Volume 27, Number 2, June 2026 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

ISSN 1234-7655

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

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

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

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

CONTENTS

STATISTICS IN TRANSITION new series, *June 2026*
Vol. 27, No. 2, 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, *June 2026*
Vol. 27, No. 2, pp. V–VI

Policy statement

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

***

Abstracting and indexing databases

Statistics in Transition new series is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

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

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

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

Electronic Journals Library

Elsevier – Scopus

Genamics JournalSeek

Google Scholar

Index Copernicus

J-Gate

JournalGuide

JournalTOCs

Keepers Registry

MIAR

Microsoft Academic

OpenAIRE

ProQuest – Summon

Publons

QOAM (Quality Open Access Market)

ReadCube

RePec

SCImago Journal & Country Rank

TDNet

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

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. VII–XI

From the Editor

The June issue of *Statistics in Transition new series* contains a set of twelve articles by twenty-eight authors from eight countries - in order of appearance: USA, Poland, Morocco, India, Slovakia, Nepal, Jordan, Uganda.

Original research papers

In the first paper, ***Mean absolute deviations for the Weibull distribution: applications in survival analysis and insurance claims***, **Weiqi Zhang**, **Zibo Wang**, and **Eugene Pinsky** present the formula for the Mean Absolute Deviation (MAD) about the median of the Weibull distribution and provided the calculations for the scale and shape parameters using three methods: the Maximum Likelihood Estimation (MLE), the Quantiles method, and the Mean Absolute Deviation (MAD) around the median method. The application of these methods to real-world data underscores the importance of selecting an appropriate statistical approach based on the dataset's characteristics. In survival analysis, where outcomes can vary widely, methods such as MAD, which reduce the influence of outliers, yield more reliable and clinically relevant predictions. Similarly, in actuarial science, modeling insurance claim amounts with MAD improves parameter estimation robustness, particularly for heavy-tailed or skewed distributions where extreme values are common.

Stanisław Maciej Kot in the article entitled ***The entanglement of attitudes toward inequality: the theoretical background and measurement for the EU countries in 2021*** assumes two types of social planners who evaluate income distributions in terms of social welfare, economic inequality, and poverty. The first type, $SP\varepsilon$, denotes individuals who have an aversion to income inequality as measured by the normative parameter ε. The second, $SP\nu$, comprises individuals who have an aversion to rank inequality, as measured by the normative parameter v. Since every member of a society may play the role of a social planner, there could be as many levels of ε and ν as there are society members. It raises the question of which ranges of v and ε values are ethically sensible when conducting empirical welfare studies. This paper proposes an answer to this question by introducing the concepts of inequality-entangled $SP\nu$ and $SP\varepsilon$.

In the next paper, ***Multivariate statistical analysis of the seismic activity in Morocco using PCA and K-Means clustering,*** prepared by **Achraf Chakir Baraka**, **Kaoutar Baraka**, **Mehdi Rahmaoui**, **Nada Yamoul**, **Yassine Bahi**, and **Hamid Khalifi,** authors propose the integration of statistical analyses, using principal component

analysis to reduce the number of dimensions and detect the multidimensional structure, in addition to applying the K-means algorithm to classify seismic motions according to their magnitude. The study thus preserved redundant information based on two principal components: the first component is characterized by a strong correlation with magnitude and significance, while the attitude and time variables are strongly correlated with the second component. This demonstrates that the first component reflects the intensity of seismic motions. The objective of statistical analysis is to reduce the number of dimensions by highlighting the dependencies between different variables in order to ensure effective risk management related to seismic events.

Hosenur Rahman Prodhani's and **Rama Shanker's** article ***Power quasi Sujatha distribution with properties and applications in real lifetime data*** presents a three-parameter power quasi Sujatha distribution. Statistical properties including the survival function, hazard function, reverse hazard function, mean residual life function and stochastic ordering have been discussed. Moments of the proposed distribution have been obtained. The estimation of the parameters using the maximum likelihood method and maximum product spacing estimation has been explained and a simulation study has been presented to determine the efficiency of the maximum likelihood estimate of the parameters. The bootstrap confidence interval method has been used to estimate the confidence interval of the parameters. Finally, two examples of real lifetime datasets have been presented to demonstrate the applications of the proposed distribution

Lívia Krajčíková and **Mária Vojtková** describe ***Consumption patterns of Slovak households in 2021 and 2022*** to analyze the similarities and differences in the structure of consumption expenditure of various types of Slovak households in 2021 and 2022. The study also focuses on the impact of demographic factors on spending behavior, examining how households in different income groups allocate their expenses. To identify and profile individual household segments, they applied cluster analysis, which allowed to distinguish homogeneous groups of households based on their spending patterns. The analysis was based on data from the Household Budget Survey. The results indicate that Slovak households can be divided into six main segments, with four segments displaying stable spending patterns over both analyzed years, and two where spending patterns varied from one of the studied years to another. The results emphasize significant differences between the expenditure structures of low- and high-income households as well as among households with varying demographic compositions.

The paper prepared by **Agata Boratyńska**, ***Bayesian sensitivity of insurance premium in collective risk model under bivariate prior with dependent frequency and severity of claims***, deals with the problem of robustness of the collective and Bayes premiums under uncertainty of prior knowledge. The inaccuracy of the prior

knowledge concerns the disturbance of independence between variables describing the frequency and average value of claims. Traditionally, these variables are independent, but in applications it is not always the case. Two classes of priors are presented: in the first class, the FGM copula is applied, while in the second one, the dependence between two contaminated priors is shown. In both classes, priors have the form of a linear combination of known bivariate probability distributions. The ranges of collective and Bayes premiums are calculated and prior and posteriori regret gamma-minimax premiums are presented as the optimal premiums. Despite the very mild or small dependence, its influence on the premiums, especially on the bonus-malus factor, is relatively significant.

Arjun Kumar Gaire in the article ***New version of Log-Logistic distribution: properties and applications to survival time and demographic data*** proposes a Deflation-Inflation Log-Logistic (DILLog) distribution as a sub-model of the Deflation-Inflation Distributed (DID) family, introduced by Alodat and Al-Rawwash (2021). The proposed model offers greater flexibility than the original model in fitting data from real-world problems, especially for survival times and demographic data. The DILLog model is characterized by unimodal right-tailed density and hazard rate functions, and its key statistical properties, including the cumulative distribution function and a closed-form quantile function, are derived. To test the performance of the distribution, a simulation study has been used as well as an application to two real datasets: the age at menarche of Nepalese girls and the survival times of patients suffering from melanoma disease. To illustrate the usefulness and application of the proposed distribution, its parameters were estimated by using the maximum likelihood estimation method.

