedures predicated upon graphical representations and correlation metrics,
2. Goodness-of-fit assessment methodologies,
3. Tests formulated based on skewness and kurtosis statistics,
4. Consistent methodologies derived from the empirical characteristic function.

The first of these categories involves visual data analysis using graphs, such as quantile-quantile charts (Royston, 1982). Perspective and contour charts can also be used for two-dimensional data (Korkmaz, Goksuluk, and Zararsiz, 2014). These charts allow one to assess whether the data follow the expected normal distribution (Koziol, 1993; Liang and Ng, 2009). For example, deviations from linearity in a qq-plot suggest deviations from normality.

Generalized goodness-of-fit tests make it possible to compare the empirical distribution of data with a theoretical normal distribution. Examples of such tests include the extension of the Anderson-Darling test proposed by Hawkins (Domański, 2011), the Kolmogorov-Smirnov test (Kesemen et al., 2021), the Cramér-von Mises test (Hernandez, 2021), and the chi-square test.

Tests based on measures of skewness and flattening use multivariate generalizations of skewness and flattening statistics to assess normality. Tests based on the measures mentioned above include Mardia (Mardia, 1975; Domański, 2009), the Doornik and Hansen omnibus test (Doornik and Hansen, 2008), and tests based on a multivariate version of the Shapiro-Wilk test (Domański, Gadecki, and Wagner, 1989).

Procedures based on empirical characteristic functions use an empirical characteristic function to compare the sample distribution with a normal distribution. An example of such a test is the Henze-Zirkler test (Szekely and Rizzo, 2013).

Many multivariate normality tests are extensions of one-dimensional normality tests. The appropriate test choice depends on the specifics of the research problem, sample size, and expected deviations from normality (Joenssen and Vogel, 2014).

1.3. Testing multivariate normality

Multivariate normality tests allow us to assess whether a multivariate dataset follows a multivariate normal distribution. The multivariate normal distribution is an essential assumption in many statistical modeling and inference methods. Many multivariate normality tests are available, each with its advantages and disadvantages.

Researchers have extensively explored various statistical approaches to evaluate the multivariate normality assumption due to its critical role in ensuring the validity of statistical inferences. The literature on multivariate normality tests encompasses a range of methodologies developed to assess the assumption of normality in multivariate distributions. Traditional tests such as Mardia's test (Mardia, 1974) focus on multivariate skewness and kurtosis, providing a measure that helps identify deviations from normality based on these higher moments. Henze-Zirkler's test (Henze and Zirkler, 1990) employs statistics based on the Mahalanobis distance, offering robustness against sample size variations and good power against alternatives. The Royston test (Royston, 1983) for multivariate normality extends the Shapiro-Wilk test for univariate normality. It involves transforming the Shapiro-Wilk statistic into one that Royston claims to be approximately chi-squared distributed with equivalent degrees of freedom, calculated based on the cumulative distribution function for the standard normal distribution. The test statistic combines the Shapiro-Wilk statistics for the separate variables.

The energy multivariate normality test (Székely and Rizzo, 2013, 2017; Móri, Székely, and Rizzo, 2021) is a distance-based test that assesses multivariate normality by comparing the energy of the observed data with the energy of data sampled from a multivariate normal distribution. The energy of a dataset is a measure of dispersion or spread, calculated based on the distances between observations in the dataset. This test computes a test statistic based on the difference in energy between the observed data and simulated multivariate normal data. Departures from multivariate normality are detected if the observed energy significantly differs from the expected energy under the null hypothesis.

Recent approaches include copulas and bootstrap methods, allowing for more flexible testing in complex data structures. These modern techniques address the limitations of classical tests, such as their sensitivity to outliers and dependency on large sample sizes. The literature also discusses the practical applications of these tests in fields ranging from finance and economics to the biological sciences, where the correct identification of data normality significantly impacts the conclusions of empirical research. Overall, the evolution of multivariate normality tests reflects a broader trend toward more computationally intensive and more accurate statistical methodologies in the face of increasingly complex data.

Commonly employed methods include multivariate extensions of univariate normality, distance-based, and transformation-based tests. Multivariate normality tests typically examine the distributional properties of multivariate data while considering the interrelationships among variables. Among these, tests such as Mardia's, Henze-Zirkler, and Royston's tests are widely used, leveraging sample moments and distributional properties to assess departures from multivariate normality. On the other hand, distance-based tests rely on measures of distances or dissimilarities between observations to evaluate deviations from multivariate normality. Examples include the Mahalanobis distance criterion and the correlation ratio criterion, which assess the adequacy of the observed data distribution compared to the multivariate normal distribution.

Furthermore, transformation-based tests involve transforming multivariate data to conform more closely to a multivariate normal distribution and assessing the goodness of fit of the transformed data. Despite the availability of these diverse methodologies, challenges persist in their application, including the sensitivity to sample size, dimensionality, and underlying distributional characteristics. Additionally, the choice of a proper test depends on the specific characteristics of the dataset and the research context. As such, further research is warranted to develop robust and versatile methodologies for assessing multivariate normality across diverse research domains.

In statistical research, the inference of unknown population parameters is often carried out based on random samples. In the case of characteristics measured on strong scales (interval and ratio), samples of relatively small sizes are sufficient to carry out the inference effectively.

The multivariate goodness-of-fit multivariate normality test is proposed in the paper. The hypothesis regarding the distribution of the multivariate random variable will be examined. The idea of this proposal is based on the empty cells test.

2. The proposal of a goodness-of-fit multivariate normality test based on the idea of David-Hellwig empty cells

The empty cells test (David, 1950; Hellwig, 1965; Domański, 2012) is one of the goodness-of-fit tests. The hypothesis regarding the form of the distribution of random variables can be assessed with this test. A random variable's entire area of variability is divided into m cells, and the number of elements from the random sample in each cell is counted. Then, the number of empty cells is counted. This number of empty cells is compared to the critical value. Domański and Pruska (2000) presented the interpolated critical values for the empty cells test statistic.

Let us assume that an n-elements random sample was taken from the population. We will test the hypothesis about the form of the distribution

$$H_0: F(x) = F_0(x)$$

against the alternative

$$H_1: F(x) \neq F_0(x)$$

where F_0 is the specified distribution.

First, we divide the random variable's variability area $\mathbb{X}$ into m disconnected cells Mi. It could be written as (Kończak, 2005, 2008):

$$\mathbb{X} = \bigcup_{i=1}^{m} M_i \tag{1}$$

where $\mathbb{X}$ is the area of variability $\mathbb{X}$ of the random variable X, and

$M_i \cap M_j = \emptyset, for\ , i, j\ \in \{1, 2, \dots, m\}, i \neq j$ for $i = 1, 2, \dots, m$, and

$$P(x \in M_i) = \frac{1}{m}, \tag{2}$$

for $i = 1, 2, \dots, m$, if the hypothesis H_0 is true.

The test statistic in the empty cells test has the following form:

$$K_n = card\{j: m_j = 0\}$$

where m_j means the number of elements from a random sample in j-th cell.

The critical area of the empty target test is right-sided.

Under the assumption of the truth of the hypothesis H_0, for a division into m targets, the probability of k empty cells, when an n-element sample is taken, is expressed by the following formula (Hellwig, 1965):

$$p_k(n, m) = \binom{m}{k} \sum_{r=0}^{m-k} (-1)^r \binom{m-k}{r} \left(1 - \frac{k+r}{m}\right)^n \tag{3}$$

and the cumulative probability function has the following form:

$$P_k(n, m) = \sum_{s=0}^{k} \binom{m}{s} \sum_{r=0}^{m-s} (-1)^r \binom{m-s}{r} \left(1 - \frac{s+r}{m}\right)^n \tag{4}$$

For the assumed significance level α the rejection region could be denoted as:

$$K = \{k: k \geq K_{n,\alpha}\}$$

where $K_{n,a}$ are taken from tables (e.g. David (1950), Hellwig (1965), Domański and Pruska (2000).

Domański (2011) points out the superiority of the power of the empty cells test over the Shapiro-Wilk test in testing unidimensional normality for sample sizes of n≤20.

The one-dimensional and multivariate normality tests presented above overwhelmingly do not consider the case of known parameters. Among the few exceptions is the Kolmogorov-Smirnov test, which checks for conformity to a distribution, particularly a normal distribution, with fixed parameters. There are few goodness-of-fit tests for multivariate normality, and those that exist are highly constrained. McAssey (2013) presented a multivariate goodness-of-fit test based on Mahalanobis distances. Some tests are extremely complex to implement, so much so that they can handle only three or four p-variates (Romeu and Ozturk, 1993). Hanusz, Tarasińska, and Zieliński (2012) modified the Shapiro-Wilk test for the case of normality with a known mean.

Let $\boldsymbol{X} = (X_1, X_2, \cdots, X_k)$ be the k-dimensional continuous vector random variable with the cumulative distribution function $F(\boldsymbol{x})$. Let $x_1, x_2, \cdots, x_n$ where $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ik})$ for $i = 1, 2, \cdots, n$, be an n-element sample.

Let us consider the simple null hypothesis of multivariate normality denoted by the form:

$$H_0\colon F_p(\boldsymbol{x}) = F_{0p}(\boldsymbol{x})$$

where $F_{0p}(\boldsymbol{x}) \sim N_p(\boldsymbol{x};\ \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$, $\boldsymbol{\mu_0}$ and $\boldsymbol{\Sigma_0}$ are given vector of means and covariance matrix,

against the alternative hypothesis

$$H_1\colon \sim H_0$$

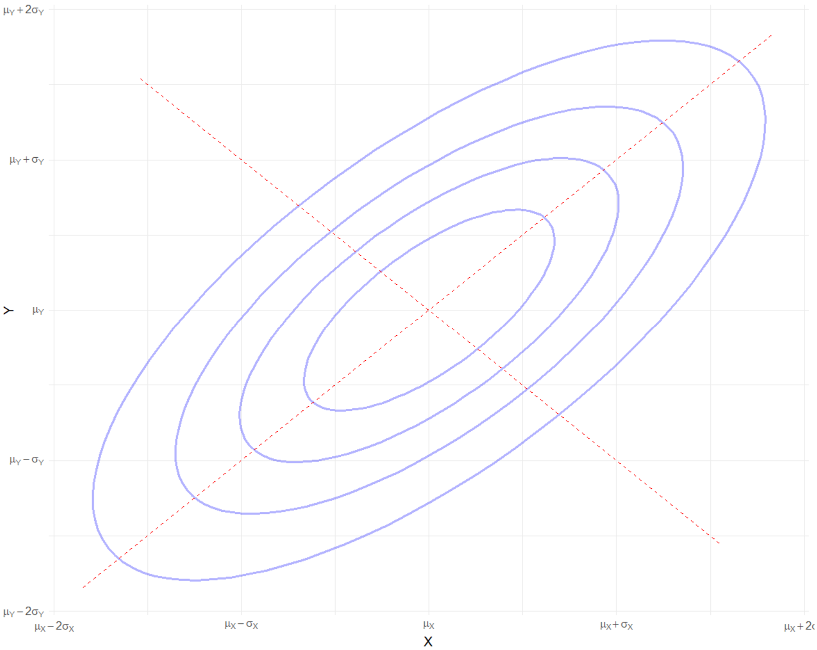


Figure 1. The idea of constructing cells for goodness-of-fit test for bivariate normality (m = 20)

For the application of the empty cells test for multivariate normality, it is necessary to identify disjoint cells that satisfy conditions (1) and (2). The idea of cells construction is based on using confidence ellipsoids and then dividing the created areas into 2^p targets, where p is the dimension of the space. A visualization of how the area of variation is divided into cells is shown only for a two-dimensional distribution (p = 2). Figure 1 shows the idea of target construction for the case of a two-dimensional normal distribution with a vector of expected values $\boldsymbol{\mu_0} = (\mu_X, \mu_Y)$ and a covariance matrix $\boldsymbol{\Sigma_0} = \begin{bmatrix} \sigma_X^2 & \sigma_X\sigma_Y \\ \sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$. The division of the area of variation into m = 20 cells is shown. For the indicated cell design, the probabilities of observations occurring in all cells under H_0 are the same.

The p-dimensional ellipsoids are determined by:

$$(\boldsymbol{x}-\boldsymbol{\mu_0})^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu_0}) \leq \chi^2_{p,1-\alpha} \tag{5}$$

where

$\boldsymbol{x} \in \mathbb{R}^p$,

$\boldsymbol{\mu_0}$ - given vector of expected values,

$\boldsymbol{\Sigma}_0$ - given covariance matrix (dimension $p \times p$),

$\chi^2_{p,1-\alpha}$ - quantile of the chi-squared distribution with p degrees of freedom at confidence level $1-\alpha$.

The division of the elliptic strip into 2$^{\mathrm{p}}$ cells is determined by the eigenvectors of the $\boldsymbol{\Sigma}_0$ matrix.

Figure 2 shows the partitioning into $m = 20$ cells for a two-dimensional normal distribution with an expectation vector $\boldsymbol{\mu_0} = (0{,}0)$ and a covariance matrix $\boldsymbol{\Sigma_0} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$. The figure shows $n = 20$ element random samples from a two-dimensional normal distribution. In the general case, the number of cells m may differ from the sample size n. If $m \neq n$, the critical values can be determined based on (3) and (4). In the case of $m = n$, critical values can be obtained from Domanski and Pruska (2000). The theoretical contour plot represent the bivariate normal distribution with parameters $\boldsymbol{\mu_0}$ and $\boldsymbol{\Sigma_0}$. The empirical contour plots in Figure 2 show estimates from an $n = 20$ element sample of the density of the distribution. Two variants of expected values of X and Y variables were considered:

a) $\boldsymbol{\mu} = (1, 0)$

b) $\boldsymbol{\mu} = (2, 2)$

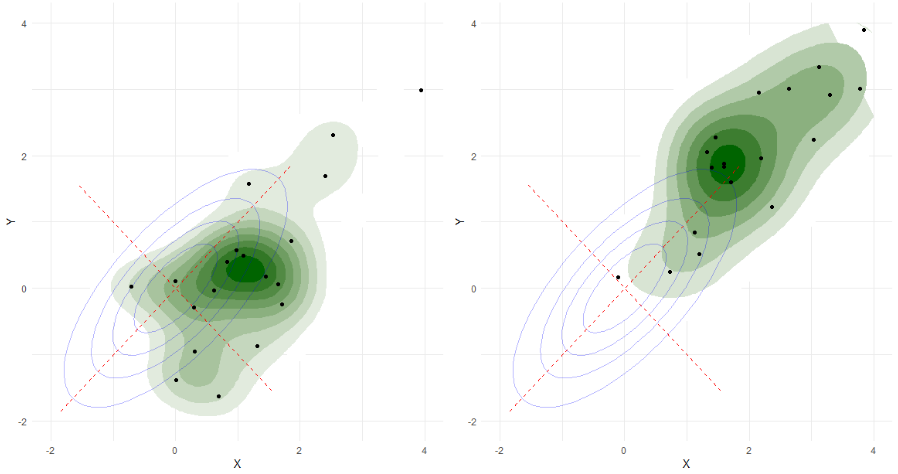


Figure 2. Sample of n = 20 elements with empirical and theoretical, under H_0, contour plot

For the sample size n = 20, the number of cells m = 20, and the significance level α=0.05, the interpolated critical value in the empty cells test equals 9.96 (Domański and Pruska, 2000). This means that hypothesis H_0 is rejected if the number of empty cells is greater than or equal to 10. In both cases shown in Figure 2, the number of empty cells is k = 11 (left) and k = 15 (right). In both cases presented in Figure 2, the hypothesis H_0 should be rejected for the assumed significance level $\alpha = 0.05$.

3. The properties of the test - Monte Carlo study

The simulation analyses carried out evaluated the probability of rejecting the H_0 hypothesis. The size and power of this test were analyzed. Simulation analyses were conducted for a two-dimensional normal distribution. Two variants were considered:

a) $N_2(x;\ \mu_0, \Sigma_0)$, where $\mu_0 = (0{,}0), \Sigma_0 = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$

b) $N_2(x;\ \mu_0, \Sigma_0)$, where $\mu_0 = (0{,}0), \Sigma_0 = \begin{bmatrix} 1.0 & 0.7 \\ 0.7 & 1.0 \end{bmatrix}$

Random values were generated for variant (a) or (b) assuming changes in expected values

$$\delta_X = 0.0, 0.1, \dots, 1.5;\ \delta_Y = 0.0, 0.1, \dots, 1.5.$$

If $\delta_X = \delta_Y = 0.0$, then the hypothesis H_0 was true, otherwise H_0 was false. The estimated probabilities of rejecting H_0 were obtained by a Monte Carlo study. The sample sizes considered were n = 12, 20, and 28. In all considered cases, the number of cells was set to $m = n$, and the number of simulations was established to N = 1,000. The largest possible standard error when considering rejection proportions of multivariate normality with 1,000 simulations is $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = \sqrt{\frac{0.5(1-0.5)}{1000}} \approx 0.0158$.

In generating random samples from a two-dimensional normal distribution, the *rmvnorm* function from the **mvtnorm** package in *R* was used (mvtnorm 2025). The function allows to generate random values from a multivariate normal distribution with a given vector of expected values $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_0$.

The probabilities of rejecting H_0 were estimated for the multivariate Kolmogorov-Smirnov test (KS) and the described proposal multivariate normality test (mvnDH). Figure 3 shows the power function of the mvnDH test of concordance and the KS test for a two-dimensional normal distribution for independent variables (variant a). The presentation includes sample sizes of n = 12, 20, and 28. It is noticeable that the power of both tests increases as the sample size increases. In all cases, the KS test features a slightly higher power. Figure 4 shows the power function of the mvnDH concordance test and the KS test for a two-dimensional normal distribution for the dependent variables (variant b). The presentation includes sample sizes of n = 12, 20, and 28. It is noticeable that the power of both tests increases as the sample size increases. For simultaneous increases in the expected values of both variables (X and Y), the probability of rejecting the H_0 hypothesis for the KS test is higher than for the mvnDH test. However, when increasing the expected value of only one variable (X or Y), the probability of rejecting the H_0 hypothesis for the mvnDH test is greater than for the KS test. Depending on the type of deviation from H_0, the effectiveness of the analyzed tests is not the same.

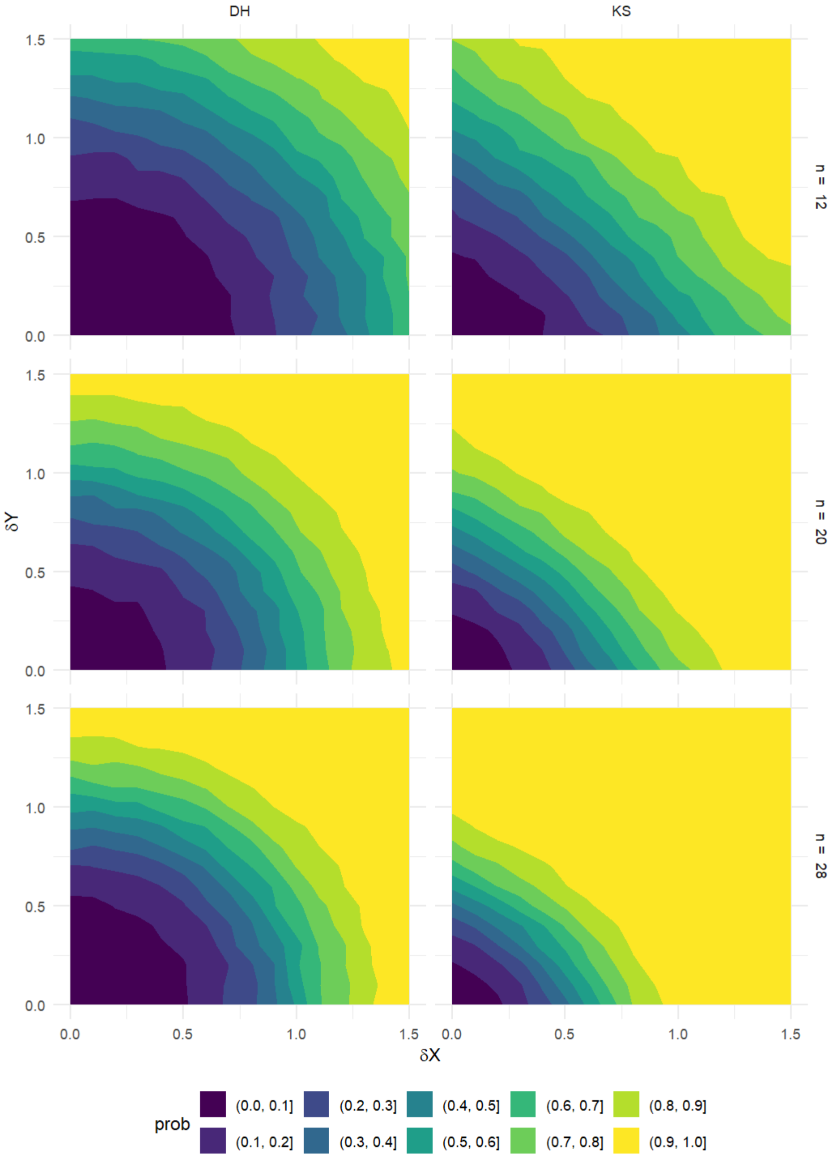


Figure 3. Estimated probabilities of rejecting H_0 for the case a) - independent variables

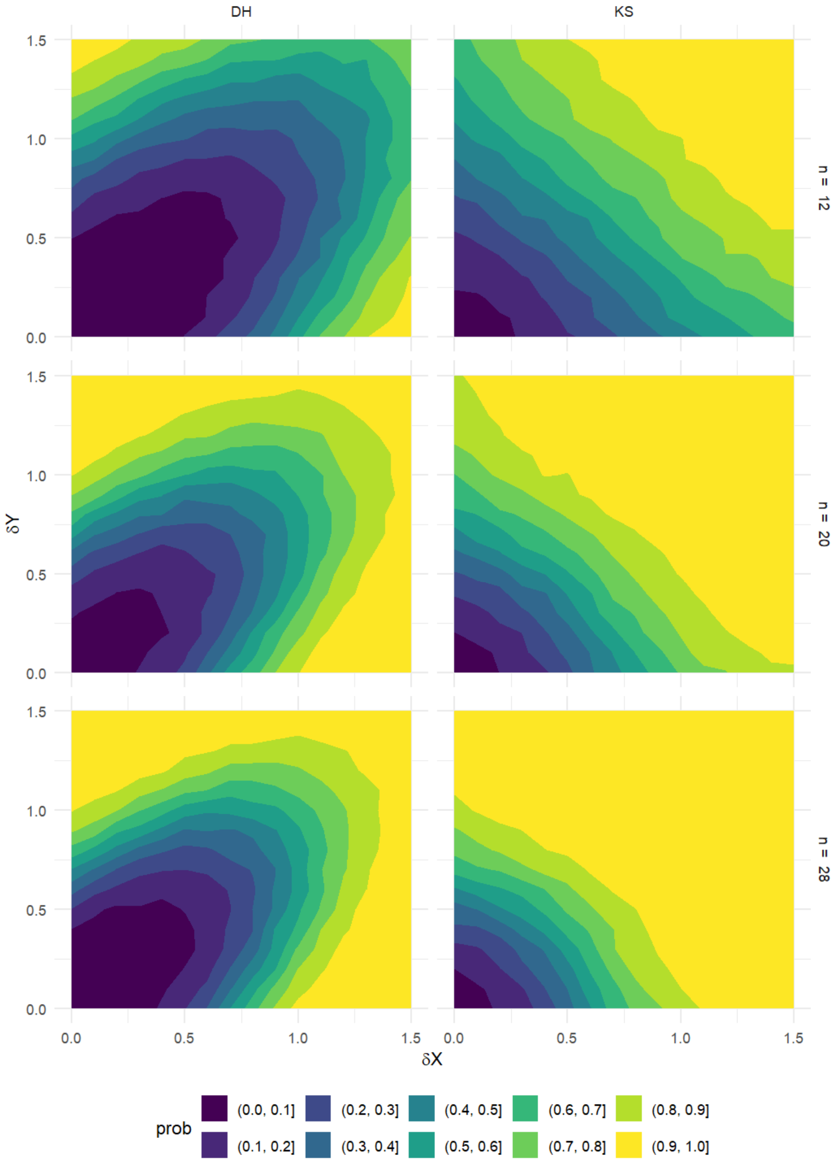


Figure 4. Estimated probabilities of rejecting H_0 for the case b) - dependent variables

4. Conclusions

In reality, many economic, social, or medical issues are complex and involve multiple dimensions. This article addresses the issue of statistical inference for multivariate normality tests. A proposal for using the empty cells test to assess multivariate normality is presented. The idea of this test is based on the concept of empty cells. In conducting an empty cells test, it is crucial to partition the area of variation into disjoint cells, which is particularly significant in multivariate analysis. In the proposed testing procedure, cells are defined by confidence ellipsoids and are further segmented along the eigenvectors of the variance-covariance matrix. This division of the area of variation ensures equal probabilities of observations in the cells under H_0. The proposed test validates the conformity of the empirical distribution to a preset, precisely specified, multivariate normal distribution. Accordingly, it reacts not only to changes in the parameters of the distribution, but also to deviations from normality, in particular to the occurrence of a distribution type other than normal.

In simultaneous increases in the expected values of both variables (X and Y), the probabilities of rejecting the H_0 hypothesis for the KS test are greater than those for the mvnDH test. However, increasing the expected value of either variable (X or Y) enhances the likelihood of rejecting the H_0 hypothesis in the mvnDH analysis compared to the KS test. The performance of the analyzed tests is inconsistent, contingent upon the form of deviation from H_0. In some cases of deviation from multivariate normality, the mvnDH test has greater power than the KS test. Due to the discrete nature of the test statistic (number of empty cells), it is recommended that the researcher specify a larger number of cells.

The advantage of the proposed mvnDH test is its intuitive idea and the possibility of relatively easy programming of the relevant functions.

References

Anderson, T. W., Darling, D. A., (1952). Asymptotic Theory of Certain Nonparametric Tests, *Annals of Mathematical Statistics*, t. 23, 3, pp. 193–212.

Cramer, D., Howitt, D., (2004). *The Sage dictionary of statistics: A practical resource for students in the social sciences*. Sage Publications.

David, F. N., (1950). Order Statistics. J. Wiley & Sons Inc. New York.

Domański, C., (2009). Attempt to Assess Multivariate Normality Tests. *Acta Universitatis Lodziensis. Folia Oeconomica*, 225, pp. 76–90.

Domański, C., (2011). Własności testu Davida-Hellwiga. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 165, pp. 40–49.

Domański, C., (2012). Statistical Tests based on Empty Cells. *Acta Universitatis Lodziensis. Folia Oeconomica*, 269, pp. 39–47.

Domański Cz., Gadecki H. and Wagner W. (1989). Wartości krytyczne uogólnionego testu normalności Shapiro-Wilka. *Przegląd Statystyczny*, 36, pp. 107–112.

Domański, Cz., Pruska, K., (2000). Nieklasyczne metody statystyczne. *Polskie Wydawnictwo Ekonomiczne*, Warszawa.

Doornik, J. A., Hansen, H., (2008). An Omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics,* 70, pp. 927–939.

Ebner, B., Henze, N., (2020). Tests for Multivariate Normality – a Critical Review with Emphasis on Weighted L^2-Statistics. *TEST (2020)*, 29, pp. 845–892, https://doi.org/10.1007/s11749-020-00740-0

Hanusz, Z., Tarasińska, J. and Zieliński, W., (2012). Adaptation of Shapiro-Wilk Test to the Case of Known Mean. *Colloquium Biometricum*, 42, pp. 43–50.

Hellwig, Z., (1965). *Test zgodności dla małej próby*. Przegląd Statystyczny, 12, pp. 99–112.

Hernandez, H., (2021). *Testing for Normality: What is the Best Method?*

Henze, N., Zirkler, B., (1990). A Class of Invariant Consistent Tests for Multivariate Normality. *Commun. Statist.-Theor. Meth.*, 19(10), pp. 35953618.

Jarque, C. M., Bera, A. K., (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3), 1980, pp. 255–259.

Joenssen, D. W., Vogel, J., (2014). A power study of goodness-of-fit tests for multivariate normality implemented in R. *Journal of Statistical Computation and Simulation*, 84(5), pp. 1055–1078. https://doi.org/10.1080/00949655.2012.739620.

Kankainen, A, Taskinen, S. and Oja, H., (2007). Tests of multinormality based on location vectors and scatter matrices. *Statistical Methods & Applications,* 16, pp. 357–379.

Kesemen, O., Ti̇Ryaki̇, B. K., Tezel, Ö. and Özkul, E., (2021). A new goodness of fit test for multivariate normality. *Hacettepe Journal of Mathematics and Statistics*, 50(3), pp. 872–894.

Kołmogorov A., (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, pp. 83–91.

Kończak, G., (2005). On the Modification of David-Hellwig Test. [in:] *Innovation in classification, data science and information systems. Proceedings of the 27th Annual Conference of the Gesellschaft fur Klassifikation, Brandenburg University of Technology, Cottbus, March 12-14, 2003,* pp. 138–145.

Kończak, G., (2008). On the Multivariate Goodness-of-Fit test. Compstat. Proceedings in Computational Statistics. 18th Symposium held in Porto, Portugal. *Physica-Verlag*, pp. 751–758.

Korkmaz, S., Goksuluk, D. and Zararsiz, G., (2014). MVN: An R Package for Assessing Multivariate Normality. *The R Journal*, 6(2), 151. https://doi.org/10.32614/RJ-2014-031

Koziol, J. A., (1993). Probability Plots for Assessing Multivariate Normality. *The Statistician*, 42(2), 161. https://doi.org/10.2307/2348980.

Liang, J., Ng, K. W., (2009). A Multivariate Normal Plot to Detect Nonnormality. *Journal of Computational and Graphical Statistics*, 18(1), pp. 52–72. https://doi.org/10.1198/jcgs.2009.0004.

Liang, J., He, P. and Yang, J., (2022). Testing Multivariate Normality Based on t-Representative Points. *Axioms*, 11(11), p. 587. https://doi.org/10.3390/axioms11110587.

Lilliefors, H. W., (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318), pp. 399–402.

Lin, C.-C., Mudholkar, G. S., (1980). A simple test for normality against asymmetric alternatives. *Biometrika*, 67, pp. 455–61.

Malkovich, J. F., Afifi, A. A., (1973). On Tests for Multivariate Normality. *Journal of the American Statistical Association*, *68*(31), pp. 176–179.

Mardia, K. V., (1974). Applications of some measures of multivariate skewness and kurtosis for testing normality and robustness studies. *Sankhy A*, 36, pp. 115–128.

Mardia, K. V., (1975). Assessment of Multinormality and the Robustness of Hotelling's T^2 Test. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2), pp. 163–171.

Mardia, K. V., (2024). *Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies.*

McAssey, M. P., (2013). An empirical goodness-of-fit test for multivariate distributions. *Journal of Applied Statistics*, 40(5), pp. 1120–1131. https://doi.org/10.1080/02664763.2013.780160.

Móri, T. F., Székely, G. J. and Rizzo, M. L., (2021). On energy tests of normality, *Journal of Statistical Planning and Inference*, Vol. 213, pp. 1–15, https://doi.org/10.1016/j.jspi.2020.11.001.

Mudholkar, G. S., McDermo, M. and Srivastava, D. K., (2024). *A Test of p-Variate Normality.*

Mvtnorm, (2025). https://cran.r-project.org/web/packages/mvtnorm/mvtnorm.pdf [25.05.2025].

Rizzo M. L., Székely G. J., (2016). Energy Distance. *WIRES Computational Statistics*, *Wiley*, Vol. 8, Issue 1, pp. 27–38.

Romeu, J. L., Özturk, A., (1993). A Comparative Study of Goodness-of-Fit Tests for Multivariate Normality. *Journal of Multivariate Analysis*, 46(2), pp. 309–334. https://doi.org/10.1006/jmva.1993.1063.

Royston, J. P., (1982). An Extension of Shapiro and Wilks W Test for Normality to Large Samples. *Applied Statistics*, 31(2), p. 115124.

Royston, J. P., (1983). Some Techniques for Assessing Multivariate Normality Based on the Shapiro-Wilk W. *Applied Statistics*, 32(2).

Shapiro, S. S., Wilk, M. B., (1965). An Analysis of Variance Test for Normality. *Biometrika*, Vol. 52, No. 3/4, pp. 591–611.

Smirnow, N., (1939). Estimate of the goodness of fit for a continuous distribution. *Bulletin de l'Académie des Sciences de l'URSS. Série Mathématique*, 7, pp. 2–14.

Székely, G. J., Rizzo, M. L., (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference.*

Székely, G. J., Rizzo, M. L., (2017). The Energy of Data. *The Annual Review of Statistics and Its Application,* 4, pp. 447–79. https://doi.org/10.1146/annurev-statistics-060116-054026.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 167–183, https://doi.org/10.59139/stattrans-2025-045
Received – 16.12.2024; accepted – 04.06.2025

The concept of a behavioral model of decision-making under risk

Ewa Falkiewicz[1]

Abstract

This study outlines the decision-making process under risk considering the psychological aspects of the decision-maker. The aim is to construct a principle of an optimal decision for an individual decision-maker. The study considers a finite, discrete set of acceptable decisions, a set of possible world states and a system of probabilities of these states (where the probabilities are either known or subjectively estimated by the decision-maker), and a utility matrix of making each decision in particular world states. The proposed process of optimizing the decision addresses not only the rationality of the person making but also emotional aspects of the person making the decision. Rationality is represented by the value of the utility function of the benefits resulting from making a decision in a possible world state. The behavioral part of the model involves two emotions important to decision-making: regret over making a decision that brought less utility than possible in the given conditions and satisfaction with the choice which proved better than the worst option. The first emotion is represented by the regret function and the other by the satisfaction function. New notions are defined: relative utility and expected relative utility of particular decisions used to construct the principle of an optimal decision under risk.

The presented theory thus supplements the prospect theory, as it accounts for regret and satisfaction in the decision-making process under risk. This idea is part of behavioral economics, yet not standing in opposition to classical economics. Its advantage is that it considers both psychological and rational factors in the decision-making process.

Key words: utility function, prospect theory, decisions under risk, regret function, satisfaction function, expected relative utility.

1. Introduction

This paper is inspired by the prospect theory (Kahneman and Tversky, 1979; Jajuga, 2008), which is part of behavioral economics and states that the actions of an individual making economic decisions are guided not only by rationality, but also by psychological factors.

A decision-making model is examined in which one decision-maker desires to make a decision that is optimum under risk; that is, he knows the possible states of the outside world and the distribution of their probabilities (these probabilities can be understood in the classical sense and treated as objective) or does not know the probabilities of the particular world states but estimates them subjectively (these are subjective probabilities in the circumstances (Sadowski, 1981), a concept first introduced by Savage (1954)). Savage pointed

[1] Department of Business and International Finance, Faculty of Economics and Finance, Casimir Pulaski Radom University, Radom, Poland. E-mail: e.falkiewicz@urad.edu.pl ORCID: https://orcid.org/0000-0002-2263-9476.

out that the probability of certain events (e.g., wars, natural disasters, economic crises) is unknown in decision-making situations involving uncertainty. Subjective probability, introduced by Savage, expresses the degree of someone's conviction that a given event may take place (Tyszka et al., 2004). Here, subjective probability should be understood as the extent of belief that a given state will occur; although this degree, for a given state, in certain circumstances, may differ among various decision-makers.

Let the following be given:

- a set of acceptable decisions $M = \{d_1, \dots, d_k\}$, which, for the sake of simplicity, is discrete and finite,

- a set of external world states $S = \{s_1, \dots, s_n\}$ with a corresponding system of probabilities $p_1, \dots, p_n$, where $p_j = P(S = s_j)$ is the probability of the state s_j for $j = 1, \dots, n$, with $p_j > 0$ and $\sum_{j=1}^{n} p_j = 1$.