Raed R. Abu Awwad, **Ghassan K. Abufoudeh**, **Samer Alokaily**, and **Maalee Almheidat** in their work ***Progressive Type II censored exponential data analysis: the method comparison with a breakdown voltage case study*** provide a comparative analysis of frequentist and Bayesian estimation methods for the parameter, reliability and hazard functions of the exponential distribution using progressive type II censored data. The theoretical methods presented are well established in the statistical literature; the article's contribution lies in the systematic empirical comparison of these approaches. The maximum likelihood estimates and Bayes estimates of the parameter, reliability and hazard functions have been computed using standard procedures implemented via "Mathematica 12" software. Two loss functions have been considered: squared error and Kullback-Leibler for Bayesian computation under two priors: weakly informative and informative.

The paper by **Bernard Kokczyński** and **Dorota Witkowska**, ***Effectiveness of bankruptcy prediction models constructed for differently selected diagnostic variables*** compares the effectiveness of discriminant models for predicting corporate bankruptcy, constructed using different methods of diagnostic variables selection. Several

methods, such as arbitrary selection, the forward stepwise method applied after the initial selection of variables by means of a significance test of differences between group averages, the Hellwig method, the t-statistics and the backward stepwise method were compared. The authors also assessed the models' accuracy in terms of the synthetic measure. It was constructed by applying eight measurements of the classification effectiveness, such as the values of Wilks' lambda statistic and AUC together with the percentage of correctly identified companies, i.e. total, bankrupts and non-bankrupts in training and testing sets.

In the article ***ARIMA-LSTM hybrid model for forecasting urban temperature dynamics in Ugandan cities,*** **Lekia Nkpordee**, **Yusuf Abass Aleshinloye**, **Ejidokun Adekunle Olugbenga**, and **Ikpotokin Osayomore** develop and evaluate an ARIMA-LSTM hybrid model with the goal of capturing both linear trends and nonlinear fluctuations within a unified and interpretable framework. Monthly temperature data from 2017 to 2023 obtained from the Uganda National Meteorological Authority, alongside long-term urban population data from the World Bank, were used to support robust urban climate analysis. Data quality was ensured through systematic preprocessing and outlier assessment, providing reliable inputs for model estimation. The proposed hybrid approach applies ARIMA to explicitly model linear and seasonal temperature structures, while an LSTM network learns the remaining nonlinear patterns embedded in the residuals. Model performance was evaluated against a wide range of benchmark models, including standalone statistical models, deep learning architectures, machine learning methods, and Facebook Prophet used strictly for comparison.

Conference papers

XXXXII Multivariate Statistical Analysis 2024, Lodz, Poland

Joanna Landmesser-Rusek's paper, entitled ***Comparison of two types of topological networks for the foreign exchange market: one based on correlation coefficients and the other on the concept of causality*** compares two types of topological networks for the foreign exchange market: those based on correlation coefficients and those based on Granger's concept of causality. The networks were constructed in a stepwise manner for the most important world currencies in the period from 03/01/2020 to 18/10/2024. The comparison was carried out using certain topological characteristics of the networks, such as density, average distance, diameter, centralization index, and degrees of vertices. Properties of both approaches were highlighted. The study also aimed to demonstrate that causality networks act as a complementary framework to traditional correlation models. Accordingly, currency networks were built taking into account the strength of relationships between currency pairs and the directionality of these links.

Research Communicates and Letters

Tofan Kumar Biswal in the paper ***R-optimal design strategies for logistic regression models with complementary log-log link*** investigates optimal experimental design strategies, specifically R-optimality, for two-parameter logistic regression (2PLR) models using the complementary log-log link function based on two- and three-support point designs. The study seeks to establish efficient designs that minimize the average width of confidence bands across the range of predictor variables. The general equivalence theorem validates the necessary and sufficient conditions of this optimality criterion. This program is employed to quantitatively determine the support points of the optimal designs and the corresponding weights assigned to these points.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 1–14, https://doi.org/10.59139/stattrans-2026-013
Received – 06.03.2025; accepted – 02.01.2026

Mean absolute deviations for the Weibull distribution: applications in survival analysis and insurance claims

Weiqi Zhang[1], **Zibo Wang**[2], **Eugene Pinsky**[3]

Abstract

The Weibull distribution is widely applied in fields such as survival analysis, reliability engineering, failure analysis, and extreme value theory. Traditionally, Maximum Likelihood Estimation (MLE) has been commonly used to estimate the parameters of this distribution. In this paper, we derive a new formula for the mean absolute deviation (MAD) about the median. We use this formula to derive a MAD-based parameter-estimation method that is computationally simpler than MLE. We apply our results to analyze the survival times of breast cancer patients and insurance claim amounts, providing evidence from biomedical and actuarial domains. We estimate parameters using MLE, MAD, and quantile-based approaches. The results show that the proposed MAD-based approach is superior to the other two methods. It demonstrates the practical application of MAD methods in survival analysis and financial risk modeling of insurance claims, where accurate modeling is crucial for understanding extreme outcomes.

Key words: Weibull.

1. Introduction

In various fields, from healthcare to manufacturing, accurate statistical models are essential for evaluating risks and understanding the reliability and survival of systems or patients (Carroll, 2003). These models play a crucial role in quantifying the probabilities of events occurring over time. By capturing the temporal dynamics of such events, they enable a deeper understanding of risk factors, reliability patterns (Almeida, 1999), and longevity predictions across various domains, including engineering (Kang 2018), finance (Chen, 2011; Gebizlioglu 2011), and healthcare (Quiroz, 2024), to name just a few. The Weibull distribution is widely utilized in survival analysis and reliability engineering due to its flexibility in modeling various failure rates, making it a versatile tool for analyzing time-to-event data and estimating the probability of survival or failure under different conditions (Fogliatto, 2019), (Yoosefi, 2018).

For the Weibull distribution, the Maximum Likelihood Estimation (MLE) is the main method for parameter estimation (Cohen, 1965; Balakrishnan, 2008). However, MLE has some limitations, and the MLE method is subject to errors when the data distribution is

[1]Department of Computer Science, Metropolitan College, Boston University, United States.
ORCID: https://orcid.org/0009-0009-2536-9529.
[2]Department of Computer Science, Metropolitan College, Boston University, United States.
ORCID: https://orcid.org/0009-0008-6243-3131.
[3]Department of Computer Science, Metropolitan College, Boston University, United States.
E-mail: epinsky@bu.edu. ORCID: https://orcid.org/0000-0002-3836-1851.

skewed, or outliers are present (Jacquelin, 1993). In such cases, researchers are exploring alternative methods to provide more accurate and robust estimates (Cohen,1982; Teimouri, 2011; Pobockova, 2014). Some researchers use the Quantiles method as an alternative to maximum likelihood estimation (Jokiel, 2024). The use of Quantiles offers an insightful approach for analyzing the variability in the estimated values of α, revealing important aspects of the data's structure and concentration patterns (Pinsky, 2024). This method, which estimates parameters based on specific percentiles, captures the distribution's characteristics in a simplified manner but still fails to eliminate the effects of outliers.

Research has shown that MAD exhibits greater resistance to extreme values and delivers more reliable estimates, even in the presence of data irregularities, thus reducing the impact of outliers (Hortobagyi, 2022). In healthcare, building accurate and robust statistical models of patient survival can inform treatment strategies, guide resource allocation, and improve patient care by accurately estimating survival probabilities (Cancho, 2020). Accurate survival estimates are critical for predicting outcomes and assessing risk probabilities.