- a benefit function
$$k : M \times S \to \mathbf{R},$$
which assigns each possible pair (d_i, s_j) of a decision $d_i \in M$ and external world state $s_j \in S$ to a real number $x_{ij} \in \mathbf{R}$, to be known as the benefit of a decision d_i, $i = 1, \dots, k$ in the state s_j, $j = 1, \dots, n$:
$$k(d_i, s_j) = x_{ij}. \tag{1}$$

The values of the benefit function (1) can be represented by a matrix of benefits (Table 1). Let
$$K := k(M \times S)$$
be a set of benefit function values (1).

Table 1. A matrix of benefit function

Probabilities	p_1	p_2	$\dots$	p_j	$\dots$	p_n
Decisions/states	s_1	s_2	$\dots$	s_j	$\dots$	s_n
d_1	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1n}
d_2	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
d_i	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{in}
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
d_k	x_{k1}	x_{k2}	$\dots$	x_{kj}	$\dots$	x_{kn}

It is the decision-maker's aim to choose a decision that is optimal from his point of view.

The principle of maximizing the expected value (benefit), formulated by Pascal, is the oldest method for selecting an optimum decision under risk. For any decision d_i, $i = 1, \dots, k$, the expected value of benefits needs to be calculated, considering all the known states of the external world and their probabilities:

$$\mathrm{E}X_i = \sum_{j=1}^{n} x_{ij} p_j, \tag{2}$$

The decision is optimal for which the expected value of benefits formulated with (2) is maximum:

$$\max_i \mathrm{E}X_i. \tag{3}$$

Although the theory of maximizing expected value was the starting point for Markowitz's portfolio theory (Markowitz, 1952), it has faced criticism. Its weakness consists in the fact that it fails to address the risk or the decision-maker's subjective approach, for instance. It was already noted by Daniel Bernoulli, in 1738, in the so-called St. Petersburg paradox, in which the value of the expected payoff is infinite (the game involves flipping a coin and ends when tails are up, and the payoff is 2^n, where n is the number of tosses). Thus, someone guided by the principle of maximizing expected profit should pay an infinite sum to take part in the game, which sounds like a paradox. Bernoulli pointed out that people, in practice, are driven in their decisions not by the rule of maximizing the expected payoff, but by maximizing the expected utility, which he understood as the psychological value of money. Based on this observation, Bernoulli suggested a novel approach to evaluating bets. As Daniel Kahneman (2012) writes: "His idea was simple: human choices are based not on the monetary values of possible choices but on their psychological value, or utility. The psychological value of a bet is therefore not a weighted average of possible financial results, but an average of their utilities, where the utility of each result is weighted based on its probability":

$$\mathrm{E}u(X_i) = \sum_{j=1}^{n} u(x_{ij}) p_j, \tag{4}$$

where $u(x_{ij})$ is the utility of benefits from making the decision d_i given the world's state s_j. The optimum decision is the one for which the expected value of utility formulated as (4) is maximized:

$$\max_i \mathrm{E}u(X_i). \tag{5}$$

The utility of a certain amount of money is a subjective value that can vary among individuals. In order to arrive at a reliable decision-making criterion, therefore, one should be guided not only by the rule (5), but also take into account the value of standard deviation, for example, a measure of risk, or the coefficient of variation. The classic version of the theory of maximizing the expected utility is available in the monograph by von Neumann and Morgenstern (1947). New theories emerged over successive years of selecting optimum decisions under risk that began to consider psychological factors, for instance, Savage's theory of subjective expected utility (Savage, 1954) and the prospect theory (Kahneman and Tversky, 1979) – for which Daniel Kahneman was awarded the Nobel Prize in 2002 – are based on the assumption that human choices are unstable and context-dependent, chiefly on whether a decision is made when losses are sustained or profits are generated. The theory of regret, introduced independently by Bell (1982) and Loomes and Sugden (1982) and developed by Quiggin (1994), continues the concept of decision-making by addressing behavioral factors. Regret is expressed as the difference between the utility of a given decision's result and the utility of the best result of all decisions in a given situation (Zatoń, 2010).

2. The behavioral model of decision-making under risk considering regret and satisfaction

In this study, a method for choosing the optimum decision under risk is proposed. This method is an extension of the methods described in Chapter 1. It refers back to the theory of regret (Bell, 1982; Loomes and Sugden, 1982) by considering the role of this emotion in the decision-making process. It is also an expansion: an emotion contrary to regret – joy, satisfaction – is taken into account. In addition, the model uses the utility function that meets the assumptions of the prospect theory (Kahneman and Tversky, 1979). However, as Daniel Kahneman (2012) states: "The prospect theory and the theory of utility are unable to explain the phenomenon of regret. They both assume that all available options are evaluated separately and independently when making a decision; then, the option with the maximum value is selected. (...) such an assumption is certainly wrong."
Besides rationality, the considered model takes behavioral factors into account. Two strong emotions are analyzed from the viewpoint of decision-making: regret at making a decision that brings lesser utility than is possible in a given world state (this section refers back to the regret theory, Quiggin (1994))–and an opposite emotion–satisfaction with making a choice that is better than the worst in a given world state. Each emotion is represented by their respective functions of regret and satisfaction.

Let $u : K \to \mathbf{R}$ be the function of utility that, for each benefit $x_{ij} = k(d_i, s_j) \in K \subset \mathbf{R}$ of making decision d_i, $i = 1, \ldots, k$ in a world state s_j, $j = 1, \ldots, n$ (described by the function of benefit (1) and present in the matrix of benefits in Table 1), ascribes a certain subjective value, which is individual for each decision-maker:

$$x_{ij} \mapsto u(x_{ij}). \tag{6}$$

Since the function of utility depends on decision-makers' individual preferences, it may vary for different persons. It is assumed, nonetheless, that the function has certain characteristics shared by all decision-makers. Namely, the utility curve $u : K \to \mathbf{R}$ is assumed to meet the assumptions of the value function from the prospect theory (Kahneman and Tversky, 1979); that is, its shape is different for gains and for losses.

Assuming $0 \in K$ is the reference point for the decision-maker's gains and losses, the utility function can be described as follows:

1. $u \in C^2(K)$, i.e., it is continuous and has first and second order derivatives,

2. $\forall_{x \in K, x>0}\ u(-x) < 0 < u(x) \wedge u(0) = 0$ - the utility of losses is negative, that of gains is positive, and that of zero benefit is zero,

3. $\forall_{x_1, x_2 \in K}(x\ _1 < x_2 \to u(x_1) < u(x_2))$ - the utility is an increasing function,

4. $\forall_{x \in K, x>0}\ |u(-x)| > u(x)$ - the function is steeper for losses than for gains (a loss is more painful than a gain is enjoyable),

5. $\forall_{x \in K, x \leq 0}\ \frac{\partial u}{\partial x} > 0 \wedge \frac{\partial^2 u}{\partial x^2} > 0$ - the utility function for losses rises in a convex manner,

6. $\forall_{x\in K, x\geq 0}\ \frac{\partial u}{\partial x} > 0 \wedge \frac{\partial^2 u}{\partial x^2} < 0$ - the utility function for gains rises in a concave manner.

Where the sets of decisions M and external world states S are finite, the values of the utility functions (6) can be illustrated with a matrix of utility (Table 2), as in the case of the benefit function (1):

Table 2. Utility matrix

Probabilities	p_1	p_2	$\dots$	p_j	$\dots$	p_n
Decisions/states	s_1	s_2	$\dots$	s_j	$\dots$	s_n
d_1	$u(x_{11})$	$u(x_{12})$	$\dots$	$u(x_{1j})$	$\dots$	$u(x_{1n})$
d_2	$u(x_{21})$	$u(x_{22})$	$\dots$	$u(x_{2j})$	$\dots$	$u(x_{2n})$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
d_i	$u(x_{i1})$	$u(x_{i2})$	$\dots$	$u(x_{ij})$	$\dots$	$u(x_{in})$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
d_k	$u(x_{k1})$	$u(x_{k2})$	$\dots$	$u(x_{kj})$	$\dots$	$u(x_{kn})$

Definition 2.1 ***The general model for decision-making under risk (GMR)*** *is called the system (M,S,P,U), consisting of:*

- *the set $M = \{d_1,\dots,d_k\}$ of acceptable decisions, discrete and finite for the sake of simplicity,*
- *the set $S = \{s_1,\dots,s_n\}$ of external world states,*
- *the probability distribution $P = (p_1,\dots,p_n)$ of the external world states, where $p_j = P(s_j)$ for $j = 1,\dots,n$, with $\forall_{j=1,\dots,n} p_j > 0$ and $\sum_{j=1}^{n} p_j = 1$ (as part of the model, the probabilities p_j may be either objective or subjectively estimated by the decision-maker),*
- *the utility matrix $U = [u(x_{ij})]_{i=1,\dots,k,\ j=1,\dots,n}$, where $u(x_{ij})$ is the utility of making decision d_i, $i = 1,\dots,k$ in a world state s_j, $j = 1,\dots,n$, i.e. the value of utility function u that satisfies the above assumptions 1.-6. of the prospect theory.*

Remark 1 *Insofar as, there is precisely one benefit function (1) for all decision-makers for a given decision-making system (M,S,P,U), there can be a different utility function (6) for every decision-maker (the function of utility depends on the individual preferences of persons making decisions).*

It is the decision-maker's intention to choose an optimum decision in probable world states $s_1,\dots,s_n$.

Remark 2 *Note that the rows of the benefit matrix (Table 1) can be considered as the values of random variables $X_1,\dots,X_k$, where X_i is the benefit of making a decision d_i, $i = 1,\dots,k$. Therefore, for every $i = 1,\dots,k$, the distribution of random variable X_i is known (Table 3).*

Table 3. Benefits distribution

x_{ij}	x_{i1}	x_{i2}	$\dots$	x_{in}
$P(X_i = x_{ij})$	p_1	p_2	$\dots$	p_n

Since the utility function $u : K \to \mathbf{R}$ is continuous, from the theory of probability (Evans and Rosenthal, 2009), it is known that $u(X_i)$ is also a random variable for every $i = 1,\dots,k$. What is more, since u is a monotonous function, the probability of the particular benefits' utility is equal to the probabilities of these benefits, i.e.

$$P(u(X_i) = u(x_{ij})) = P(X_i = x_{ij}) = p_j \text{ for } i = 1,\dots,k,\ j = 1,\dots,n. \tag{7}$$

Therefore, the probability distribution of the random variable $u(X_i)$ denoting the utility of making i-th decision for every $i = 1,\dots,k$ is the same as that for the random variable X_i (Table 4).

Table 4. Utility distribution

$u(x_{ij})$	$u(x_{i1})$	$u(x_{i2})$	$\dots$	$u(x_{in})$
$P(u(X_i) = u(x_{ij}))$	p_1	p_2	$\dots$	p_n

The present model of decision-making under risk is designed to consider not only the rationality of a decision-maker but also psychological factors. The rational determinants are the values of the utility function (from the utility matrix, Table 2) for the specific gains in the subsequent world states. The behavioral determinants are two strong emotions associated with decision-making: regret at selecting a suboptimal result possible in a given world state, and satisfaction with choosing a result better than the worst in a given world state. Two functions are defined for describing emotions: one of regret and one of satisfaction. Let U be the set of values of the utility function $u : K \to \mathbf{R}$, i.e.

$$U := u(K). \tag{8}$$

Definition 2.2 *Let*

$$r : U \to \mathbf{R}_- \cup \{0\} \tag{9}$$

be a continuous function that attributes to the utility $u(x_{ij})$ of making i-th decision in j-th world state the difference between this utility and the maximum utility that can be derived from any decision in a known j-th world state:

$$r(u(x_{ij})) = u(x_{ij}) - \max_{1 \le i \le k} u(x_{ij}) := r_{ij}. \tag{10}$$

*The function r defined with (10) will be called **the regret function**.*

Note that the values of the regret function are not positive; they are negative if, in a given world state, the utility of gains from a decision under consideration is lower than the greatest utility of another decision. The value of this negative number serves as a measure of regret. They are zero if the utility of benefits from the considered decision is maximum in a given world state. In the case of a finite, discrete set of decisions and a finite, discrete set of world states, the values of the regret function for the particular decisions can be grouped into a matrix that will be referred to as ***the matrix of regret***:

$$M_r = [r_{ij}]_{i=1,\dots,k,\ j=1,\dots,n}. \tag{11}$$

Example 1 *Let a set of acceptable decisions $M = \{d_1, d_2, d_3, d_4\}$, a set of external world states $S = \{s_1, s_2, s_3\}$ and their corresponding probabilities $p_1 = 0.8, p_2 = 0.15, p_3 = 0.05$, and a matrix of utility of making the particular decisions in the subsequent world states be given:*

Table 5. *Utility matrix - example 1,2*

Probabilities	0.8	0.15	0.05
Decisions/states	s_1	s_2	s_3
d_1	*8*	*14*	*20*
d_2	*10*	*10*	*10*
d_3	*15*	*0*	*-40*
d_4	*-10*	*80*	*100*

Considering the first world state s_1, it is clear that the maximum utility of 15 can be attained in the case of the third decision d_3, namely

$$\max_{1 \leq i \leq 4} u(x_{i1}) = 15.$$

Thus, relying on equation (10), the first column of the regret matrix, or the value of regret in the first world state s_1, can be calculated as:

$$\begin{aligned} r_{11} &= u(x_{11}) - \max_{1 \leq i \leq 4} u(x_{i1}) = 8 - 15 = -7, \\ r_{21} &= u(x_{21}) - \max_{1 \leq i \leq 4} u(x_{i1}) = 10 - 15 = -5, \\ r_{31} &= u(x_{31}) - \max_{1 \leq i \leq 4} u(x_{i1}) = \ 15 - 15 = \ 0, \\ r_{41} &= u(x_{41}) - \max_{1 \leq i \leq 4} u(x_{i1}) = -10 - 15 = -25. \end{aligned} \tag{12}$$

Proceeding likewise for the second state s_2, with its maximum utility of 80 ($\max_{1 \leq i \leq 4} u(x_{i2}) = 80$) that can be reached for the decision d_4, the second column of the regret matrix is generated:

$$\begin{aligned} r_{12} &= u(x_{12}) - \max_{1 \leq i \leq 4} u(x_{i2}) = 14 - 80 = -66, \\ r_{22} &= u(x_{22}) - \max_{1 \leq i \leq 4} u(x_{i2}) = 10 - 80 = -70, \\ r_{32} &= u(x_{32}) - \max_{1 \leq i \leq 4} u(x_{i2}) = \ 0 - 80 = \ -80, \\ r_{42} &= u(x_{42}) - \max_{1 \leq i \leq 4} u(x_{i2}) = 80 - 80 = 0 \end{aligned} \tag{13}$$

and for the third state s_3, where the maximum utility of 100 ($\max_{1 \leq i \leq 4} u(x_{i3}) = 100$) can be reached for decision d_4, we generate the third column of the regret matrix:

$$\begin{aligned} r_{13} &= u(x_{13}) - \max_{1\leq i\leq 4} u(x_{i3}) = 20 - 100 = -80, \\ r_{23} &= u(x_{23}) - \max_{1\leq i\leq 4} u(x_{i3}) = 10 - 100 = -90, \\ r_{33} &= u(x_{33}) - \max_{1\leq i\leq 4} u(x_{i3}) = -40 - 100 = -140, \\ r_{43} &= u(x_{43}) - \max_{1\leq i\leq 4} u(x_{i3}) = 100 - 100 = 0. \end{aligned} \tag{14}$$

Substituting the values (12), (13), and (14) of the regret function into (11), the following regret matrix results:

$$M_r = \begin{bmatrix} -7 & -66 & -80 \\ -5 & -70 & -90 \\ 0 & -80 & -140 \\ -25 & 0 & 0 \end{bmatrix}. \tag{15}$$

The elements of the matrix M_r, the numerical values of the regret function, express the scale of regret at not choosing a better option. For instance, $r_{11} = -7$ is equivalent to the size of regret in the first world state relative to the decision d_1 producing the utility $u(x_{11}) = 8$, with the maximum possible utility in this world state being $\max_{1\leq i\leq 4} u(x_{i1}) = 15$. In the world state s_1 for the decision d_3, the value of regret is zero (i.e., this feeling is absent), since the value of utility is $u(x_{31}) = 15$; that is, it is the maximum possible utility to be achieved in this world state.

Besides regret, satisfaction with choosing a potentially better option than the worst in a given external world state is another strong emotion. It is contrary to the feeling of regret. Like in the case of regret, the level of satisfaction can be measured.

Definition 2.3 *Let*

$$s : U \to \mathbf{R}_+ \cup \{0\}, \tag{16}$$

be a continuous function that attributes to the utility $u(x_{ij})$ of making i-th decision in j-th world state the difference between this utility and the minimum possible value of the utility in the j-th world state:

$$s(u(x_{ij})) = u(x_{ij}) - \min_{1\leq i\leq k} u(x_{ij}) := s_{ij}. \tag{17}$$

The function s, defined with (17), will be called ***the satisfaction function****.*

Note that the values of the satisfaction function are non-negative; that is, they are either positive, where the utility of a decision is greater than the minimum utility of another decision (this positive number is then a measure of satisfaction) in a given world state, or zero if the utility of benefit from a given decision is the lowest in a given world state (satisfaction is absent in this case). In the case of a finite, discrete set of decisions and a finite, discrete set of external world states, the values of the satisfaction function for the particular decisions can be grouped into a matrix that will be referred to as ***the satisfaction matrix***:

$$M_s = [s_{ij}]_{i=1,\dots,k,\ j=1,\dots,n}. \tag{18}$$

Example 2 *As far as example 1 is concerned, satisfaction levels will be determined when considering a decision from the set of acceptable decisions $M = \{d_1, d_2, d_3, d_4\}$ given the set of external world states $S = \{s_1, s_2, s_3\}$, where the utilities of making the particular decisions in the successive world states are given in the utility matrix (Table 5).*

Since the utility of benefits from the decision d_4 is the minimum utility in the first world state s_1:

$$\min_{1\le i\le 4} u(x_{i1}) = -10,$$

relying on (17), the first column of the satisfaction matrix can be derived:

$$\begin{aligned} s_{11} &= u(x_{11}) - \min_{1\le i\le 4} u(x_{i1}) = 8-(-10) = 18,\\ s_{21} &= u(x_{21}) - \min_{1\le i\le 4} u(x_{i1}) = 10-(-10) = 20,\\ s_{31} &= u(x_{31}) - \min_{1\le i\le 4} u(x_{i1}) = 15-(-10) = 25,\\ s_{41} &= u(x_{41}) - \min_{1\le i\le 4} u(x_{i1}) = -10-(-10) = 0. \end{aligned} \tag{19}$$

For the second world state s_2, where utility is minimum for decision d_3 and equals zero:

$$\min_{1\le i\le 4} u(x_{i2}) = 0,$$

the second column of the satisfaction matrix is the same as the second column of the utility matrix (Table 5):

$$\begin{aligned} s_{12} &= u(x_{12}) - \min_{1\le i\le 4} u(x_{i2}) = 14,\\ s_{22} &= u(x_{22}) - \min_{1\le i\le 4} u(x_{i2}) = 10,\\ s_{32} &= u(x_{32}) - \min_{1\le i\le 4} u(x_{i2}) = \ 0,\\ s_{42} &= u(x_{42}) - \min_{1\le i\le 4} u(x_{i2}) = 80. \end{aligned} \tag{20}$$

For the third world state s_3, in which utility is minimum for decision d_3:

$$\min_{1\le i\le 4} u(x_{i3}) = -40,$$

the third column of the satisfaction matrix results:

$$\begin{aligned} s_{13} &= u(x_{13}) - \min_{1\leq i\leq 4} u(x_{i3}) = 20 - (-40) = 60, \\ s_{23} &= u(x_{23}) - \min_{1\leq i\leq 4} u(x_{i3}) = 10 - (-40) = 50, \\ s_{33} &= u(x_{33}) - \min_{1\leq i\leq 4} u(x_{i3}) = -40 - (-40) = 0, \\ s_{43} &= u(x_{43}) - \min_{1\leq i\leq 4} u(x_{i3}) = 100 - (-40) = 140. \end{aligned} \tag{21}$$

On substituting the calculated values (19), (20), and (21) of the satisfaction function into (18), the following satisfaction matrix is generated:

$$M_s = \begin{bmatrix} 18 & 14 & 60 \\ 20 & 10 & 50 \\ 25 & 0 & 0 \\ 0 & 80 & 140 \end{bmatrix}. \tag{22}$$

The elements of the matrix M_s, the numerical values of the satisfaction function, express the extent of satisfaction with choosing a better-than-the-worst option. For example, $s_{11} = 18$ is the scale of satisfaction in the first world state s_1 when considering the decision d_1 that produces the utility $u(x_{11}) = 8$, where the lowest possible utility in this world state is $\min_{1\leq i\leq 4} u(x_{i1}) = -10$. Satisfaction is zero in the world state s_1 for decision d_4, since utility is $u(x_{41}) = -10$, i.e., the lowest possible utility in this world state.

Some new notions are defined in the decision-making model introduced here: relative utility and expected relative utility.

Definition 2.4 ***Relative utility of i-th decision*** d_i ***in the j-th world state*** s_j *is the total sum of utility* $u(x_{ij})$ *and the values of regret function* $r(u(x_{ij}))$ *and the sum total* $s(u(x_{ij}))$ *with making the decision:*

$$u_w(x_{ij}) = u(x_{ij}) + \alpha \cdot r(u(x_{ij})) + (1-\alpha)\cdot s(u(x_{ij})), \tag{23}$$

where $0 \leq \alpha \leq 1$ *is loss-aversion factor,* $i = 1,\dots,k,\ j = 1,\dots,n$*. The expression* $u(x_{ij})$ *will be referred to as* ***the rational part*** *and* $\alpha \cdot r(u(x_{ij})) + (1-\alpha)\cdot s(u(x_{ij}))$ *-* ***as the behavioral part*** *of the relative utility of i-th decision in the j-th world state (or, to be more brief - if it does not give rise to misunderstandings - the rational and behavioral parts of the relative utility).*

Denoted by

$$\mathrm{Ra}(u_w(x_{ij})) := u(x_{ij}) \text{ and } \mathrm{Be}(u_w(x_{ij})) := \alpha \cdot r(u(x_{ij})) + (1-\alpha)\cdot s(u(x_{ij})), \tag{24}$$

relative utility can be expressed briefly as the sum of the rational and behavioral parts:

$$u_w(x_{ij}) = \mathrm{Ra}(u_w(x_{ij})) + \mathrm{Be}(u_w(x_{ij})). \tag{25}$$

The behavioral part of relative utility

$$\mathrm{Be}(u_w(x_{ij})) = \alpha \cdot r(u(x_{ij})) + (1-\alpha) \cdot s(u(x_{ij})), \tag{26}$$

for $0 \leq \alpha \leq 1$, is a convex combination of the values of regret and satisfaction function when making decision d_i in the world state s_j.

Remark 3 *The loss-aversion factor α, depending on its value in the range $[0;1]$, may enhance or weaken the values of regret and satisfaction functions. Thus, when considering a given decision-making variant in a specified world state:*

1. *if $0 \leq \alpha < 0.5$, the feeling of satisfaction prevails over regret in the decision-making process for the decision-maker,*
2. *if $0.5 < \alpha \leq 1$, the feeling of regret prevails over satisfaction,*
3. *if $\alpha = 0.5$, the feelings of regret and satisfaction are balanced when making decisions.*

Remark 4 *In some economic studies, the $\alpha : (1-\alpha)$ ratio is like 2:1 (Kahneman, 2012).*

Once the values of regret $r(u(x_{ij}))$ and satisfaction $s(u(x_{ij}))$ functions are replaced with the formulas from (10) and (17) in the equation (23) as appropriate, the relative utility of the i-th decision ($i = 1, \ldots, k$) in the j-th world state ($j = 1, \ldots, n$) becomes:

$$u_w(x_{ij}) = u(x_{ij}) + \alpha \cdot (u(x_{ij}) - \max_{1 \leq i \leq k} u(x_{ij})) + (1-\alpha) \cdot (u(x_{ij}) - \min_{1 \leq i \leq k} u(x_{ij})). \tag{27}$$

Relative utility is thus understood as utility relative to the maximum or minimum utility, with reference to the extent of regret or satisfaction in a given world state.

Transforming (27) produces yet another equivalent form of relative utility of i-th decision in the j-th world state:

$$u_w(x_{ij}) = 2u(x_{ij}) - \alpha \cdot \max_{1 \leq i \leq k} u(x_{ij}) - (1-\alpha) \cdot \min_{1 \leq i \leq k} u(x_{ij}). \tag{28}$$

Definition 2.5 *The system (M, S, P, U_w), where $U_w = [u_w(x_{ij})]_{i=1,\ldots,k,\ j=1,\ldots,n}$ is a relative utility matrix, will be called* ***the behavioral model of decision-making under risk*** *or shortly* ***BMR****.*

Based on definition 2.4 of relative utility of i-th decision in the j-th world state, another notion is constructed that is needed to build the optimization criterion of behavioral decisions:

Definition 2.6 ***Expected relative utility of i-th decision*** *d_i for $i = 1, \ldots, k$ is defined as:*

$$Eu_w(d_i) = \sum_{j=1}^{n} u_w(x_{ij}) \cdot p_j, \tag{29}$$

where p_j is the probability of the world state s_j, $j = 1, \ldots, n$.

Remark 5 *In the BMR model, the simple probabilities $p_1,\ldots,p_n$ can be replaced by the so-called weighting function (from the prospect theory, (Jajuga 2008)) transforming the probabilities:*

$$\pi : [0;1] \rightarrow [0;1],$$

$$\pi(p_j) = \frac{p_j^b}{\sum_{j=1}^n p_j^b},\ \ 0 \le b \le 1, \tag{30}$$

Kahneman and Tversky (1979) noted that low probabilities close to zero are normally inflated, whereas those high, approaching one, are depressed. The parameter b in (30) of the weighting function fulfills the role of subjectively distorting the probabilities:

1. *for $b = 0$, the weighting function (30) distorts the probabilities into a uniform distribution, i.e.*
$$\forall_{j=1,\ldots,n}\ \ \pi(p_j) = \frac{1}{n},$$
2. *for $b = 1$, the distribution remains unchanged, that is*
$$\forall_{j=1,\ldots,n}\ \ \pi(p_j) = p_j.$$

The criterion of decision optimization under risk in the BMR model:

Let the behavioral decision-making model under risk, described above, (M,S,P,U_w), be given. For any decision d_i, $i = 1,\ldots,k$, the expected relative utility can be calculated (eq. (29)). The decision d_0 for which the expected relative utility is maximum is the optimum decision:

$$E_0 = \max_{i=1,\ldots,k} Eu_w(d_i). \tag{31}$$

The foregoing criterion, besides rational factors represented in relative utility u_w by utility in its classic sense, also considers behavioral factors represented by the values of the regret and satisfaction functions. If, in equation (23), the relative utility of the i-th decision in the j-th world state described by (25) is adopted, it becomes clear that the optimum decision is d_0, for which

$$E_0 = \max_{i=1,\ldots,k} Eu_w(d_i) = \max_{i=1,\ldots,k} \sum_{j=1}^n u_w(x_{ij}) \cdot p_j = \max_{i=1,\ldots,k} \sum_{j=1}^n (\mathrm{Ra}(u_w(x_{ij})) + \mathrm{Be}(u_w(x_{ij}))) \cdot p_j. \tag{32}$$

Remark 6 *If there exist two different decisions $d_l \neq d_m$, $l,m = 1,\ldots,k$, for which*

$$E_0 = \max_{i=1,\ldots,k} Eu_w(d_i) = Eu_w(d_l) = Eu_w(d_m), \tag{33}$$

then, in order to decide which decision is optimal for the decision-maker, the standard deviations for both decisions should be calculated and, using their analysis, the optimal decision should be selected. The standard deviation for i-th decision is understood here as the *standard deviation from the relative expected value:*

$$su_w(d_i) = \sqrt{\sum_{j=1}^{n} (u_w(x_{ij}) - Eu_w(d_i))^2 p_j}. \tag{34}$$

3. BMR model in the matrix form

The behavioral decision-making model under risk described above, addressing the emotions of regret and satisfaction, can be expressed in an equivalent matrix form.
Let (M,S,P,U_w) be the behavioral decision-making model under risk (BMR), where $M = \{d_1,\ldots,d_k\}$ is a set of acceptable decisions, $S = \{s_1,\ldots,s_n\}$ a set of external world states with a corresponding system of probabilities $p_1,\ldots,p_n$, where $p_j = P(s_j)$, $j = 1,\ldots,n$, and $\sum_{j=1}^{n} p_j = 1$ are given. Let $u : K \to \mathbf{R}$ be the utility function for making a decision d_i, $i = 1,\ldots,k$ in the world state s_j, $j = 1,\ldots,n$, with properties 1. - 6. referring to Kahneman's and Tversky's prospect theory. Assuming that, for any benefit $x_{ij} \in K$ of making the i-th decision, $i = 1,\ldots,k$, in the j-th world state, $j = 1,\ldots,n$, the benefit's utility is formulated as

$$u(x_{ij}) = u_{ij}, \tag{35}$$

a utility matrix can be created:

$$M = [u_{ij}]_{i=1,\ldots,k,\ j=1,\ldots,n} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & & \vdots \\ u_{k1} & u_{k2} & \cdots & u_{kn} \end{bmatrix} \in M_{k\times n}(\mathbf{R}). \tag{36}$$

On defining the regret function $r : U \to \mathbf{R}_- \cup \{0\}$ by means of (10), where U is the set of utility functions, the values of regret r_{ij} at making the i-th decision, $i = 1,\ldots,k$, in the j-th world state, $j = 1,\ldots,n$, are inputs to the regret matrix described with (11), expanded as follows:

$$M_r = [r(u_{ij})]_{i=1,\ldots,k,\ j=1,\ldots,n} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{k1} & r_{k2} & \cdots & r_{kn} \end{bmatrix} \in M_{k\times n}(\mathbf{R}). \tag{37}$$

By analogy, when defining the satisfaction function $s : U \to \mathbf{R}_+ \cup \{0\}$ using the formula (17), the values of satisfaction s_{ij} resulting from the i-th decision $i = 1,\ldots,k$ in the j-th world state $j = 1,\ldots,n$ are inputs to the satisfaction matrix described by (18), which is expanded as follows:

$$M_s = [s(u_{ij})]_{i=1,\ldots,k,\ j=1,\ldots,n} = \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1n} \\ s_{21} & s_{22} & \ldots & s_{2n} \\ \vdots & \vdots & & \vdots \\ s_{k1} & s_{k2} & \ldots & s_{kn} \end{bmatrix} \in M_{k\times n}(\mathbf{R}). \tag{38}$$

Considering definition 2.4 of the relative utility of the i-th decision d_i, in the j-th world state s_j and (23), ***the matrix of relative utility*** is given by:

$$U_w = [u_w(x_{ij})]_{i=1,\ldots,k,\ j=1,\ldots,n} = M + \alpha\cdot M_r + (1-\alpha)\cdot M_s \in M_{k\times n}(\mathbf{R}), \tag{39}$$

where $0 \leq \alpha \leq 1$. In (39), the matrix M accounts for the rational part of the decision, whereas the convex combination of the matrices M_r and M_s, i.e., $\alpha\cdot M_r + (1-\alpha)\cdot M_s$, accounts for the behavioral portion. On developing (39), the matrix of relative utility becomes:

$$\begin{aligned} U_w &= \begin{bmatrix} u_w(x_{11}) & u_w(x_{12}) & \ldots & u_w(x_{1n}) \\ u_w(x_{21}) & u_w(x_{22}) & \ldots & u_w(x_{2n}) \\ \vdots & \vdots & & \vdots \\ u_w(x_{k1}) & u_w(x_{k2}) & \ldots & u_w(x_{kn}) \end{bmatrix} = \\ &= \begin{bmatrix} u_{11} & u_{12} & \ldots & u_{1n} \\ u_{21} & u_{22} & \ldots & u_{2n} \\ \vdots & \vdots & & \vdots \\ u_{k1} & u_{k2} & \ldots & u_{kn} \end{bmatrix} + \alpha\cdot \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ r_{21} & r_{22} & \ldots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{k1} & r_{k2} & \ldots & r_{kn} \end{bmatrix} + \\ &\quad + (1-\alpha)\cdot \begin{bmatrix} s_{11} & s_{12} & \ldots & s_{1n} \\ s_{21} & s_{22} & \ldots & s_{2n} \\ \vdots & \vdots & & \vdots \\ s_{k1} & s_{k2} & \ldots & s_{kn} \end{bmatrix}. \end{aligned} \tag{40}$$

Based further on definition 2.6 of the expected relative utility of the i-th decision d_i in the j-th world state s_j, and using (29), ***the vector of the expected relative utilities of decisions*** $d_1,\ldots,d_k$ is defined:

$$E_w = [Eu_w(d_i)]_{i=1,\ldots,k} = U_w\cdot p = (M + \alpha\cdot M_r + (1-\alpha)\cdot M_s)\cdot p \in \mathbf{R}^k, \tag{41}$$

or in its expanded form:

$$E_w = \begin{bmatrix} Eu_w(d_1) \\ Eu_w(d_2) \\ \vdots \\ Eu_w(d_k) \end{bmatrix}, \tag{42}$$

where $p \in \mathbf{R}^n$ is the vector of the probabilities $p_1, \ldots, p_n$ of the world states $s_1, \ldots, s_n$:

$$p = [p_j]_{j=1,\ldots,n} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}. \tag{43}$$

The optimum decision d_0 is signaled by the maximum coordinate of the vector of the expected relative utilities, i.e.:

$$E_0 = \max_{i=1,\ldots,k} Eu_w(d_i). \tag{44}$$

Example 3 *With reference to examples 1 and 2, the optimum decision will be determined for the decision-maker considering the choice of a decision from the set $M = \{d_1, d_2, d_3, d_4\}$ in the external world states from $S = \{s_1, s_2, s_3\}$, whose utilities from making the particular decision in the successive world states are given in the matrix of utilities (Table 5).*

The matrix of relative utilities U_w derived from (39), where the matrices of regret M_r and satisfaction M_s are (15) and (22), respectively, is:

$$U_w = M + \alpha \cdot M_r + (1-\alpha) \cdot M_s =$$

$$= \begin{bmatrix} 8 & 14 & 20 \\ 10 & 10 & 10 \\ 15 & 0 & -40 \\ -10 & 80 & 100 \end{bmatrix} + \alpha \cdot \begin{bmatrix} -7 & -66 & -80 \\ -5 & -70 & -90 \\ 0 & -80 & -140 \\ -25 & 0 & 0 \end{bmatrix} + (1-\alpha) \cdot \begin{bmatrix} 18 & 14 & 60 \\ 20 & 10 & 50 \\ 25 & 0 & 0 \\ 0 & 80 & 140 \end{bmatrix},$$

where $0 \leq \alpha \leq 1$. Depending on the value of α, various matrices of relative utilities may be generated. Taking, for instance, $\alpha = 0.66$, which means that according to the classification in Remark 3, the feeling of regret prevails over satisfaction (2:1), the following matrix of relative utilities results:

$$U_w = \begin{bmatrix} 9.5 & -24.8 & -12.4 \\ 13.5 & -32.8 & -32.4 \\ 23.5 & -52.8 & -132.4 \\ -26.5 & 107.2 & 147.6 \end{bmatrix}.$$

Further, relying on (41), the vector of expected subjective relative utilities is computed:

$$E_w = U_w \cdot p = \begin{bmatrix} 9.5 & -24.8 & -12.4 \\ 13.5 & -32.8 & -32.4 \\ 23.5 & -52.8 & -132.4 \\ -26.5 & 107.2 & 147.6 \end{bmatrix} \cdot \begin{bmatrix} 0.8 \\ 0.15 \\ 0.05 \end{bmatrix} = \begin{bmatrix} 3.26 \\ 4.26 \\ 4.26 \\ 2.26 \end{bmatrix}.$$

There are two coordinates of the maximum value 4.26, corresponding to decisions d_2 and d_3:

$$Eu_w(d_2) = Eu_w(d_3) = 4.26.$$

Thus, using the formula (34) from Remark 6, we calculate the standard deviations from the relative expected values for decisions d_2 and d_3:

$$su_w(d_2) = \sqrt{\sum_{j=1}^{3}(u_w(x_{2j}) - 4.26)^2 p_j} = \sqrt{341.5164} = 18.48,$$
$$su_w(d_3) = \sqrt{\sum_{j=1}^{3}(u_w(x_{3j}) - 4.26)^2 p_j} = \sqrt{1718.316} = 41.45.$$

Comparing the standard deviations from the relative expected values of 4.26, it can be concluded that, from the perspective of a risk-averse decision-maker, decision d_2 is optimum because the standard deviation is lower. However, for a risk-taking decision-maker, decision d_3 appears to be optimal; due to the higher standard deviation, a higher payout is possible.