In this paper, we derive formulas for calculating MAD around the median of the Weibull distribution. The median is particularly important in cancer survival studies because it provides a robust measure of central tendency that is less susceptible to extreme values, making it a reliable reference point for survival outcomes. In addition, we derive three methods for calculating parameters using MLE, quantiles, and MAD. Finally, we used all three methods to predict survival in breast cancer patients and model insurance claim amounts, demonstrating the method's versatility across biomedical and actuarial contexts. It was found that MAD provided more accurate predictions with less error than MLE and quantiles. The aim of this study is to provide a method for calculating Weibull distribution parameters using MAD to improve the validity of statistical analysis by enabling more reliable estimation techniques and modeling methods in survival analysis and financial risk assessment.

2. Mean absolute deviations for the Weibull distribution

The Weibull distribution is defined by the following probability density function (PDF) and cumulative distribution function (CDF):

$$f(x) = \frac{k}{\alpha}\left(\frac{x}{\alpha}\right)^{k-1} e^{-(x/\alpha)^k} \quad \text{and} \quad F(x) = 1 - e^{-(x/\alpha)^k}, \quad x \geq 0 \tag{1}$$

The quantile function $Q(p)$ is $Q(p) = \alpha\left(-\log(1-p)\right)^{1/k}$. The quartiles for Weibull distributions are $Q_1 = \alpha(\log 4/3)^{1/k}$, $M = \alpha(\log 2)^{1/k}$ and $Q_3 = \alpha(2\log 2)^{1/k}$. The mean μ and variance σ^2 are given by:

$$\mu = \alpha\Gamma\left(1+\frac{1}{k}\right), \quad \sigma^2 = \alpha^2\left[\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right] \tag{2}$$

where $\Gamma(\cdot)$ denotes the Gamma function (Olver, 2010):

$$\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}\,dt \tag{3}$$

To compute mean absolute deviations, we will find it convenient to introduce the auxiliary integral:

$$I(z) = \int_{x \le z} x\, dF(x) \tag{4}$$

We can consider $I(z)$ as the partial mean of X computed over all $x \le z$. The mean absolute deviation (MAD) of X from a constant c is defined as:

$$H(X,c) = \int_{x \le c} (c-x)\, dF(x) + \int_{x>c} (x-c)\, dF(x) \tag{5}$$

This can be rewritten as:

$$H(X,c) = c\,(2F(c)-1) + \mu - 2I(c) \tag{6}$$

We can express $I(c)$ as:

$$I(c) = \int_{x \le c} x\, dF(x) = cF(c) - \int_{x \le c} F(x)\, dx \tag{7}$$

Substituting this into equation (6) gives:

$$\begin{aligned} H(X,c) &= (\mu - c) + 2\alpha \int_0^{c/\alpha} \left(1 - e^{-z^k}\right) dz \\ &= (\mu + c) - 2\alpha \int_0^{c/\alpha} e^{-z^k}\, dz \end{aligned} \tag{8}$$

From the definition of the upper incomplete Gamma function (Olver, 2010)

$$\Gamma(s,x) = \int_x^{\infty} t^{s-1} e^{-t}\, dt \tag{9}$$

we obtain from equation (8)

$$H(X,c) = (\mu + c) + \frac{2\alpha}{k}\Gamma\left(\frac{1}{k},\ \left(\frac{c}{\alpha}\right)^k\right) - \frac{2\alpha}{k}\Gamma\left(\frac{1}{k},\ 0\right) \tag{10}$$

For the Weibull distribution, the mean $\mu = \alpha\Gamma(1 + 1/k) = (\alpha/k)\Gamma(1/k)$. Thus, the mean absolute deviation around the mean is:

$$H(X,\mu) = \frac{2\alpha}{k}\Gamma\left(\frac{1}{k},\ \Gamma^k\left(1 + \frac{1}{k}\right)\right) \tag{11}$$

The median M of the Weibull distribution is $M = \alpha\,(\log 2)^{1/k}$. Therefore, the mean absolute deviation around the median is:

$$H(X,M) = \alpha\left[(\log 2)^{1/k} - \Gamma\left(1 + \frac{1}{k}\right) + \frac{2}{k}\Gamma\left(\frac{1}{k},\ \log 2\right)\right] \tag{12}$$

Example 1: For $k = 1$, we have the exponential distribution with rate $1/\alpha$. The mean $\mu = \alpha$

and median $M = \alpha \log 2$. The corresponding mean absolute deviations are: $H(X,\mu) = (2\alpha/e)$ and $H(X,M) = \alpha \log 2$

Example 2: For $k = 2$, we have the Rayleigh distribution with scale $\sigma = (\alpha/\sqrt{2})$. The mean is $\mu = \alpha(\sqrt{\pi}/2)$. Using the identities $\Gamma(3/2) = (\sqrt{\pi}/2)$ and $\Gamma(1/2,x) = \sqrt{\pi}\,\mathrm{erfc}(\sqrt{x})$, we compute:

$$H(X,\mu) = \alpha\sqrt{\pi}\,\mathrm{erfc}\left(\frac{\pi}{2}\right) \tag{13}$$

The median is $M = \alpha\sqrt{\log 2}$. Therefore, the MAD around the median for the Rayleigh distribution is:

$$H(X,M) = \alpha\left[\sqrt{\log 2} - \frac{\sqrt{\pi}}{2} + \sqrt{\pi}\,\mathrm{erfc}\left(\sqrt{\log 2}\right)\right] \tag{14}$$

Example 3: Consider the Frechet (inverse Weibull) distribution with shape $k > 0$, scale $\alpha > 0$ and location 0 [**?**]. Its density function (PDF) $fx)$ and cumulative distribution function (CDF) $F(x)$ given by:

$$f(x) = \frac{k}{\alpha}\left(\frac{x}{\alpha}\right)^{-k-1} e^{-(x/\alpha)^{-k}} \quad \text{and} \quad F(x) = 1 - e^{-(x/\alpha)^{-k}}, \quad x > 0 \tag{15}$$

The quantile function $Q(p)$ is $Q(p) = \alpha\,(-\log p)^{-1/k}$. The quartiles for this distribution are $Q_1 = \alpha\,(\log 4)^{-1/k}$, $M = \alpha\,(\log 2)^{-1/k}$, and $Q_3 = \alpha(\log 4/3)^{-1/k}$. The mean and variance are defined only for $k > 1$ and $k > 2$ respectively and are:

$$\mu = \alpha\Gamma\left(1 - \frac{1}{k}\right), \qquad \sigma^2 = \alpha^2\Gamma\left(1 - \frac{2}{k}\right) - \Gamma^2\left(1 - \frac{1}{k}\right) \tag{16}$$

For $k > 1$, the mean absolute deviations exist, and we can derive mean absolute deviations in a manner similar to Weibull in equations (6), (7), (8). For example, for $H(X,M)$ we obtain

$$H(X,M) = \alpha\left[(\log 2)^{-1/k} - \Gamma\left(1 - \frac{1}{k}\right) + \frac{2}{k}\Gamma\left(\frac{1}{k},\,(\log 2)^{-1}\right)\right] \tag{17}$$

in complete analogy with the Weibull distribution.

3. Confidence intervals and tail estimation

Let us examine how the MAD deviations can be used to estimate tail probabilities. The most general bound for any distribution (with finite variance) is Chebyshev's inequality (Feller, 1956):

$$P\left(|X - \mu| \geq b\sigma\right) \leq \frac{1}{b^2} \tag{18}$$

This inequality is useful for $b \geq 1$. This inequality follows from the so-called Pearson inequality (Feller, 1956) with $r = 2$:

$$P\left(|X - \mu| \geq bV_r^{1/r}\right) \leq \frac{1}{b^2}, \qquad \text{where} \qquad V_r = \mathrm{E}\left(|X - \mu|^r\right) \tag{19}$$

For $r = 1$, a much less-known inequality exists for bounds in terms of mean absolute deviation $H(X,\mu)$ from the mean, namely

$$P\left(|X-\mu| \geq bH(X,\mu)\right) \leq \frac{1}{b} \tag{20}$$

Similarly, there is an inequality in terms of the mean absolute deviation from the median $H(X,M)$ given by (PhamGia, 2001)

$$P\left(|X-M| \geq bH(X,M)\right) \leq \frac{1}{b} \tag{21}$$

Let us compute bounds based on the MAD deviation. First, define $\delta = H(X,\mu)/\sigma$. Note that for all distributions we have $H(X,M) \leq H(X,\mu) \leq \sigma$ and therefore $\delta \leq 1$. We can re-write the MAD-based inequality for $H(X,\mu)$ in equation (20) in terms of σ as follows:

$$P(|X-\mu| \geq b\sigma) = P\left(|X-\mu| \geq \frac{b\sigma}{H(\mu)} \cdot H(\mu)\right) \leq \frac{\delta}{b} \tag{22}$$

Also, we can consider the Peek inequality.

$$P(|X-\mu| \geq b\sigma) \leq \frac{1-\delta^2}{b^2 - 2b\delta + 1} \tag{23}$$

Comparing equations (18) and (22) we find that MAD-based upper bound for $H(X,\mu)$ is lower than Chebyshev's upper bound for $1 \leq b \leq 1/\delta$. Comparing equations (22) and (23) we find that the Peek inequality is lower than both MAD and Chebyshev for $b > 1/\delta$. This is shown in Figure 1.

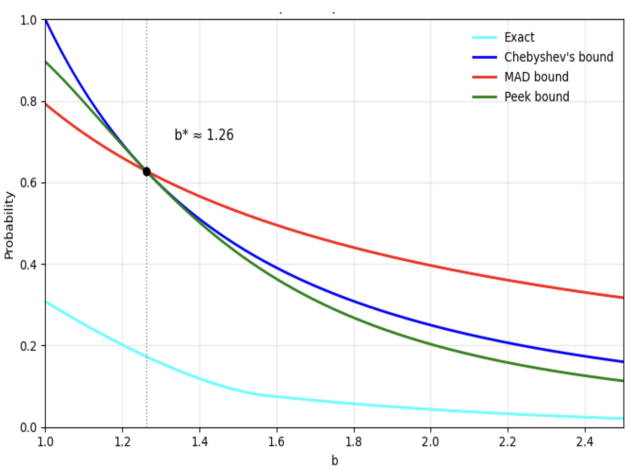


Figure 1. Comparison of Tail Probabilities for the Standard Symmetric Case

In Figure 1, we compare the exact tail probabilities with three classical bounds: the Chebyshev bound (equation (18)), the MAD-based bound (equation (22)), and the Peek bound (equation (23)). For the Weibull distribution with $k = 1.6$ and $\alpha = 1.0$, we obtain $\mu = 0.90$, $\sigma = 0.57$, $H(\mu) = 0.45$, and $\delta = 0.79$, which yields the intersection point $b^* = 1/\delta \approx 1.26$. For $1 \leq b < 1.26$, the MAD-based inequality provides a tighter upper bound than either Chebyshev or Peek. For $b > 1.26$, both Chebyshev and Peek improve upon the MAD bound, with the Peek inequality offering the sharpest estimate. Nevertheless, none of these inequalities approximate the exact tail probabilities well, particularly for larger values of b.

4. Parameter estimation

The most commonly used method for estimating the Weibull distribution parameters α (scale) and k (shape) is Maximum Likelihood Estimation (MLE). MLE efficiently estimates parameters by maximizing the likelihood function, making it statistically robust and precise with large samples. However, MLE can be computationally intensive and sensitive to small sample sizes or outliers. For this reason, we compare MLE with alternative methods, such as the Quantiles and Mean Absolute Deviation (MAD) about the median approaches, which offer simpler computations and may provide more robust estimates in certain cases.

4.1. Parameter estimation using quantiles

Another simple method for estimating the Weibull distribution's parameters is to use quantiles. This approach is based on the relationship between the median and the third quartile. The procedure is as follows:

1. for the Weibull distribution, the median M and the third quartile Q_3 are related via quantiles have the following formulas:

$$\begin{cases} M = \alpha(\log 2)^{1/k} \\ Q_3 = \alpha(2\log 2)^{1/k} \end{cases} \quad \Longrightarrow \quad Q_3 = 2^{1/k} M \tag{24}$$

2. Calculate the sample median M, and find the third quartile Q_3, the point where 75% of the data falls below.
3. Using the formula between the median and third quartile, estimate the shape parameter k as follows:

$$2^{1/k} = \frac{Q_3}{M} \quad \Longrightarrow \quad k = \frac{\log 2}{\log Q_3 - \log M} \tag{25}$$

4. Once k is estimated, compute the scale parameter α using the following equation:

$$\alpha = \frac{M}{(\log 2)^{1/k}} = \frac{M}{(\log 2)^{\frac{\log(Q_3)-\log(M)}{\log 2}}} \tag{26}$$

4.2. Parameter Estimation Using MAD (around Median)

An alternative to Maximum Likelihood Estimation is to use the Mean Absolute Deviation (MAD) around the sample median, which can be computationally simpler. The procedure to estimate the parameters α and k is as follows:

1. Compute the sample median $M = \text{median}\{x_1, \ldots, x_n\}$, which represents the central tendency of the data.
2. Compute the Mean Absolute Deviation H about the sample median M:

$$H = \frac{1}{n} \sum_{i=1}^{n} |x_i - M| \tag{27}$$

   where H measures the spread of the data around the median.
3. From the Weibull distribution properties, we use the following relationship:

$$\frac{H}{M} = 1 - \frac{1}{k(\log 2)^{1/k}} \left[\Gamma\left(\frac{1}{k}\right) + 2\Gamma\left(\frac{1}{k}, \log 2\right) \right] \tag{28}$$

   where Γ is the Gamma function, and $\Gamma(a,b)$ is the incomplete Gamma function. This equation provides an estimate of k, the shape parameter.
4. Once k is estimated, compute α, the scale parameter, from the following equation:

$$\alpha = \frac{M}{(\log 2)^{1/k}} \tag{29}$$

5. Estimation results

To assess the accuracy of the proposed parameter estimation, we will generate Weibull distributions and compare our method with maximum likelihood and quantile methods.