4. Conclusions

The BMR behavioral model of decision-making under risk demonstrates the method of selecting an optimum decision for the decision-maker wishing to choose the single best decision from a set of multiple acceptable decisions in a simple and clear manner. Introducing the concept of relative utility allows for considering not only rational factors, such as the utility of benefits, but also behavioral factors that address two opposing emotions important to decision-making: regret at a choice that is worse than the best possible option and satisfaction with a choice that is better than the worst in a given world state. This method goes beyond the framework of the rational choice theory. It does not reject rationality, although the relative utility introduced here is the total sum of the rational and behavioral parts (25). It emphasizes the fact that human choices are a complex interplay of rational and emotional elements. The study draws inspiration from the prospect theory (the utility function, which treats profits and losses differently, and the weighting function that transforms probabilities). Considering two strong emotions: regret, expressed by the values of the regret function, and satisfaction, expressed by the values of the satisfaction function, is an important addition to the prospect theory. These results are a starting-point for further development of the rational-behavioral theory of decision-making.

References

Bell, D. E., (1982). Regret in Decision Making under Uncertainty. *Operations Research*, 30, pp. 961–981.

Bernoulli, D., (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica*, 22, pp. 23–36.

Evans, M. J., Rosenthal J. S., (2009). *Probability and Statistics: The Science of Uncertainty*, second edition, W.H. Freeman, New York.

Jajuga, K., (2008). Trzydzieści lat współczesnych finansów behawioralnych. *Studia i prace Wydziału Nauk Ekonomicznych i Zarządzania*, 9, pp. 42–52.

Kahneman, D., (2012). *Pułapki myślenia. O myśleniu szybkim i wolnym*, Media Rodzina, Poznań.

Kahneman, D., Tversky, A., (1979). Prospect Theory: an Analysis of Decision under Risk. *Econometrica*, 47, pp. 263–291.

Loomes, G., Sugden, R., (1982). Regret Theory: an Alternative Theory of Rational Choice under Uncertainty. *Economic Journal*, 92, pp. 805–824.

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

Neuman von, J., Morgenstern, O., (1947). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.

Quiggin, J., (1994). Regret Theory with General Choice Sets. *Journal of Risk and Uncertainty*, 8, pp. 153–165.

Sadowski, W., (1981). *Decyzje i prognozy*, Państwowe Wydawnictwo Ekonomiczne, Warszawa.

Savage, L., (1954). *The Foundations of Statistics*, John Wiley annd Sons, New York.

Tyszka, T., (red.), (2004). *Psychologia ekonomiczna*, Gdańskie Wydawnictwo Psychologiczne, Gdańsk.

Zatoń, W., (2010). Żal i inne aspekty psychologiczne podejmowania decyzji finansowych na przykładzie problemu opcji walutowych w przedsiębiorstwach w Polsce w 2008 r. *Prace naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Prognozowanie w zarządzaniu firmą*, 103, pp. 274–284.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 185–199, https://doi.org/10.59139/stattrans-2025-046
Received – 08.08.2023; accepted – 25.02.2025

Mean estimation based on the factor-type estimator under an adaptive cluster sampling design

Narendra Singh Thakur[1], Shubhangi Chaurasia[2], Unnati Bhayare[3]

Abstract

If a sample is designated by a standard sampling strategy and if the character of the study satisfies a predetermined statement for an independent unit in the sample, then the items in the locality remain automatically in the sample. This type of method of selection of sampling units is called adaptive cluster sampling. This manuscript emphasizes the use of the factor-type estimator designed for population mean of the variable under study using the data of highly correlated auxiliary (supplementary) variable under adaptive cluster sampling. The bias, mean squared error and optimum mean squared errors up to the first order is obtained and a simulation study is performed for comparison purpose.

Key words: adaptive cluster sampling (ACS), ratio estimator, factor-type estimator, auxiliary variable, bias, mean squared error (MSE).

Mathematical Subject Code: 62D05.

1. Introduction

Thompson (1990) introduced an innovative sampling scheme called adaptive sampling, which directly incorporates the knowledge of the study variable into the selection process. This approach is distinct from traditional sampling strategies that rely solely on predetermined sampling plans. The adaptive sampling scheme was proposed to address situations where the study variable exhibits certain patterns or characteristics that can inform the sampling process. For instance, in surveys involving rare species, researchers may gather information on the number of individuals with specific characteristics. Frequently, zero abundance is encountered, but when substantial abundance is observed, it suggests that additional clusters of abundance might be found in nearby

[1] Department of Statistics, Govt. Model Girls College, Sheopur, MP, India. E-mail: nst_stats@yahoo.co.in. ORCID: https://orcid.org/0000-0001-9731-058X.

[2] Department of Mathematics and Statistics, SMS Govt. Model Science College, Gwalior, MP, India. E-mail: shubhangichaurasia22101989@gmail.com. ORCID: https://orcid.org/0009-0007-2399-7746.

[3] Department of Statistics, Govt. (Model, Autonomous) Holkar Science College, Indore, MP, India. E-mail: unnati.b80@gmail.com. ORCID: https://orcid.org/0009-0001-5561-0303.

locations. This pattern is not limited to rare species but can be observed in various domains such as whales, insects, trees, lichens, and more. The conventional approach in sample surveys involves deciding on a sampling strategy before data collection begins. However, this predetermined approach may not always be effective, especially in certain scenarios. For instance, in epidemiological studies of infectious diseases, encountering a diseased individual suggests a higher-than-expected incidence rate among nearby individuals. In such cases, ground staff may deviate on or after the predesignated selection plan and then combine adjacent or closely allied items to the sample.

Thompson's (1990) adaptive sampling scheme addresses this need for flexibility in sampling. It starts with drawing a preliminary sample of a predetermined size using a standard sampling strategy. The values of the sampled items are then examined, and if an elected item fulfils a specified condition, supplementary items are put in to the sample from the locality of that item. Thus, adaptive process allows for the expansion of the sample based on specific criteria or patterns observed in the study variable. The design of the adaptive sampling scheme is demonstrated in Figure 1(a) and 1(b), which likely provide visual representations of how the sampling process unfolds.

The Figure 1(a) illustrates the preliminary sampling stage of the adaptive sampling scheme. A sample of 12 units is selected using a probability sampling procedure, which could be any conventional sampling design. The key feature of adaptive sampling is that when one or more units in the preliminary sample satisfy a specific criterion associated to the variable under study, accompanying units from locality of those selected units are included in the sample. The neighborhood is typically defined based on spatial proximity, as indicated by the connected units on the left, right, top, and bottom in Figure 1(a).

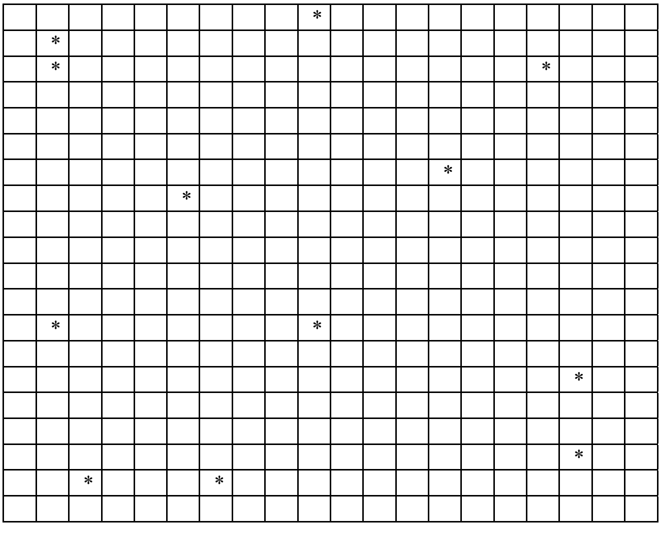


Figure 1(a). Preliminary sample of 12 units

After the adaptive procedure is finished, the sample contains 54 units, as revealed in Figure 1(b), where the symbol $\underline{*}$ represents the unit selected in preliminary sample of size 12. It should be noted that the concept of neighborhood is not limited to spatial proximity and is able to be express in numerous aspects subject to the condition and the nature of the study. In summary, in the adaptive sampling scheme, a preliminary subgroup of some units is selected using a probability sampling technique and if the variable of interest for a carefully chosen item satisfies a given criterion, subsequent units from the locality of that item are considered in the sample. This adaptive approach allows for the expansion of the sample based on specific conditions or patterns observed in the variable of interest.

			*					*	*	$\underline{*}$									
	$\underline{*}$	*	*	*					*										
	$\underline{*}$	*	*	*												$\underline{*}$			
			*																
													*						
					*							*	$\underline{*}$	*					
				*	$\underline{*}$	*							*	*					
					*														
		*																	
		*							*										
	$\underline{*}$	*						*	$\underline{*}$	*									
								*	*	*									
									*	*							$\underline{*}$		
																*			
		*				*	*									*	$\underline{*}$	*	
		$\underline{*}$				$\underline{*}$	*	*							*				
		*					*												

Figure 1(b). Adaptive cluster sample of 54 units

In the adaptive selection system, the criterion for picking additional neighboring items can be defined in various ways, depending on the nature of the study. One approach is to frame the criterion as an interval L that covers a specific range of values related to the variable of curiosity. If a unit is considered in the sample, it should meet the criteria fixed by the interval L. Mathematically, it can be represented as if $i \in L$ then the unit $i \in S$.

The provided definitions lay the foundation for understanding the structure and components of the sampling process in the adaptive cluster sampling scheme. Let's elaborate on each definition:

1) **Neighborhood of a unit:** The neighborhood of a unit i refers to a group of items that contains unit i. These neighborhoods are determined on the basis of design and selection process and are independent of the population values.

2) **Cluster:** A cluster is the assembly of all units that are detected as per an outcome of the preliminary choice of a specific item i. In ACS, the preliminary selection of a unit (seed unit) leads to the insertion of entire units in the corresponding cluster in the final sample. It is possible for a cluster to consist of the union of several neighborhoods, which means that multiple neighborhoods can be grouped together as part of the same cluster. The concept of clusters is important for understanding the unit selection process and how ACS samples are formed.
3) **Network**: It is a set of items where the inclusion of any item in the preliminary sample from that set ensures the inclusion of all units in that network in the final sample. In other words, if a single unit from a network is selected, the entire network becomes part of the sample. It is worth noting that units not satisfying the condition L are also considered network, but they consist of a single unit only. Networks play a significant role in adaptive sampling, where certain networks may be oversampled to improve estimation efficiency.
4) **Edge unit:** An edge unit is a population item that does not satisfy the network requirements but is in the neighborhood of an item that satisfies the condition L. Essentially, edge units are on the boundary of clusters or networks. They play a crucial role in ACS because their selection may influence the inclusion of entire clusters or networks in the final sample.

The estimators taken into consideration as a relationship between neighborhoods, clusters, networks, and edge units to obtain a reliable and efficient population estimate and to make a valid statistical inference under adaptive cluster sampling. For further study some useful and valuable contributions for readers are advise as Chao (2004), Chutiman and Kumphon (2008), Dryver and Thompson (1998), Pochai (2008) etc. This manuscript is concerned with to develop the factor-type estimator in adaptive cluster sampling and discuss its properties. Furthermore, how we were motivating for writing this manuscript is explained in Section 3 of this paper.

1.1. Notations

Let y be the variable under study based on a population U and let it consists of a set of N units indexed by their labels $U = \{1, 2, \dots N\}$. The population mean of y is $\bar{Y} = \mu_y = \frac{1}{N}\sum_{\mathrm{i}=1}^{N} y_i$. Let $\bar{Y}_{\mathrm{ac}} = \bar{w}_y = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} w_{yi}$ be the estimate of the mean in adaptive cluster sampling. Let us consider:

n = Size of preliminary sample,

A_i = Network which consists of the item i,

m_i = The amount of the items in the network to which i^{th} item belongs.

Let w_{yi} and w_{xi} represent the mean of y and the mean of x in the network which consist of unit i, viz., $w_{yi} = \frac{1}{m_i}\sum_{j\epsilon A_i} y_j$ and $w_{xi} = \frac{1}{m_i}\sum_{j\epsilon A_i} x_j$ respectively. According to

Dryver and Chao (2007) adaptive cluster sampling is considered as SRSWOR when the means of networks are considered under study. Let us use the notations $\bar{w}_y$ and $\bar{w}_x$ to denote the sample means of the study and supplementary variables in the transformed sample respectively. We calculate $\bar{w}_y$ and $\bar{w}_x$ as

$$\bar{w}_y = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} w_{yi} \text{ and } \bar{w}_x = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} w_{xi}$$

For simplicity we write,

$$s_{wy}^2 = \frac{1}{n-1}\sum_{\mathrm{i}=1}^{n}(w_{yi} - \bar{w}_y)^2,\ s_{wx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(w_{xi} - \bar{w}_x)^2$$

are unbiased estimators of

$$S_{wy}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(w_{yi} - \bar{Y})^2,\ S_{wx}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(w_{xi} - \bar{X})^2 \quad \text{respectively,}$$

and $\mu_{rs} = \frac{1}{N}\sum_{i=1}^{N}\left(w_{yi} - \bar{Y}\right)^r (w_{xi} - \bar{X})^s$; where r, s are positive integers.

Also, $C_{wy}^2 = \frac{S_{wy}^2}{\bar{Y}^2}$ and $C_{wx}^2 = \frac{S_{wx}^2}{\bar{X}^2}$ are the coefficient of variations of w_y and w_x respectively,

and $\rho_{wyx} = \frac{\mu_{11}}{S_{wy}S_{wx}}$ is the coefficient of correlation between w_y and w_x.

Using the concept of large sample approximations, let $\varepsilon = \frac{\bar{w}_y}{\bar{Y}} - 1$ and $\eta = \frac{\bar{w}_x}{\bar{X}} - 1$, for specified $\bar{w}_y$, $\bar{w}_x$ respectively, then

$$E(\varepsilon) = E(\eta) = 0,\ E(\varepsilon^2) = \frac{C_{wY}^2}{n},\ E(\eta^2) = \frac{C_{wX}^2}{n},\ E(\varepsilon\eta) = \rho_{wyx}C_{wY}C_{wX}$$

and $E(\varepsilon^i\eta^j) = 0$ if $i + j > 2;\ i, j = 0, 1, 2, \dots$.

The expectations as derived above under the concept of large sample approximations will be used for further mathematical treatments.

Remark: Assume, $\alpha = \frac{1}{1+C_{wx}^*}$; $\omega = \frac{\beta_2(w_x)}{\beta_2(w_x)+ C_{wx}^*}$; $\delta = \frac{1}{1+ \beta_2^*\,(w_x)}$;

$$\theta_1 = \frac{fB}{A+fB+C};\ \theta_2 = \frac{C}{A+fB+C};\ P = (\theta_1 - \theta_2);\ V = \rho_{wyx}\frac{C_{wy}}{C_{wx}}.$$

The above symbols will be used for further algebraic treatments.

2. Existing estimators and their characteristics in adaptive cluster sampling

Some existing estimators under adaptive cluster sampling are considered in this section. Note that, the constants $C_{wx}^* = \frac{C_{wx}}{\bar{X}}$, $\beta_2^*(w_x) = \frac{\beta_2(w_x)}{\bar{X}}$, $u_{wx} = \frac{\bar{w}_x}{\bar{x}}$ and $\bar{X} = \mu_x = \frac{1}{N}\sum_{\mathrm{i}=1}^{N} x_i$, C_{wx}, $\beta_2(w_x)$ of the auxiliary variable are known in advance by past experience.

The unbiased Hansen and Hurwitz estimator for the population mean of main variable is given by [see, Thompson (1990), Thompson and Seber (1996)]

$$\bar{Y}_{ac} = \frac{1}{N}\sum_{i=1}^{N}(w_y)_i = \bar{w}_y \tag{2.1}$$

where, $(w_y)_i$ is mean of the main variable in the network that contains item i of the preliminary sample.