To generate Weibull distribution, we generate n random values $p_i \in (0,1)$ and compute $x_i = Q(p_i)$ from the quantile function. We repeat this procedure 100 times and report the estimation results (averages and standard deviation SD for parameters k and α) for the Weibull distribution in Table 1. As shown in the table, the results for k are more accurate, with a much lower SD, than those obtained with the quantile method.

Figure 2 presents the relative percentage error for estimating the shape k and scale α parameters.

As can be seen from Figure 2, for estimating the shape k, the Mean Absolute Deviation from Median method (MAD-M) is very close to Maximum Likelihood, and both methods give better results than the quantile method. On the other hand, for estimating the scale α, the MAD-M method is worse than the quantile and MLE for smaller sample sizes n (up to 200), but for large n, the proposed method is similar to MLE and superior to the quantile method. One advantage of the MAD-M method is its reduced sensitivity to outliers in the estimated parameters.

Table 1. Estimation Results for the Weibull Distribution ($k = 1$, $\alpha = 2$)

Method	$n = 100$				$n = 200$			
	k		α		k		α	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MLE	1.0132	0.0830	2.0036	0.2055	1.0123	0.0564	2.0021	0.1520
MAD$-M$	1.0162	0.1013	2.0147	0.2514	1.0132	0.0698	2.0049	0.1827
Quantile	1.0468	0.1962	2.0079	0.2318	1.0287	0.1347	2.0016	0.1697
	$n = 400$				$n = 800$			
MLE	1.0031	0.0402	2.0021	0.1091	1.0015	0.0277	1.9996	0.0751
MAD$-M$	1.0021	0.0483	2.0012	0.1257	1.0011	0.0345	2.0005	0.0901
Quantile	1.0113	0.0923	1.9986	0.1220	1.0066	0.0666	1.9986	0.0852

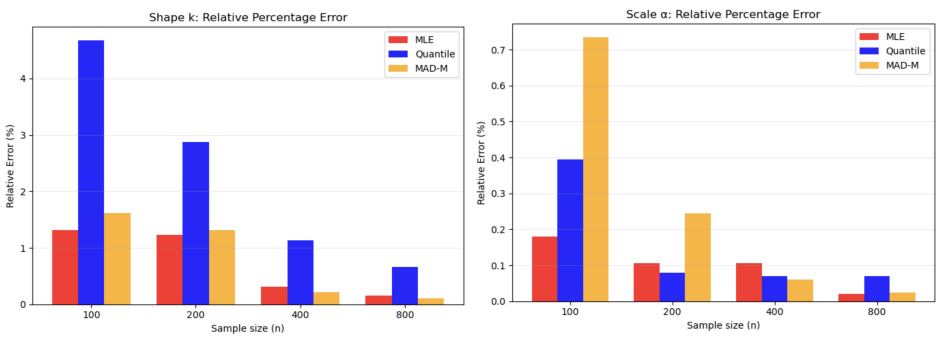


Figure 2. Relative Percentage Error Plots for the Weibull Distribution ($k = 1$ and $\alpha = 2$)

6. Case Study I: Analysis of Breast Cancer Survival Time Using MLE and MAD Methods

Data for this analysis were sourced from the cBioPortal platform, specifically the Breast Cancer (METABRIC) dataset, which contains clinical data for thousands of breast cancer patients. In this study, we analyze patient data from the METABRIC dataset, focusing on key clinical attributes related to breast cancer survival. To streamline the presentation, we employ several common abbreviations in our tabular data. Type of Surgery is abbreviated as mastectomy (Mast) and breast-conserving surgery (BCS). For Cancer Type Detailed, we use Invasive Ductal Carcinoma (IDC), Mixed Ductal and Lobular Carcinoma (MDLC), and Invasive Lobular Carcinoma (ILC). The sample is presented in Table 2.

These abbreviations are widely used in clinical oncology literature and facilitate a more concise representation of patient data, particularly when discussing multiple cases in survival analysis. The selected columns, including age at diagnosis, type of surgery, cancer subtype, chemotherapy status, and overall survival (measured in months), are crucial for understanding patient outcomes and the factors influencing survival rates.

Table 2. Sample of Breast Cancer Patients from the METABRIC Dataset
(Mast. = Mastectomy, BCS = Breast-Conserving Surgery, IDC = Invasive Ductal Carcinoma, MDLC = Mixed Ductal and Lobular Carcinoma, ILC = Invasive Lobular Carcinoma).

Patient ID	Age at Diagnosis	Type of Surgery	Cancer Type	Chemotherapy	Survival (Months)
MB-0000	75.65	Mast.	IDC	No	140.5
MB-0002	43.19	BCS	IDC	No	84.63
MB-0005	48.87	Mast.	IDC	Yes	163.70
MB-0006	47.68	Mast.	MDLC	Yes	164.93
MB-0008	76.97	Mast.	MDLC	Yes	41.37
MB-0010	78.77	Mast.	IDC	No	7.80
MB-0014	56.45	BCS	IDC	Yes	164.33
MB-0020	70.00	Mast.	ILC	Yes	22.40

In this study, we used three statistical methods—Maximum Likelihood Estimation (MLE), Quantiles, and Mean Absolute Deviation (MAD)—to estimate the parameters of the Weibull distribution for breast cancer survival data. The Weibull distribution is popular in survival analysis because it can model different survival time patterns and risk profiles (Muhammad, 2024). Our main aim was to see how well each method predicted the median survival time and scale parameter α (Hirst, 2021).

In this analysis, we use survival time data from breast cancer patients and randomly split the dataset into equal training and test sets (50%/50%). To ensure robust results, the training-testing process is repeated 1,000 times. In each iteration, considering the importance of median survival time in handling skewed data and outliers in medical and survival studies, we apply the Maximum Likelihood Estimation (MLE) and the Quantiles and Median Absolute Deviation (MAD) methods to estimate the scale parameter (α) of the Weibull distribution and predict the median survival rate of patients, using the test data for error analysis.

It is critical to analyze the scale parameter α and the mean error of the median prediction because these metrics directly affect the accuracy and reliability of survival predictions in the clinical setting. The scale parameter α determines the distribution of survival times; inaccurate estimates can lead to significant bias in predicting survival probabilities. Similarly, the predicted median survival time is a key metric in medical statistics, indicating how long half of the patients are expected to survive, and it serves as an important benchmark in survival studies. Minimizing the error in these two metrics is therefore critical to generating more accurate estimates and improving the overall reliability of survival models.