The variance of $\bar{Y}_{ac}$ is

$$V(\bar{Y}_{ac}) = \frac{(N-n)}{Nn(N-1)} \sum_{i=1}^{N} [(w_y)_i - \mu_y]^2 \quad = \frac{N-n}{Nn} S_{wy}^2 \tag{2.2}$$

The ratio type estimator under adaptive cluster sampling strategy projected by Dryver and Chao (2007) as

$$\bar{Y}_{Rac} = \bar{w}_y \frac{\bar{X}}{\bar{w}_x} = \hat{R}_{ac} \bar{X} \tag{2.3}$$

where, $\hat{R}_{ac} = \frac{\bar{w}_y}{\bar{w}_x}$. The estimator $\bar{Y}_{R_{ac}}$ in relationships of ε and η can approximate up to the first order and is expressed as

$$\bar{Y}_{R_{ac}} = \bar{Y}[1 + \varepsilon - \eta - \varepsilon\eta + \eta^2]$$

This estimator is biased and its bias up to the first order is

$$B(\bar{Y}_{R_{ac}}) = \frac{\bar{Y}}{n} \left[C_{wx}^2 - \rho_{wyx} C_{wy} C_{wx}\right] \tag{2.4}$$

The expression of MSE is

$$M(\bar{Y}_{R_{ac}}) = \frac{\bar{Y}^2}{n} \left[C_{wy}^2 + C_{wx}^2 - 2\rho_{wyx} C_{wy} C_{wx}\right] \tag{2.5}$$

Chutiman (2013) suggested some estimators in adaptive cluster sampling described as below

(A)

$$\bar{Y}_{R_{ac2}} = \bar{w}_y \left(\frac{1 + C_{wx}^*}{u_{wx} + C_{wx}^*}\right) \tag{2.6}$$

and the estimator $\bar{Y}_{R_{ac1}}$ in relationships of ε and η approximate up to the first order, can be expressed as

$$\bar{Y}_{Rac1} = \bar{Y}[1 + \varepsilon - \alpha\eta - \alpha\varepsilon\eta + \alpha^2\eta^2]$$

This estimator is biased and its bias up to the first order is

$$B\left(\bar{Y}_{R_{ac2}}\right) = \frac{\bar{Y}}{n} \left[\alpha^2 C_{wx}^2 - \alpha\rho_{wyx} C_{wy} C_{wx}\right] \tag{2.7}$$

The equation of MSE is

$$M\left(\bar{Y}_{R_{ac1}}\right) = \frac{\bar{Y}^2}{n} \left[C_{wy}^2 + \alpha^2 C_{wx}^2 - 2\alpha\rho_{wyx} C_{wy} C_{wx}\right] \tag{2.8}$$

(B)

$$\bar{Y}_{R_{ac2}} = \bar{w}_y \left[\frac{\beta_2(w_x) + C_{wx}^*}{\beta_2(w_x) u_{wx} + C_{wx}^*}\right] \tag{2.9}$$

and the estimator $\bar{Y}_{R_{ac2}}$ in relationships of ε and η approximate up to the first order, can be expressed as

$$\bar{Y}_{R_{ac2}} = \bar{Y}\,[1 + \varepsilon - \omega\eta - \omega\varepsilon\eta + \omega^2\eta^2]$$

This estimator is biased and its bias up to the first order is

$$B\left(\bar{Y}_{R_{ac2}}\right) = \frac{\bar{Y}}{n} [\omega^2 C_{wx}^2 - \omega\rho_{wyx} C_{wy} C_{wx}] \tag{2.10}$$

The equation of MSE is

$$M\left(\bar{Y}_{R_{ac2}}\right) = \frac{\bar{Y}^2}{n} \,[C_{wy}^2 + \omega^2 C_{wx}^2 - 2\omega\rho_{wyx} C_{wy} C_{wx}] \tag{2.11}$$

(C)
$$\bar{Y}_{R_{ac3}} = \bar{w}_y \left[\frac{1+ \beta_2^* (w_x)}{u_{wx}+ \beta_2^* (w_x)}\right] \tag{2.12}$$

and the estimator $\bar{Y}_{R_{ac3}}$ in relationships of ε and η, approximate up to the first order, can be expressed as

$$\bar{Y}_{R_{ac3}} = \bar{Y} \left[1 + \varepsilon - \delta\eta - \delta\varepsilon\eta + \delta^2\eta^2\right]$$

This estimator is biased and its bias up to the first order is

$$B\left(\bar{Y}_{R_{ac3}}\right) = \frac{\bar{Y}}{n}\left[\delta^2 C_{wx}^2 - \delta\rho_{wyx}C_{wy}C_{wx}\right] \tag{2.13}$$

The equation of MSE is

$$M\left(\bar{Y}_{R_{ac3}}\right) = \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + \delta^2 C_{wx}^2 - 2\delta\rho_{wyx}C_{wy}C_{wx}\right] \tag{2.14}$$

The proofs of the above expressions are simple and readers can obtain them similar manner to that described in Section 4 for the proposed factor-type estimator using large sample approximations.

3. Proposed Estimator in adaptive cluster sampling

Singh and Shukla (1987) proposed a factor-type (F-T) estimator for estimating population mean and Singh and Shukla (1993) derived an efficient factor-type estimator family for estimating the similar population mean. Shukla (2002) suggested factor-type estimator for estimation in two-phase sampling. Also, Shukla and Thakur (2008), Thakur and Shukla (2022) developed factor-type estimator as a device of imputation used for dealing with missingness of the data.

Deriving motivation from all of these, we advocate the modified factor-type estimator for adaptive cluster sampling is

$$\bar{Y}_{FTc} = \bar{w}_y \frac{(A+C)\bar{X}+fB\bar{w}_x}{(A+fB)\bar{X}+C\bar{w}_x} \tag{3.1}$$

where, $A = (k-1)(k-2);\ B = (k-1)(k-4);$

$C = (k-2)(k-3)(k-4);\ f = \frac{n}{N}$ and $0 < k < \infty$ is a constant.

The estimator $\bar{Y}_{FTc}$ is biased and the expressions of bias, MSE and optimum MSE up to the first order of approximations are obtained ahead in Section 4. Theoretical and numerical comparisons of different estimators as discussed earlier, with $\bar{Y}_{FTc}$ is presented in Section 5 and Section 7 respectively.

For some specified values of k, the estimator $\bar{Y}_{FTc}$ provides some well-known estimators like – ratio, product, dual to ratio and unbiased unit mean estimator for population mean, i.e. at k = 1, 2, 3 and 4, the estimator $\bar{Y}_{FTc}$ is as special case in the following Table 3.1.

Table 3.1. Adaptive factor-type estimator as special cases

Value of *k*	**Estimators**	**Value of *k***	**Estimators**
$k = 1$	$\bar{y}_{FTc} = \bar{w}_y \frac{\bar{X}}{\bar{w}_x}$	$k = 2$	$\bar{y}_{FTc} = \bar{w}_y \frac{\bar{w}_x}{\bar{X}}$
$k = 3$	$\bar{y}_{FTc} = \bar{w}_y \frac{N\bar{X}-n\bar{w}_x}{(N-n)\,\bar{X}}$	$k = 4$	$\bar{y}_{FTc} = \bar{w}_y$

For $k = 1$ the estimator $\bar{y}_{FTc}$ provides the ratio estimator for mean, for $k = 2$ the proposed estimator is termed in product estimator, for $k = 3$ the estimator $\bar{y}_{FTc}$ is converted as dual to ratio estimator and for $k = 4$ the factor of auxiliary information vanishes and the estimator $\bar{y}_{FTc}$ is the same as unbiased unit mean estimator in adaptive cluster sampling.

4. Bias, MSE and optimum MSE of the proposed estimator

Let $B(\hat{\theta})$, $M(\hat{\theta})$ and $M(\hat{\theta})_{min}$ represents the bias, MSE and minimum MSE of the estimator θ. Further, the equations of bias, MSE and minimum MSE of the proposed estimators in terms of population parameters and other constants (as available) up to the first order are represented in the subsequent theorems using the concept of large sample approximations.

Theorem 4.1. The estimator $\bar{Y}_{FTc}$ in terms of ε and η up to the first order, could be stated as

$$\bar{Y}_{FTc} = \bar{Y}[1 + \varepsilon + P(\eta + \varepsilon\eta - \eta^2\theta_2)] \tag{4.1}$$

Proof: From equation (3.1), we have

$$\bar{Y}_{FTc} = \bar{w}_y \; \frac{(A+C)\bar{X}+fB\bar{w}_x}{(A+fB)\bar{X}+C\bar{w}_x}$$

and by using the concept of large sample approximation as discussed in Section 1.1

$$\begin{aligned}\bar{Y}_{FTc} &= \bar{\mathrm{Y}}\,(1+\varepsilon)\left[\frac{(A+C)\bar{X} + fB\bar{X}(1+\eta)}{(A+fB)\bar{X} + C\bar{X}\,(1+\eta)}\right]\\ &= \bar{Y}(1+\varepsilon)(1+\theta_1\eta)\left(1+\theta_2{}^{-1}\right)\\ &= \bar{Y}(1+\varepsilon)(1+\theta_1\eta)(1+\theta_2\eta+\theta_2^2\eta^2-\cdots)\\ &= \bar{Y}[1+\varepsilon+P(\eta+\varepsilon\eta-\eta^2\theta_2)]\end{aligned}$$

Theorem 4.2. Bias of $\bar{Y}_{FTc}$ in the relationships of population parameters is

$$B(\bar{Y}_{FTc}) = -\frac{\bar{Y}}{n} P[\theta_2 C_{wx}^2 + \rho_{wyx} C_{wy} C_{wx}] \tag{4.2}$$

Proof: $B(\bar{Y}_{FTc}) = E(\bar{Y}_{FTc} - \bar{Y})$

$$= -\frac{\bar{Y}}{n} P\left[\theta_2 C_{wx}^2 - \rho_{wyx} C_{wy} C_{wx}\right]$$

Theorem 4.3. MSE of $\bar{Y}_{FTc}$ in the relationships of population parameters is

$$M(\bar{Y}_{FTc}) = \frac{\bar{Y}^2}{n}\,[C_{wy}^2 + P^2\, C_{wx}^2 + 2P\rho_{wyx} C_{wy} C_{wx}] \tag{4.3}$$

Proof: $M(\bar{Y}_{FTc}) = E(\bar{Y}_{FTc} - \bar{Y})^2$

$$= \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + P^2 C_{wx}^2 + 2P\rho_{wyx}C_{wy}C_{wx}\right]$$

Theorem 4.4. The minimum MSE of $(\bar{Y}_{FTc})$, when $P = -V$ is

$$M(\bar{Y}_{FTc})_{min} = \frac{S_{wy}^2}{n}(1 - \rho_{wyx}^2) \tag{4.4}$$

Proof: Minimum MSE occurs when

$$\frac{d}{dP}M(\bar{Y}_{FTc}) = 0$$

or

$$2PC_{wx}^2 + 2\rho_{wyx}C_{wy}C_{wx} = 0$$

the optimal condition is

$$P = -\rho_{wYX}\frac{C_{wy}}{C_{wx}} = -V \quad \text{(let)} \tag{4.5}$$

Hence,

$$\mathrm{M}(\bar{Y}_{FTc})_{min} = \left(1 - \rho_{wyx}^2\right)\frac{S_{wy}^2}{n}$$

By simplifying the optimality condition $P = -\rho_{wYX}\frac{C_{wy}}{C_{wx}} = -V$, we will get a cubic equation in terms of k and the roots of this equation will provide us best choice of parameter k for minimum mean squared error with lowest bias.

4.1. Bias control estimator $\bar{Y}_{FTc}$

The condition of optimality provides from equation (4.5)

$$AV + (V + 1)fB + (V - 1)C = 0 \tag{4.6}$$

The equation (4.6) is an equation of degree 3 in terms of k.

Obviously, at most three values of k (k_1, k_2, k_3) are possible for which MSE is optimum.

The choice criteria for best estimation is:

1) Compute

$$\left|B(\bar{Y}_{FTc})_{k_j}\right| \text{ for } j = 1, 2, 3$$

2) From computed values, choose k_j as

$$\left|B(\bar{Y}_{FTc})_{k_j}\right| = min\left[\left|B(\bar{Y}_{FTc})_{k_j}\right|\right]; j = 1, 2, 3.$$

So, it is clear that the estimator $\bar{Y}_{FTc}$ is bias control for the optimum MSE.

5. Comparisons

This section compares the proposed estimator $\bar{y}_{FTc}$ with existing estimators as discussed in Section 2 of this manuscript. Let $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$,…$\hat{\theta}_k$ be k estimators of the population parameter θ and there exist an estimator $\hat{\theta}$ of the population parameter θ such that

$$\forall\, i, \quad \text{var}(\hat{\theta}) < \text{var}(\hat{\theta}_i), i = 1, 2, 3, \dots, k$$

i.e., variance of $\hat{\theta}$ is minimum among all existing estimators $\hat{\theta}_i$, i =1, 2, 3,…, k, then $\hat{\theta}$ is best estimator of the population parameter θ.

The theoretical comparison between the existing and proposed estimators, has been performed in this section, and the conditions of better performance of $\bar{Y}_{FTc}$ have been derived.

[A]: The variance of $\bar{Y}_{ac}$ in SRSWOR is given by

$$V(\bar{Y}_{ac}) = \frac{N-n}{Nn} S_{wy}^2$$

and the MSE of $\bar{Y}_{FTc}$ is

$$M(\bar{Y}_{FTc})_{min} = \frac{S_{wy}^2}{n}(1 - \rho_{wyx}^2)$$

Now, let

$$D_1 = V(\bar{Y}_{ac}) - M(\bar{Y}_{FTc})_{min}$$
$$= \frac{N-n}{Nn} S_{wy}^2 - \frac{S_{wy}^2}{n}(1 - \rho_{wyx}^2)$$

$\bar{Y}_{FTc}$ is better than $\bar{Y}_{ac}$ if $D_1 > 0$

i.e. $$\frac{N-n}{Nn} S_{wy}^2 - \frac{S_{wy}^2}{n}(1 - \rho_{wyx}^2) > 0$$
$$\frac{N-n}{N} > (1 - \rho_{wyx}^2)$$
$$n < N\rho_{wyx}^2$$

If the condition $n < N\rho_{wyx}^2$ holds, then $\bar{Y}_{FTc}$ is always better than $\bar{Y}_{ac}$.

[B]: The mean squared error of $\bar{Y}_{R_{ac}}$ is

$$\text{MSE}(\bar{Y}_{R_{ac}}) = \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + C_{wx}^2 - 2\rho_{wyx}C_{wy}C_{wx}\right]$$

Let $D_2 = \left[M(\bar{Y}_{R_{ac}}) - M(\bar{Y}_{FTc})_{min}\right]$

$$= \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + C_{wx}^2 - 2\rho_{wyx}C_{wy}C_{wx}\right] - \frac{\bar{Y}^2}{n}\left[C_{wy}^2 - C_{wy}^2\rho_{wyx}^2\right]$$

$\bar{Y}_{FTc}$ is better than $\bar{Y}_{R_{ac}}$, if $D_2 > 0$, i.e. $\rho_{wyx} < \frac{C_{wx}}{C_{wy}}$.

If the above condition satisfies then $\bar{Y}_{FTc}$ is better than $\bar{Y}_{Rac}$.

[C]: The mean squared error of $\bar{Y}_{R_{ac1}}$ is

$$M(\bar{Y}_{R_{ac1}}) = \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + \alpha^2 C_{wx}^2 - 2\alpha\rho_{wyx}C_{wy}C_{wx}\right]$$

Let $D_3 = \left[M(\bar{Y}_{R_{ac1}}) - M(\bar{Y}_{FTc})_{min}\right]$

$$= \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + \alpha^2 C_{wx}^2 - 2\alpha\rho_{wyx}C_{wy}C_{wx}\right] - \frac{\bar{Y}^2}{n}\left[C_{wy}^2 - C_{wy}^2\rho_{wyx}^2\right]$$

$\bar{Y}_{FTc}$ is better than $\bar{Y}_{R_{ac1}}$, if $D_3 > 0$

i.e. $\rho_{wyx} < \alpha\frac{C_{wx}}{C_{wy}}$

If the above condition satisfies then $\bar{Y}_{FTc}$ is better than $\bar{Y}_{R_{ac1}}$.

[D]: The mean squared error of $\bar{Y}_{R_{ac2}}$ is

$$M(\bar{Y}_{R_{ac2}}) = \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + \omega^2 C_{wx}^2 - 2\omega\rho_{wyx}c_{wy}c_{wx}\right]$$

Let $D_4 = \left[M\left(\bar{Y}_{R_{ac2}}\right) - M(\bar{Y}_{FTc})_{min}\right]$

$$= \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + \omega^2 C_{wx}^2 - 2\omega\rho_{wyx}c_{wy}c_{wx}\right] - \frac{\bar{Y}^2}{n}\left[C_{wy}^2 - C_{wy}^2\rho_{wyx}^2\right]$$

$\bar{Y}_{FTc}$ is better than $\bar{Y}_{R_{ac2}}$, if $D_4 > 0$

i.e. $\rho_{wyx} < \omega\frac{C_{wx}}{C_{wy}}$

If above condition satisfies then $\bar{Y}_{FTc}$ is better than $\bar{Y}_{R_{ac2}}$.

[E]: The mean squared error of $\bar{Y}_{R_{ac3}}$ is

$$M(\bar{Y}_{R_{ac3}}) = \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + \delta^2 C_{wx}^2 - 2\delta\rho_{wyx}C_{wy}C_{wx}\right]$$

Let $D_5 = \left[M\left(\bar{Y}_{R_{ac2}}\right) - M(\bar{Y}_{FTc})_{min}\right]$

$$= \frac{\bar{Y}^2}{n}\left[C_{wy}^2 + \delta^2 C_{wx}^2 - 2\delta\rho_{wyx}C_{wy}C_{wx}\right] - \frac{\bar{Y}^2}{n}\left[C_{wy}^2 - C_{wy}^2\rho_{wyx}^2\right]$$

$\bar{Y}_{FTc}$ is better than $\bar{Y}_{R_{ac3}}$, if $D_5 > 0$

i.e. $\rho_{wyx} < \delta\frac{C_{wx}}{C_{wy}}$

If the above condition satisfies then $\bar{Y}_{FTc}$ is better than $\bar{Y}_{R_{ac3}}$.

6. Empirical study

Appendix A displays an engendered simulated population covering amounts of y and x respectively. Summary of population is calculated as

$N = 400$	$\bar{Y} = 1.2275$	$\bar{X} = 0.56500$	$S_{wy}^2 = 12.6791$	$S_{wx}^2 = 3.79790$
$C_{wy} = 2.9008$	$C_{wx} = 3.44920$	$\rho_{wxy} = 0.80710$	$\beta_2(w_x) = 92.63470$	
$\beta_2(w_y) = 23.0357$	$S_{wxy} = 5.60070$	$V = 0.67877$		

Taking random sample of size 5, 10, 20 and 40 by SRSWOR and, by solving optimum condition (4.6) i.e., the equation of degree 3 in terms of k we got three k-values for different sample sizes as shown below in Table 6.1.

Table 6.1. Values of k for different sample sizes

Sample Size	k_1	k_2	k_3
$n = 5$	7.1827013	1.8390406	2.1566182
$n = 10$	7.2310521	1.7829198	2.2299408
$n = 20$	7.3283849	1.7091097	2.3370733
$n = 40$	7.5273670	1.6156457	2.4928638

Table 6.1 reveals that for sample size n = 5, by simplifying the optimality condition (4.6) we obtained a cubic equation of variable k and using the available constants the three values of k are k_1, k_2 and k_3. A similar procedure is used for sample sizes 10, 20 and 40.

7. Simulation study

In this section, we conducted a simulation using the population data of appendix A. The population was visualized through appendix A, and data were provided to describe the population. The simulation process involved the following method.

A preliminary sample of n units was carefully chosen by SRSWOR. Once the preliminary sample was selected, the y-values and x-values were obtained for each unit in the sample. The criterion for inclusion of items in the sample is $L = \{y: y > 0\}$. Every estimator was obtained for 5,000 iterations. The estimators were then applied to these samples to estimate population parameters of interest. By repeating this process, the study aimed to obtain accuracy estimates for the estimators.