The results of the analysis are summarized in Table 3, showing the differences in performance between the Maximum Likelihood Estimation MLE, Quantiles, and MAD methods.

The median survival time predicted using the MLE method was 104.58 months with a mean prediction error of 12.15 months. The Quantiles method improved in prediction accuracy, predicting a median of 116.68 months with a mean error of 4.22 months. The MAD method, on the other hand, provided the closest prediction to the actual median, with a median prediction of 116.58 months and a significantly lower mean error of 3.79 months.

Regarding the Weibull distribution parameters, the MAD method also showed higher accuracy than the other methods. The shape parameter k was estimated to be 1.56 by the

Table 3. MLE, Quantile, and MAD results for Breast Cancer Survival Data.

Metric	MLE	Quantiles	MAD
Shape parameter k	1.56	1.51	1.51
Median Survival Time from prediction method (months)	104.58	116.68	116.58
Error in Predicted Median (months)	12.15	4.22	3.79
Error in Predicted Median (%)	10.75	3.70	2.46
Scale parameter α	132.18	148.83	148.73
Error in Scale parameter α	7.50	3.83	3.39
Error in Scale parameter α (%)	5.47	2.46	2.33

MLE method and 1.51 by both the Quantiles and MAD methods. For the scale parameter α, which determines the distribution of the survival time, the estimates were 132.18 by the MLE method, 148.83 by the Quantiles method, and 148.73 by the MAD method. For the mean error of α, the MAD method had the lowest error of 3.39, the Quantiles method had 3.83, and the MLE method had the highest error of 7.50. These results suggest that the MAD method is more robust than the MLE and Quantiles methods for predicting median survival time and scale parameters in cancer research and, therefore, can provide more accurate survival predictions in the clinical setting.

This finding has important implications for practical research. Accurate prediction of median survival time is critical in clinical survival analysis as it is a key metric for assessing treatment efficacy and patient prognosis (Hazim, 2025). The MAD method's ability to efficiently handle outliers ensures that predictions are not overly influenced by extreme survival times, making it a more robust tool for predicting patient outcomes. In contrast, despite its widespread use, the MLE method may not be as robust as the MAD method in biased datasets, leading to less accurate predictions in some cases.

7. Case Study II: Analysis of Insurance Claims Using MLE and MAD Methods

Data for this analysis come from an insurance claims dataset that captures demographic and clinical attributes, as well as claim amounts. The columns include patient age, gender, BMI (Body Mass Index), blood pressure, diabetic status, number of children, smoking status, region, and total claim amount. These factors are commonly used in actuarial modeling to understand claim severity and to identify risk drivers across demographics and geographic areas. For brevity, we employ several abbreviations in our tabular presentation: regions are abbreviated as Southeast (SE), Northwest (NW), and Southwest (SW); diabetic status and smoking status are recorded as Yes/No (Y/N). In this sample slice, *children* are all 0 and *smoker* is No for all rows; other values may appear in the full dataset. The sample is presented in Table 4. The selected columns are pertinent to claim severity modeling and facilitate comparisons across sex, metabolic factors (e.g., BMI, diabetic status), and region.

The results of the analysis are summarized in Table 5, showing the differences in performance between the Maximum Likelihood Estimation (MLE), Quantiles, and MAD methods.

Table 4. Sample of Insurance Claims (Selected Records)

Patient ID	Age	Gender	BMI	Blood Pressure	Diabetic	Children	Smoker	Region	Claim
1	39.0	Male	23.2	91	Yes	0	No	SE	1,121.87
2	24.0	Male	30.1	87	No	0	No	SE	1,131.51
8	19.0	Male	41.1	100	No	0	No	NW	1,146.80
19	49.0	Male	35.4	97	Yes	0	No	SW	1,263.25
25	50.0	Female	20.8	85	Yes	0	No	SE	1,607.51
29	58.0	Female	31.1	87	No	0	No	SE	1,621.88
31	29.0	Male	20.4	80	Yes	0	No	NW	1,625.43
40	49.0	Female	39.8	100	Yes	0	No	SE	1,633.96

Table 5. MLE, Quantile, and MAD results for Insurance Claims Data.

Metric	MLE	Quantiles	MAD
Shape parameter (k)	1.15	1.23	1.15
Median Claim from prediction method (USD)	9,641.87	9,341.35	9370.01
Absolute Error in Predicted Median (USD)	483.84	450.84	427.56
Relative Error in Predicted Median (%)	5.16	4.81	4.56
Scale parameter (α)	13,235.47	12,655.27	11,356.77
Absolute Error in Scale parameter (α) (USD)	812.25	873.21	519.36
Relative Error in Scale parameter (α) (%)	6.13	6.90	4.57

In this study, we applied Maximum Likelihood Estimation (MLE), Quantiles, and Mean Absolute Deviation (MAD) to fit a Weibull distribution to insurance claim amounts. The Weibull distribution is widely used in actuarial modeling because it accommodates skewed and heavy-tailed data often observed in claim severity (Hamza, 2023). We focused on two key metrics: the median claim amount, which provides a robust measure of central tendency, and the scale parameter α, which governs overall claim dispersion and impacts pricing and capital planning.

The dataset was split into equal training and test subsets, and the estimation process was repeated multiple times to ensure stable averages. As shown in Table 5, the MAD method yielded the smallest relative errors for both the predicted median (4.56%) and scale parameter α (4.57%), outperforming MLE and Quantiles. These results highlight MAD's robustness in handling variability and outliers, making it a practical alternative for improving the reliability of insurance risk models and premium estimation.

8. Conclusion

In this paper, we presented the formula for the Mean Absolute Deviation (MAD) about the median of the Weibull distribution and provided the calculations for the scale and shape parameters using three methods: Maximum Likelihood Estimation (MLE), the Quantiles method, and the Mean Absolute Deviation (MAD) around the median method.

The application of these methods to real-world data underscores the importance of selecting an appropriate statistical approach based on the dataset's characteristics. In survival analysis, where outcomes can vary widely, methods such as MAD that reduce the influence

of outliers yield more reliable and clinically relevant predictions. Similarly, in actuarial science, modeling insurance claim amounts with MAD improves parameter estimation robustness, particularly for heavy-tailed or skewed distributions where extreme values are common. While MLE and Quantiles remain valuable techniques, they are more sensitive to outliers and may produce less stable estimates. Therefore, for future research involving survival data or financial risk modeling of insurance claims, particularly when the data follow a Weibull distribution, the MAD method based on the median is recommended as the most accurate and robust approach for parameter estimation and prediction.

Acknowledgements

Conflict of Interest: We declare that there are no conflicts of interest regarding the publication of this paper.

Author Contributions: All the authors contributed equally to the effort.

Funding: This research was conducted without any external funding. All aspects of the study, including design, data collection, analysis, and interpretation, were carried out using the resources available within the authors' institution.

Data Availability (including Appendices): All the relevant data, Python code for analysis, detailed annual tables, and graphs are available via: https://github.com/vickyzhang7/Mean-Absolute-Deviation-for-Weibull-Distribution

References

Almeida, J. B., (1999). Application of Weibull Statistics to the Failure of Coatings. *Journal of Materials Processing Technology*. https://doi.org/10.1016/S0924-0136(99)00177-6.

Balakrishnan, N., Kateri, M., (2008). On the Maximum Likelihood Estimation of Parameters of Weibull Distribution Based on Complete and Censored Data. textitStatistics & Probability Letters. https://doi.org/10.1016/j.spl.2008.05.019.

Bloomfield, P., Steiger, W. L., (1984). *Least Absolute Deviations*, https://doi.org/10.1007/978-1-4684-8574-5.

Cancho, V. G., Macera, M. A. C., Suzuki, A. K., Louzada, F., and Zavaleta, K. E. C., (2020). A New Long-term Survival Model with Dispersion Induced by Discrete Frailty. *Lifetime Data Analysis*, 26(2). https://doi.org/10.1007/s10985-019-09472-2.

Carroll, K .J., (2003). On the Use and Utility of the Weibull Model in the Analysis of Survival Data. *Controlled Clinical Trials*. https://doi.org/10.1016/S0197-2456(03)00072-2.

Chen, Q., Gerlach, R., (2011). *The Two-sided Weibull Distribution and Forecasting Financial Tail Risk*, University of Sydney Business School, Discipline of Business Analytics.

Cohen, A. C., (1965). Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and Censored Samples. *Technometrics*. https://doi.org/10.1080/00401706.1965.10490300.

Cohen, C. A., Whitten, B., (1982). Modified Maximum Likelihood and Modified Moment Estimators for the Three-Parameter Weibull Distribution. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610928208828412.

Feller, J., (1956). *Probability Theory and Applications*. https://doi.org/10.1016/S0895-7177(01)00109-1.

Fogliatto, M. S. S., Santos, T. M. O., Bessani, M., and Maciel, C., (2019). *Survival Analysis of Electrical Power Distribution Systems Using Weibull Regression*, Proceedings of the 2019 SBAI. https://doi.org/10.17648/sbai-2019-111513.

Gebizlioglu, O. L., Senoglu, B., and Kantar, Y. M., (2011). Comparison of Certain Value-at-Risk Estimation Methods for the Two-Parameter Weibull Loss Distribution. *Journal of Computational and Applied Mathematics*. https://doi.org/10.1016/j.cam.2011.01.044.

Hamza, A., (2023). A Bayesian Approach to Weibull Distribution with Application to Insurance Claims Data. textitJournal of Reliability and Statistical Studies. https://doi.org/10.13052/jrss0974-8024.1611.

Hazim, K., Abdul Sada, M. T., (2025). The Survival Power Weibull Distribution With Application, *Statistics, Optimization & Information Computing*. https://doi.org/10.19139/soic-2310-5070-2318.

Hirst, T. C., Sena, E. S., and Macleod, M. R., (2021). Using Median Survival in Meta-analysis of Experimental Time-to-Event Data. *Systematic Reviews*, 10. https://doi.org/10.1186/s13643-021-01824-0.

Hortobagyi, G. N., Stemmer, S. M., Burris, H. A., Yap, Y.-S., and Sonke, G. S., (2022). Overall Survival with Ribociclib plus Letrozole in Advanced Breast Cancer. *New England Journal of Medicine*, 386(10). https://doi.org/10.1056/nejmoa2114663.

Imran, M., Alsadat, N., (2024). The Development of an Extended Weibull Model with Applications to Medicine, Industry and Actuarial Sciences. *Scientific Reports*. https://doi.org/10.1038/s41598-024-61308-8.

Jacquelin, J., (1993). Generalization of the Method of Maximum Likelihood (Insulation Testing). *IEEE Transactions on Electrical Insulation*. https://doi.org/10.1109/14.192241.

Jokiel-Rokita, A., Piatek, S., (2024). Estimation of Parameters and Quantiles of the Weibull Distribution. *Statistical Papers*. https://doi.org/10.1007/s00362-022-01379-9.

Kang, D., Ko, K. and Huh, J., (2018). Comparative Study of Different Methods for Estimating Weibull Parameters: A Case Study on Jeju Island, South Korea. *Energies*. https://doi.org/10.3390/en11020356.

Olver, F. W., Lozier, D. W., Boisvert, R. F., and Clark, C. W., (2010). NIST Handbook of Mathematical Functions, *Cambridge University Press*, New York.

Pham-Gia, T., Hung, T. L., (2001). *The Mean and Median Absolute Deviations*. https://doi.org/10.1090/S0002-9947-1956-0090927-3.

Pinsky, E., Zhang, W. and Wang, Z., (2024). Pareto Distribution of the Forbes Billionaires. *Computational Economics*. https://doi.org/10.1007/s10614-024-10730-1.

Pobockova, I., Sedliackova, Z., (2014). Comparison of Four Methods for Estimating the Weibull Distribution Parameters. *Applied Mathematical Sciences*. https://doi.org/10.12988/ams.2014.45389.

Quiroz Flores, A., (2024). Machine Learning for Survival Analysis. *In Oxford Handbook of Engaged Methodological Pluralism in Political Science*, Vol. 1. https://doi.org/10.1093/oxfordhb/9780192868282.013.48.

Teimouri, M., Hoseini, S. M. and Nadarajah, S., (2011). Comparison of Estimation Methods for the Weibull Distribution. textitStatistics. https://doi.org/10.1080/02331888.2011.559657.

Welch, B. L., Johnson, N. L. and Kotz, S., (1972). Distributions in Statistics: Continuous Univariate Distributions. *Journal of the Royal Statistical Society. Series A (General)*, 135(3). https://doi.org/10.2307/2344623.

Yoosefi, M., Baghestani, A. R. and Khadembashi, N., (2018). Survival Analysis of Colorectal Cancer Patients Using Exponentiated Weibull Distribution. *International Journal of Cancer Management*, 11(3). https://doi.org/10.5812/ijcm.8686.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 15–30, https://doi.org/10.59139/stattrans-2026-014
Received – 27.03.2025; accepted – 03.12.2025

The entanglement of attitudes toward inequality: the theoretical background and measurement for the EU countries in 2021

Stanislaw Maciej Kot[1]

Abstract

This paper assumes two types of social planners who evaluate income distributions in terms of social welfare, economic inequality, and poverty. The first type, SP_ε, denotes individuals who have an aversion to income inequality as measured by the normative parameter ε. The second, SP_ν, comprises individuals who have an aversion to rank inequality, as measured by the normative parameter ν. Since every member of a society may play the role of a social planner, there could be as many levels of ε and ν as there are society members. It raises the question of which ranges of v and ε values are ethically sensible when conducting empirical welfare studies. This paper proposes the answer to this question by introducing the concepts of inequality-entangled SP_ν and SP_ε. If a randomly selected SP_ν had ν_i, one could automatically find ε_i of the inequality-entangled SP_ε, and vice versa. The inequality-entangled SP_ν and SP_ε consistently evaluate inequality, social welfare, and poverty. This paper moreover proposes a method for eliciting the pairs (ν_i, ε_i), $i = 1, 2, \ldots, n$, from empirical income distributions. Moreover, a single pair (ν^*, ε^*) exists, representing all n pairs. Additionally, the study applies the inequality-entanglement methodological framework to assess social welfare, inequality and poverty for 27 European Union member countries in 2021.