The determination of this simulation is likely to assess the presentation of different estimators under adaptive sampling (ACS) scheme using various preliminary sample sizes. The bias is calculated by the formula

$$B(\hat{y}) = \frac{1}{5000} \sum_{i=1}^{5000}[(\hat{y}_{\mathrm{i}}) - \bar{Y}] \tag{7.1}$$

and the mean squared error is calculated by the formula

$$M(\hat{y}) = \frac{1}{5000} \sum_{i=1}^{5000}[(\hat{y}_{\mathrm{i}}) - \bar{Y}]^2 \tag{7.2}$$

The size of the preliminary sample is considered as n = 5, 10, 20 and 40 and repeated 5,000 times for each and every sample size n.

Table 7.1. Bias and MSE of Existing and Proposed Estimators

$\hat{y}$	Bias ($\hat{y}$)				MSE ($\hat{y}$)			
	n = 5	n = 10	n = 20	n = 40	n = 5	n = 10	n = 20	n = 40
$\bar{Y}_{ac}$	0	0	0	0	2.53582	1.26791	0.63395	0.31697
$\bar{Y}_{R_{ac}}$	-0.25727	-0.24622	-0.25663	-0.27817	0.35221	0.26310	0.12222	0.11167
$\bar{Y}_{R_{ac1}}$	1.61925	1.78461	1.89083	1.88892	4.87287	4.41820	4.18421	4.04436
$\bar{Y}_{R_{ac2}}$	3.94233	3.85695	3.87888	3.88283	30.22787	21.38539	18.33143	17.80784
$\bar{Y}_{R_{ac3}}$	-0.21504	-0.19859	-0.20757	-0.28289	0.34885	0.24523	0.10370	0.11321
$\bar{Y}_{FTc_1}$	0.03349	0.08942	0.11107	0.12835	0.32494	0.23383	0.10168	0.07645
$\bar{Y}_{FTc_2}$	0.21346	0.30079	0.36296	0.41172	0.48638	0.39157	0.25440	0.27200
$\bar{Y}_{FTc_3}$	-0.34290	-0.54661	-0.78513	-1.56488	0.57626	0.63471	0.91916	3.16843

By observing Table 7.1 the proposed estimator $\bar{Y}_{FTc}$ has minimum mean squared error for $k = k_1$ and minimum bias as well. The proposed estimator is better over all the estimators under consideration and the estimator $\bar{Y}_{FTc_1}$ is best overall.

8. Discussion and conclusion

In the present manuscript some estimators are discussed with their properties using the concept of large sample approximations in adaptive cluster sampling [see Chutiman (2013)] and discussed about factor-type estimator of Singh and Shukla (1987), Shukla and Thakur (2008), etc. Then raising idea from these all we experimented on the factor-type estimator under the same sampling design and found that the factor-type estimator of Singh and Shukla (1987) performed excellently overall in the adaptive cluster sampling design. The bias and mean squared error (*MSE*) of factor-type estimator are obtained up to the first order in terms of population parameters. The condition of optimality is derived as well.

From the results of simulation (table 7.1), it is clear that modified ratio estimators in ACS design have more bias and mean squared error as compared to the factor-type estimator at optimum value of k, i.e. $\bar{Y}_{FTc_1}$, $\bar{Y}_{FTc_2}$ and $\bar{Y}_{FTc_3}$. Also, it is proved that the proposed estimator is closer to the true value of average cases. The proposed estimator $\bar{Y}_{FTc_1}$ results in the lowest MSE as compared to all the estimators considered in this article. This proves that the proposed factor-type estimator has a greater efficiency than all the estimators under consideration.

Acknowledgement

We are thankful for the editorial team of the journal Statistics in Transition and the reviewers of this article for their efforts, comments and suggestions to make this article more valuable.

References

Chao, C. T., (2004). Ratio estimation on adaptive cluster sampling. *Journal of Chinese Statistical Association*, 42, pp. 307–327.

Chutiman, N., Kumphon, B., (2008). Ratio estimator using two auxiliary variables for adaptive cluster sampling. *Thailand Statistician*, 6(2), pp. 241–256.

Chutiman, N., (2013). Adaptive cluster sampling using auxiliary variable. *Journal of Mathematics and Statistics*, 9(3), pp. 249–255.

Dryver, A. L., Thompson, S. K., (1998). Improving unbiased estimators in adaptive cluster sampling. *ASA Proceedings of the Section of Survey Research Methods*, pp. 727–731.

Dryver, A. L., Chao C., (2007). *Ratio estimators in adaptive cluster sampling. Environmetrics*, 18, pp. 607–620. https://doi.org/10.1002/env.838.

Pochai, N., (2008). Ratio estimator using two auxiliary variables for adaptive cluster sampling. *J. Thai Statist. Assoc.*, 6, pp. 241–256.

Shukla, D., (2002). F-T estimator under two-phase sampling. *Metron*, 59, 1–2, pp. 253–263.

Shukla, D., Thakur, N. S., (2008). Estimation of mean with imputation of missing data using factor-type estimator. *Statistics in Transition*, 9(1), pp. 33–48.

Singh, V. K., Shukla, D., (1987). One parameter family of factor-type ratio estimator. *Metron*, 45, 1-2, pp. 273–283.

Singh, V. K., Shukla, D., (1993). An efficient one parameter family of factor-type estimator in sample survey. *Metron*, 51, 1-2, pp. 139–159.

Thompson, S. K., (1990). Adaptive cluster sampling. J. Am. Statist. Assoc., 85, pp. 1050-1059. DOI: 10.1080/01621459.1990.10474975.

Thompson, S. K., Seber, G. A. F., (1996). Adaptive Sampling. 1st Edn., Wiley, New York, ISBN-10:0471558710, p. 265.

Thakur, N. S., Shukla, D., (2022). Missing data estimation based on the chaining technique in survey sampling. *Statistics in Transition*, 23(4), pp. 91–111. https://doi.org/10.2478/stattrans-2022-0044.

Appendix A

Population for Empirical Study

Observations of Study Variable (Y)

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	2	24	5	0	0	0	0	0	0	0	0	0	0	0
0	0	1	22	5	4	5	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	4	8	0	0	0	0	0	33	0	0	0	27	0	0
0	0	0	0	0	0	0	0	0	0	0	0	7	6	7	1	0	5	0	0
0	0	0	0	0	0	0	21	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	4	0	5	7	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	5	7	0	7	7	6	3	0	0	0	0	0	0
0	0	0	5	4	3	0	5	8	4	5	1	0	5	0	0	0	0	0	0
0	7	65	0	4	5	0	9	0	0	0	0	0	3	1	0	0	0	0	0
0	1	4	5	0	7	3	3	0	0	0	0	0	0	0	0	0	0	0	0
0	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	21
0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Corresponding Observations of Auxiliary Variable (X)

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	11	3	0	0	0	0	0	0	0	0	0	0	0
0	0	0	11	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	2	4	0	0	0	0	0	12	0	0	0	15	0	0
0	0	0	0	0	0	0	0	0	0	0	0	3	2	3	0	0	2	0	0
0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	0	2	3	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	2	2	0	3	3	2	1	0	0	0	0	0	0
0	0	0	2	2	1	0	2	3	2	2	0	0	2	0	0	0	0	0	0
0	3	18	0	2	2	0	4	0	0	0	0	0	1	0	0	0	0	0	0
0	0	2	2	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	12
0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 201–204

About the Authors

Agu Friday I. is a doctoral researcher at the Mathematical Institute of the Slovak Academy of Sciences in Slovakia. He also serves as an expert and researcher in the Department of Sensory Information Systems and Technologies at the Institute of Informatics, Slovak Academy of Sciences. His research interests include probability distributions, copula models and dependence measures, statistical machine learning for cybersecurity and sensor intelligence, and risk and insurance claims modeling. He is a registered member of several professional bodies, including the American Statistical Association (ASA), Institute of Mathematical Statistics (IMS), Chartered Institute of Statisticians of Nigeria (CISON), and the Professional Statistician Society of Nigeria (PSSN). Friday has received multiple scholarships and research grants and has authored over 20 publications in reputable journals and conference proceedings. He continues to advance research.

Bhayare Unnati is an Assistant Professor at the Department of Statistics, Govt. Holkar (Model, Autonomous) Science College, Indore, Madhya Pradesh, Bharat. Simultaneously she holds the position of Head in the Department of Statistics at the College. She has over 12 years of experience in teaching and research. Her main areas of interest include: sampling theory, statistical quality control and inference. She has published a number of research articles in different reputed journals.

Bohlourihajjar Soghra is a statistician with expertise in survival analysis, Bayesian statistical modeling, and advanced data analytics in public health research. She has earned her PhD in Statistics from Razi University, where her doctoral work focused on Bayesian nonparametric survival analysis and developing and evaluating Bayesian methods, particularly in epidemiological applications. Soghra Bohlourihajjar has worked as a statistical analyst at the Department of Public Health, Faculty of Health Sciences, Bielefeld University, where she collaborated on interdisciplinary research involving health inequalities, epidemiological data analysis, and evidence-based public health. Her broader research interests include small-area estimation, reproducible statistical workflows, and the use of machine learning techniques in biomedical data. She is committed to connecting innovative statistical methodology with practical applications to support informed decision-making in public health.

Chaurasia Shubhangi is a Research Scholar of the Department of Mathematics and Statistics, Jiwaji University, Madhya Pradesh, Bharat. She is doing her research under the supervision of Dr. Narendra Singh Thakur. Her main areas of interest include:

sampling theory, missing data, imputation methods and reliability. She has published a number of research articles in different reputed journals.

Falkiewicz Ewa is an Assistant Professor at the Department of Business and International Finance, Faculty of Economics and Finance, Casimir Pulaski Radom University. Simultaneously she is an Assistant Professor at the Institute of Mathematics and Cryptology, Faculty of Cybernetics, Military University of Technology. Her main areas of interest include: the application of mathematical methods in economic sciences, in particular in behavioral finance, decision-making theory, and methods of differential geometry.

Girul Agata, PhD in Economics, specializes in statistics and data analysis, applying advanced statistical and mathematical methods to social and economic research. She focuses her work on improving the quality of life for people with special needs, such as older adults and individuals with disabilities, by identifying their unmet needs and developing reliable assessment tools. She also pursues research interests in sustainable development. She has many years of professional experience in official statistics, developing and interpreting socio-economic data. Dr Girul is also involved in social and educational initiatives that promote science and the development of analytical skills, as well as interdisciplinary collaboration.

Khazaei Soleiman is an Associate Professor of Statistics in the Faculty of Science at Razi University. He also serves as the Secretary of the Committee for Supervision, Evaluation, and Quality Assurance of Higher Education in Kermanshah Province. His research interests include Bayesian nonparametrics, simulation-based density estimation, survival analysis, and statistical inference. He has published more than twenty scientific papers and has translated a scholarly book on machine learning. He teaches a range of courses at the undergraduate, master's, and PhD levels, including Probability, Nonparametric Methods, and Inferential Statistics.

Kończak Grzegorz is a Full Professor at the Department of Statistics, Econometrics and Mathematics at the University of Economics in Katowice, Poland. His research interests include statistical inference, data analysis, Monte Carlo studies, and permutation tests, in particular. Professor Kończak has published more than 100 research papers in international/national journals and conferences. He has also published twelve books/monographs as an author or co-author. Professor Kończak is an active member of the Polish Statistical Association.

Kubus Mariusz is an Assistant Professor at the Centre of Education and Mathematics Applications, Opole University of Technology, Poland. The main areas of his current scientific interest are supervised learning and time series, in statistical and machine learning approach. The research focuses mainly on the problem of feature selection, dimensionality reduction and classification of imbalanced data.

Mach Łukasz has received his PhD in Economics and Finance from the Warsaw School of Economics, Poland. He is a professor, researcher and lecturer at the Faculty of Economics, University of Opole, Poland. His research interests include quantitative methods in economic research, time series analysis, modelling of the residential real estate market and analysis of its cyclicality.

Misiurski Przemysław is EngD in Economic Sciences, an Assistant Professor at the Faculty of Economics and Management of the Opole University of Technology, currently at the Department of Enterprise Management, E-business, and Electronic Economy. The main area of his scientific and research interests are issues in the field of transport economics, concerning modern concepts and strategies for developing transport in Poland and the world. In recent years, the author's research projects' main goal is to focus on the efficiency of rolling stock investments in public transport companies.

Nymphas E. F. is an Associate Professor in the Department of Physics, University of Ibadan. His research areas of interests are atmospheric modelling, micrometeorology, radio communication, and surface energy fluxes. Dr Nymphas has published many research papers in international/national journals and conferences. He also has two chapters in two books. Dr Nymphas is an active member of many scientific professional bodies.

Ogunde A. A. is a lecturer in the Biometry Unit, Department of Statistics, University of Ibadan. His research interests are demography, mathematical statistics, statistical inference and data analysis, in particular. He has published more than 35 research papers in scholarly international/national journals and conferences. Dr A. A. Ogunde is an active member of many scientific professional bodies.

Swanson David A. is Distinguished Professor Emeritus, University of California Riverside. He served as a member of the U.S. Census Bureau's Scientific Advisory Committee for six years (2004-10) and has produced 135 refereed sole- and co-authored journal articles and nine books, mainly dealing with demography. He also has edited or co-edited five additional books. Among other professional recognitions, he is a member of the Washington State Academy of Sciences, a Fellow of the Mississippi Academy of Sciences, and served as a "summer at census" scholar at the U.S. Census Bureau in 2019. He is also recognized as one of the world's leading demographers. His PhD was earned at the University of Hawai'i and he holds a Graduate Diploma in Social Sciences from the University of Stockholm.

Tayman Jeff retired as the Director of Technical Service for the San Diego Associations of Governments. He was also a lecturer in the Department of Economics, University of California San Diego for 20 years, 2003–2023. He has sole- and co-authored more than 30 refereed journal articles, four books, and 10 chapters in books and conference proceedings. He is a four-time recipient of the Southern Demographic Association's "Terrie Award," which recognizes the best paper presented at the SDA Annual Meeting on an applied topic, especially one relating to state and local demography. His PhD was earned at Florida State University.

Thakur Narendra Singh is an Assistant Professor of Statistics in the Department of Higher Education, Govt. of Madhya Pradesh. He has been serving at Govt. Model Girls College, Sheopur, Madhya Pradesh, Bharat since 2019. Previously, he served as an Assistant Professor (Statistics) at the Department of Mathematics and Statistics, Banasthali University, Rajasthan around eight years. His research interests are statistical analysis, sampling theory, probability, measurement errors, imputation methods, in particular. Dr. Thakur has published around 45 research papers in international/national journals and conferences. He has also published 2 books. Dr. Thakur is an active member of many scientific professional bodies and a member of editorial boards of reputed journals.

Wesołowska Monika is a PhD candidate in Economics at the Department of Macroeconomics and Development Research at the Poznań University of Economics and Business. Her research interests include income inequality, especially in Central and Eastern Europe, as well as demographic and migration-related challenges.

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 205–210

Acknowledgements to Reviewers

The Editor and Editorial Board of Statistics in Transition new series wish to thank the following persons who served as peer-reviewers of manuscripts for the *Statistics in Transition new series* – Volume 26, Numbers 1–4. The authors' work has benefited from their feedback.

Abdullahi, Umar, Kabir, Department of Statistics, University of Nigeria, Nigeria

Abu-Shawiesh, Moustafa, Omar, Ahmed, Department of Mathematics, Hashemite University, Jordan

Acha, Chigozie K., Department of Statistics, Michael Okpara University of Agriculture Umudike, Nigeria

Adekeye, Kayode, Samuel, Department of Mathematical Sciences, Redeemer's University, Nigeria

Ahmad, Hanan, Haj, Department of Statistics, Hail University, Saudi Arabia

Akarsu, Mahmut, Zeki, Faculty of Economic Sciences, University of Warsaw, Poland

Al Ameri, Mohammed, Department of Civil Engineering, Monash University, Australia

Alziadeh, Morad, Department of Statistics, Persian Gulf University, Iran

Alzoubi, Loai M., Department of Mathematics, Al al-Bayt University, Jordan

Appenzeller, Dorota, Department of Econometrics, Poznań University of Economics and Business, Poland

Archontakis, Fragiskos, School of Humanities, International Hellenic University, Greece

Awwad, Raed R. Abu, Department of Mathematics, Faculty of Arts and Sciences, University of Petra, Amman, Jordan

Ayiah-Mensah Francis, Department of Mathematics, Statistics and Actuarial Science, Takoradi Technical University, Ghana

Bayoud, Husam A., College of Sciences and Humanities, Fahad Bin Sultan University, Saudi Arabia

Castle, Jennifer, Department of Economics, University of Oxford, United Kingdom

Cierpiał-Wolan, Marek, President of Statistics Poland, Poland

Ciołek, Dorota, Faculty of Economics, University of Gdańsk, Poland

Cruze, Nathan B., NASA Langley Research Center, USA

Dańska-Borsiak, Barbara, Department of Spatial Econometrics, University of Łódź, Poland

Deepawansa, Diana Dilshanie, Department of Census and Statistics, Sri Lanka

Denkowska, Anna, Department of Mathematics, Cracow University of Economics, Poland

Dudek, Hanna, Institute of Economics and Finance, Warsaw University of Life Sciences, (SGGW), Poland

Dziechciarz, Marta, Department of Econometrics and Operational Research, Wroclaw University of Economics and Business, Poland

Elfaki, Faies, Department of Mathematics, Statistics & Physics, Qatar University, Quatar

Ghailani, Inssaf, Faculty of Sciences of Tetouan, Abdelmalek Essaadi University, Morocco

Giacalone, Massimiliano, Department of Statistics, University of Naples, Italy

Goodman, Bill, Department of Statistics, Ontario Tech University, Canada

Gomes Antonio Eduardo, Department of Statistics, University of Brasília, Brazil

Górecki, Tomasz, Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland

Grzenda, Wioletta, Department of Statistical Methods and Business Analysis, Warsaw School of Economics (SGH), Poland

Gurgul, Henryk, Department of Applications of Mathematics in Economics, AGH University of Science and Technology, Poland

Gurjar, Hariom, Department of Statistics, Chandigarh University, India

Hagemejer, Jan, Faculty of Economic Science, University of Warsaw, Poland

Hamedani, Gholamhossein, Department of Mathematics, Statistics and Computer Science, Marquette University, USA

Haron, Sharifah A., Department of Economics, Universiti Putra Malaysia, Malaysia

Hastenteufel, Jessica, Department of Statistics, International University of Applied Science (IU), Germany

Hugo Hernandez, Hugo, Independent Researcher, Colombia

Hurairah, Ahmed, Department of Statistics, Sana'a University, Yemen

Jabłoński, Łukasz, Department of Macroeconomics, Cracow University of Economics, Poland

Jajuga, Krzysztof, Department of Financial Investments and Risk Management, Wroclaw University of Economics and Business, Poland

Jankiewicz, Jacek, Department of Microeconomics, Poznań University of Economics and Business, Poland

Jędrzejczak, Alina, Department of Statistical Methods, University of Łódź, Poland

Kalton, Graham, Westat, USA

Kamil, Anton, Abdulbasah, Department of Statistics, Istanbul Gelisim University, Turkey

Khatun, Nasrin, Department of Statistics and Data Science, Jahangirnagar University, Bangladesh

Kim, Woosuk, Department of Mathematics & Statistics, Slippery Rock University, USA

Kokot, Sebastian, Department of Econometrics and Statistics, University of Szczecin, Poland

Kończak, Grzegorz, Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Poland

Kossovsky, Alex, Ely, Independent Researcher, USA

Kowalski, Arkadiusz, Michał, Department of East Asian Economic Studies, Warsaw School of Economics (SGH), Poland

Krzyśko, Mirosław, Professor Emeritus, Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznan, Poland

Kuc-Czarnecka, Marta, Faculty of Management and Economy, Gdańsk University of Technology, Poland

Kurek, Robert, Department of Finance and Accounting, Wroclaw University of Economics and Business, Poland

Lahiri, Partha, Department of Mathematics, University of Maryland, USA

Lawson, Nuanpan, Department of Applied Statistics, King Mongkut's University of Technology North Bangkok, Thailand

Leśkow, Jacek, Department of Informatics, Cracow University of Technology, Poland & American University Kyiv, Ukraine

Lisowska, Renata, Department of Entrepreneurship and Industrial Policy, University of Łódź, Poland

Longford, Nick, Mohn Centre for Child Health and Wellbeing, Imperial College London, Poland

Maji, Reba, Department of Mathematics, Sarojini Naidu College for Women, India

Małecka, Marta, Department of Statistical Methods, University of Łódź, Poland

Marzec, Jerzy, Department of Econometrics and Operations Research, Krakow University of Economics, Poland

Mdlongwa, Precious, Department of Mathematics and Statistical Sciences, Botswana International University of Science and Technology (BIUST), Botswana

Morawski, Leszek, Department of Macroeconomics and Theory of Foreign Trade, University of Warsaw, Poland

Mularczyk, Piotr, Faculty of Cybernetics, Military University of Technology (WAT), Poland

Niedzielski, Piotr, Department of Logistics and Innovation, Pomeranian University in Słupsk, Poland

Niftiyev, Ibrhaim, Center on European Economies, Azerbaijan State University of Economics (UNEC), Azerbaijan

Nordholt, Eric Schulte, Statistics Netherlands, Netherlands

Novák, Jiří, Institute for Competitiveness and Communication ICC, Switzerland

Ocloo, Selasi, Kwaku, Department of Mathematical Sciences, University of Mines and Technology, Ghana

Okrasa, Włodzimierz, Cardinal Stefan Wyszyński University in Warsaw & Statistics Poland, Poland

Olaomi, John, O., Department of Statistics, University of South Africa (UNISA), South Africa

Panek, Tomasz, Department of Statistics and Demography, Warsaw School of Economics (SGH), Poland

Panigrahi, Archana, Department of Statistics, Ravenshaw University, India

Patra, Dipika, Department of Statistics, Seth Anandram Jaipuria College, India

Pawełek, Barbara, Department of Statistics, Cracow University of Economics, Poland

Plich, Mariusz, Department of Econometrics, University of Łódź, Poland

Postek, Łukasz, Faculty of Economic Sciences, University of Warsaw, Poland

Prasad, Shakti, Department of Basic & Applied Science, National Institute of Technology, India

Psarrakos, Georgios, Department of Statistics and Insurance Science, University of Piraeus, Greece

Rai, Piyush Kant, Department of Statistics, Banaras Hindu University, India

Rajae, Elkazini, National Higher School of Electricity and Mechanics, ENSEM-Hassan II University of Casablanca, Morocco

Reznikova, Nataliia, Institute of International Relations, Taras Shevchenko National University of Kyiv, Ukraine

Rozkrut, Dominik, University of Szczecin, Poland

Rubil, Ivica, Institute of Economics, Zagreb, Croatia

Salih, Ahmed, Mahdi, Department of Statistics, College of Administration and Economics, Wasit University, Iraq

Schuster, Hannes, Department of Economics, University of Saarland, Germany

Siswanto, Siswanto, Department of Mathematics Sebelas Maret University, Indonesia

Skrovankova, Katarina, Department of Statistics, Alexander Dubcek University of Trencin, Slovakia

Szymkowiak, Magdalena, Institute of Automatic Control and Robotics, Poznan University of Technology, Poland

Tharshan, Ramajeyam, Postgraduate Institute of Science, University of Peradeniya, Sri Lanka & Department of Mathematics and Statistics, University of Jaffna, Sri Lanka

Samutwachirawong, Siriporn, Department of Statistics, Maejo University, Thailand

Sartore, Luca, U.S. Department of Agriculture, National Agricultural Statistics Service, USA

Szymkowiak, Marcin, Department of Statistics, Poznań University of Economics and Business, Poland

Tarczyński, Waldemar, President of Polish Statistical Association & University of Szczecin, Poland

Verma, Ravi, Statistics Canada, Canada

Vernizzi, Achille, Department of Economics, Management, Quantitative Methods, University of Milan, Italy

Wagalla, Alphonse, Department of Mathematics, Bomet University College, Kenia

Wawrowski, Łukasz, Department of Informatics, University of Silesia, Poland

Wesołowski, Jacek, Department of Probability and Stochastic Processes, Warsaw University of Technology, Poland

Wieczorkowski, Robert, Department for Innovation, Statistics Poland, Poland

Wołyński, Waldemar, Department of Mathematical Statistics and Data Analysis, Collegium Mathematicum, Adam Mickiewicz University in Poznan, Poland

Wycinka, Ewa, Department of Statistics, University of Gdańsk, Poland

Yacyshyn, Alison, Faculty of Management, Mihalcheon School of Management, Canada

Zamanzade, Ehsan, Department of Statistics, University of Isfahan, Iran

Zelinowa, Silvia, Department of Mathematics, University of Economics in Bratislava, Slovakia

STATISTICS IN TRANSITION new series, December 2025
Vol. 26, No. 4, pp. 211–215

Index of Authors, Volume 26, 2025

Adeboye N. O., *Bayesian and frequentist modelling of West African economic growth: a dynamic panel approach – SiTns, Vol. 26, No. 3*

Adesina O. A., *see under Adeboye N. O. – SiTns, Vol. 26, No. 3*

Aghel W. E., *see under Hassan A. S. – SiTns, Vol. 26, No. 2*

Agu F.I., *The truncated Schröter recursive algorithm for the computation of aggregate claim amounts – SiTns, Vol. 26, No. 4*

Ali I., *see under Torsen E. – SiTns, Vol. 26, No. 1*

Ali M. K. B. M., *see under Mohamed S. D. – SiTns, Vol. 26, No. 1*

Alzoubi L., *see under Hardan A. – SiTns, Vol. 26, No. 3*

Anand R., *see under Dileepkumar M. – SiTns, Vol. 26, No. 1*

Bandyopadhyay A., *see under Majumder S. – SiTns, Vol. 26, No. 2*

Bantu A. F., *Normality tests for transformed large measured data: a comprehensive analysis – SiTns, Vol. 26, No. 3*

Beresovsky V., *see under Savitsky T. D. – SiTns, Vol. 26, No. 2*

Bhayare U., *see under Thakur N. S. – SiTns, Vol. 26, No. 4*

Błażej M., *Synchronization and similarity between regional and sectoral output gaps in the Polish manufacturing industry – SiTns, Vol. 26, No. 2*

Bohlourihajjar S., *see under Khazaei S. – SiTns, Vol. 26, No. 4*

Bonnini S., *Multivariate two-sample permutation test with directional alternative for categorical data – SiTns, Vol. 26, No. 3*

Borghesi M., *see under Bonnini S. – SiTns, Vol. 26, No. 3*

Borkowski M., *The integrity of the innovation process on the example of EU countries: a PLS-SEM approach – SiTns, Vol. 26, No. 2*

Brania K., *see under Gurgul H. – SiTns, Vol. 26, No. 3*

Gruszewska E., *see under Borkowski M. – SiTns, Vol. 26, No. 2*

Chakraborty S., *see under Handique L. – SiTns, Vol. 26, No. 1*

Chaurasia S., *see under Thakur N. S. – SiTns, Vol. 26, No. 4*

Chipepa F., *see under Nkomo W. – SiTns, Vol. 26, No. 1*

Dash P., *see under Panigrahi A. – SiTns, Vol. 26, No. 3*

Deepawansa D. D., *Selection criteria and targeting the poor for poverty reduction: the case of social safety nets in Sri Lanka – SiTns, Vol. 26, No. 1*

Dileepkumar M., *Reliability properties and applications of proportional reversed hazards in reversed relevation transform, SiTns, Vol. 26, No. 1*

Dunusinghe P., *see under Deepawansa D. D. – SiTns, Vol. 26, No. 1*

Dutta-Powell R., *The perils of premature evaluation: reassessing the application of Benford's Law to the USA's COVID-19 data – SiTns, Vol. 26, No. 3*

Djuraidah A., *Exploring the stochastic production frontier in the presence of outliers: a simulation study – SiTns, Vol. 26, No. 3*

Elsherpieny E. A., *see under Hassan A.S., – SiTns, Vol. 26, No. 2*

Falkiewicz E., *The concept of behavioural model of decision-making under risk – SiTns, Vol. 26, No. 4*

Garg P., *Ratio regression type estimators of the population mean for missing data in sample surveys – SiTns, Vol. 26, No. 2*

Gershunskaya J., *see under Savitsky T. D. – SiTns, Vol. 26, No. 2*

Girul A., *Factors determining the impact of the existence of degraded areas in local government units and the impact of revitalisation activities by intervention area – SiTns, Vol. 26, No. 4*

Górajski M., *see under Błażej M. – SiTns, Vol. 26, No. 2*

Gurgul H., *The impact of the COVID-19 pandemic on forecast uncertainty of macroeconomic data releases – SiTns, Vol. 26, No. 3*

Handique L., *On a family that unifies the generalized Marshall-Olkin and Poisson-G family of distribution – SiTns, Vol. 26, No. 1*

Hardan A., *Area-biased one-parameter exponential distribution with financial applications – SiTns, Vol. 26, No. 3*

Hassan A. S., *Inference of dynamic weighted cumulative residual entropy for Burr XII distribution based on progressive censoring – SiTns, Vol. 26, No. 2*

Irfan M., *Bayesian estimation of two-parameter power Rayleigh distribution and its application – SiTns, Vol. 26, No. 1*

Ismail M. T., *see under Mohamed S. D. – SiTns, Vol. 26, No. 1*

Jamal F., *see under Handique L. – SiTns, Vol. 26, No. 1*

Jaworski S., *Optimal sample size in a triangular model for sensitive questions – SiTns, Vol. 26, No. 1*

Jędrzejczak A., *Regional differentiation of income distributions in Poland – SiTns, Vol. 26, No. 3*

Józefowski T., *see under Młodak A. – SiTns, Vol. 26, No. 2*

Kalinowski S., *see under Łuczak A. – SiTns, Vol. 26, No. 1*

Khazaei S., *Bayesian nonparametric model for weighted data using mixture of Burr XII distributions – SiTns, Vol. 26, No. 4*

Koczy J., *see under Kubiczek J. – SiTns, Vol. 26, No. 2*

Kończak G., *On a new goodness of fit test for multivariate normality with fixed parameters based on David-Hellwig test idea – SiTns, Vol. 26, No. 4*

Kozyra A., *see under Bantu A.F. – SiTns, Vol. 26, No. 3*

Kubus M., *Application of the nonlinear splines model to forecast changes in the construction costs index – SiTns, Vol. 26, No. 4*

Kubiczek J., *Harnessing AI for business transformation: strategies for effective implementation and market advantage – SiTns, Vol. 26, No. 2*

Kumar J., *Exploring new mixtures of distribution to model a skewed and heavy tailed data – SiTns, Vol. 26, No. 4*

Kumar P., *see under Sharma H. – SiTns, Vol. 26, No. 1*

Łuczak A., *A fuzzy hybrid MCDM approach to the evaluation of subjective household poverty, SiTns, Vol. 26, No. 1*

Majumder S., *Formulation of estimator for population mean in stratified successive sampling using memory based information – SiTns, Vol. 26, No. 2*

Gupta A., *see under Majumder S. – SiTns, Vol. 26, No. 1*

Mach Ł., *see under Kubus M. – SiTns, Vol. 26, No. 4*

Madukaife M. S., *Testing for multinormality with goodness-of-fit tests based on phi divergence measures – SiTns, Vol. 26, No. 2*

Marszałek M., *Households' invisible input to the economy: a review of its measurement methods and results – SiTns, Vol. 26, No. 2*

Mijinyawa M., *see under Torsen E. – SiTns, Vol. 26, No. 1*

Mishra G., *see under Panigrahi A. – SiTns, Vol. 26, No. 3*

Misiurski P., *see under Kubus M. – SiTns, Vol. 26, No. 4*

Misztal M., *see under Jędrzejczak A. – SiTns, Vol. 26, No. 3*

Młodak A., *Application of statistical disclosure control methods to protect the confidentiality of the 2020 agricultural census microdata – SiTns, Vol. 26, No. 2*

Modibbo U. M., *see under Torsen E. – SiTns, Vol. 26, No. 1*

Mohamed S. D., *Improving detectability of the indicator saturation approach through winsorization: an empirical study in the cryptocurrency market – SiTns, Vol. 26, No. 1*

Morawski L., *Education expansion and income inequality: evidence from Poland (2005–2019) – SiTns, Vol. 26, No. 3*

Nain A., *see under Kumar J. – SiTns, Vol. 26, No. 4*

Nduka U. C., *see under Madukaife M. S. – SiTns, Vol. 26, No. 2*

Nkomo W., *Type I heavy-tailed family of generalized Burr III distributions: properties, actuarial measures, regression and applications – SiTns, Vol. 26, No. 1*

Oluyede B., *see under Nkomo W. – SiTns, Vol. 26, No. 1*

Ossai E. O., *see under Madukaife M. S. – SiTns, Vol. 26, No. 2*

Panigrahi A., *A minimum variance unbiased estimator of finite population variance using auxiliary information – SiTns, Vol. 26, No. 3*

Pekasiewicz D., *see under Jędrzejczak A. – SiTns, Vol. 26, No. 3*

Pranata I., *see under Djuraidah A. – SiTns, Vol. 26, No. 3*

Rasyid S., *Clustering based on poverty indicator data using K-Means cluster with density-based spatial clustering of application with noise – SiTns, Vol. 26, No. 2*

Roszko-Wójtowicz E., *see under Kubiczek J. – SiTns, Vol. 26, No. 2*

Sahriman S., *see under Rasyid S. – SiTns, Vol. 26, No. 2*

Sankaran P. G., *see under Dileepkumar M. – SiTns, Vol. 26, No. 1*

Savitsky T. D., *Thresholding nonprobability units in combined data for efficient domain estimation – SiTns, Vol. 26, No. 2*

Seknewna L. L., *see under Torsen E. – SiTns, Vol. 26, No. 1*

Sharma A. K., *see under Irfan M. – SiTns, Vol. 26, No. 3*

Sharma H., *On survival estimation of Lomax distribution under adaptive progressive type-II censoring – SiTns, Vol. 26, No. 1*

Siswanto S., *see under Rasyid S. – SiTns, Vol. 26, No. 2*

Skrodzka I., *Impact of human capital on the innovation performance of EU economies – SiTns, Vol. 26, No. 1*

Srivastava M. K., *see under Garg P. – SiTns, Vol. 26, No. 2*

Srivastava N., *see under Garg P. – SiTns, Vol. 26, No. 2*

Suproń B., *Modelling the asymmetric relationship between energy and CO2 emissions in the Visegrad group: empirical evidence from a panel NARDL approach – SiTns, Vol. 26, No. 3*

Swanson D. A., *On using Arima model confidence intervals applied to population projections based on the components of change: a case study for the world population – SiTns, Vol. 26, No. 4*

Tayman J., *see under Swanson D. A. – SiTns, Vol. 26, No. 4*

Thakur N. S., *Estimation of mean using factor-type estimator under adaptive cluster sampling design – SiTns, Vol. 26, No. 4*

Torsen E., *Analytical modelling for COVID-19 data (fatality): a case study of Nigeria for the period of February 2020-April 2022 – SiTns, Vol. 26, No. 1*

Ulrichs M., *see under Błażej M. – SiTns, Vol. 26, No. 2*

Waszkiewicz I., *see under Kubiczek J. – SiTns, Vol. 26, No. 2*

Wesołowska M., *Exploring variation in data on income inequality across databases and measures in post-socialist countries – SiTns, Vol. 26, No. 4*

Williams M. R., *see under Savitsky T. D. – SiTns, Vol. 26, No. 2*

Wiora J., *see under Bantu A. F. – SiTns, Vol. 26, No. 3*

Woś K., *see under Kubiczek J. – SiTns, Vol. 26, No. 2*

Wywiał J. L., *Generalised spatial autocorrelation coefficients – SiTns, Vol. 26, No. 1*

Xu Z., *An expectation-maximization algorithm for logistic regression based on individual level predictors and aggregate-level response – SiTns, Vol. 26, No. 1*

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 http://www.libweb.anglia.ac.uk/referencing/harvard.htm. 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).