Key words: income distribution, social welfare, inequality, poverty, inequality aversion, European Union.

1. Introduction

Applied welfare economics delegates the measurement of social welfare embodied in income distributions to an abstract *social planner* (*SP*) who uses individual *social evaluation functions* (*SEF*). Every member of society may play the role of the social planner with the same probability (Harsanyi, 1980).

This paper assumes two types of social planners: individuals who have an aversion to *income inequality* (SP_ε) and individuals who have an aversion to *rank inequality* (SP_ν). An SP_ε uses Atkinson's (1970) index of income inequality $A(\varepsilon)$, where normative parameter $\varepsilon > 0$ reflects *aversion to income inequality.* The greater the value of ε, the

[1] Gdansk University of Technology, Gdansk, Poland. E-mail: skot@zie.pg.gda.pl.
ORCID: https://orcid.org/0000-0000-0002-5875-6498.

more sensitive SP_ε is to *income differences.* An SP_ν uses the extended (generalized) Gini index $G(\nu)$ (Yitzhaki, 1983; Kakwani, 1980; Donaldson and Weymark, 1980). The normative parameter $\nu \geq 1$ reflects an *aversion to rank inequality*. The greater the value of ν, the more sensitive SP_ν is to *rank differences,* regardless of the exact value that income may take at that rank (Duclos, 2000).

As every member of society may be a social planner with the same probability, there could be as many values of ε and ν as society members. It raises the question of which ranges of ν and ε values are ethically sensible when conducting empirical welfare studies (Duclos, 2000).

This paper proposes an answer to this question by introducing the concept of *entanglement of attitudes toward inequality* (hereafter, *inequality entanglement*). Let V and $\mathcal{E}$ be the sets of admissible levels of ν and ε, respectively, and let $\boldsymbol{V}\boldsymbol{x}\boldsymbol{\mathcal{E}}$ be the Cartesian product of V and $\mathcal{E}$. The economic theory allows for (resp. does not prohibit) the existence of pairs of SP_ν and SP_ε (resp. ν and ε) who consistently assess inequality in a given income distribution, namely that the following equality holds:

$$G(\nu) = A(\varepsilon) \tag{1}$$

for all pairs $(\nu, \varepsilon) \in \boldsymbol{V}\boldsymbol{x}\boldsymbol{\mathcal{E}}$. We will call the pairs (SP_ε, SP_ν), or (ν, ε), satisfying Eq. (1), *inequality-entangled social planners.*

The concept of inequality-entangled social planners has various advantages. Such planners consistently assess income inequality. We will show that they also consistently assess social welfare and poverty in income distributions.

In applications, the concept of inequality-entangled social planners significantly narrows the range of ν and ε values. We will show that there is a unique pair $(SP_{\nu^*}, SP_{\varepsilon^*})$, resp. (ν^*, ε^*), representing all inequality-entangled pairs of social planners. We propose a method for estimating the pairs (ν, ε) and (ν^*, ε^*) from income data.

What are the rationales for applying quantum physics concepts? Orrell (2024) notices that neoclassical economics had roots in classical mechanics. The influence of mechanics persists in concepts such as the static equilibrium and the idea that people behave as independent, rational utility maximisers. However, economics shaped by uncertainty, dynamism and entanglement might be more applicable to the real world (Facco & Fracas, 2022). Section 2 explains the use of such a quantum physics metaphor in more detail.

The remainder of this paper is organized as follows. Section 2 introduces basic concepts and formulae. This Section also offers a brief literature review. Section 3 describes the method for estimating pairs (ν, ε). Section 4 comprises the first part of the empirical results. After describing the EU-SILC income data, this Section presents

estimates of the pairs (ν^*, ε^*), social welfare, and economic inequality for 27 EU countries in 2021. Section 5 offers estimates of poverty. Section 6 concludes.

2. Quantum entanglement of particles and inequality entanglement of social planners

A phenomenon in which some social planners satisfy Eq. (1) resembles, metaphorically, the *quantum entanglement of particles*. Suppose two distinct quantum states, q_ν or q_ε, characterize some subatomic particles. The quantum state of a particle is unknown before measurement.

Quantum entanglement is the phenomenon of a system of particles such that the measurement of one particle's quantum state, say q_ν, *automatically* provides the measurement of its companion's state, say q_ε, even when a vast distance separates the particles.

An example of entanglement is a subatomic particle that decays into an entangled pair of other particles. The decay events obey the various conservation laws. As a result, the measurement outcomes of one particle must be highly correlated with the measurement outcomes of its companion particle, whereas the total momenta, angular momenta, energy, or the like remain the same before and after this process (Caltech Science Exchange 2024). Many entangled particles may exist.

Regarding social planners, note that if a randomly selected person is to play the role of a social planner, we do not know in advance whether they are SP_ε or SP_ν. *Inequality entanglement of the attitudes toward inequality* (*inequality entanglement*, for short) is a metaphor for the phenomenon of a group of social planners, so that just a measurement of ν (resp. ε) *automatically* gives the measurement of ε (resp. ν) due to Eq. (1). There may exist a multitude of inequality entangled pairs SP_ν and SP_ε.

If a person becomes an actual social planner, she should provide an unambiguous assessment of inequality in the analyzed income distribution. Suppose she announces: "$G(\nu)$ equals 0.32". It reveals that she is an SP_ν with an aversion to rank inequality equal to $\nu = G^{-1}(0.32)$. Then her *inequality-entangled* companion, SP_ε, should have $\varepsilon = A^{-1}(0.32)$. On the other hand, if the selected person's answer was: "$A(\varepsilon) = 0.32$", an entangled person should have $\nu = G^{-1}(0.32)$. There may exist a multitude of *inequality-entangled* SP_ν and SP_ε.

A change from $G(\nu)$ into $A(\varepsilon)$ (and vice versa) *preserves income inequality*. In other words, Eq. (1) plays the role of a *conservation law*. Moreover, such a change also preserves social welfare and poverty. To see this, note that for an analyzed income distribution with $\mu > 0$, Eq. (1) is equivalent to:

$$\mu[1-G(\nu)] = \mu[1-A(\varepsilon)]. \tag{2}$$

In Eq. (2), $\mu[1-G(\nu)]$ and $\mu[1-A(\varepsilon)]$ are the *Social Evaluation Functions*, SEF_ε and SEF_ν implied by $G(\nu)$ and $A(\varepsilon)$, respectively. Thus, *inequality-entangled* social planners consistently assess social welfare. It is worth adding that $\mu[1\text{-}G(\nu)]$ and $\mu[1-A(\varepsilon)]$ in Eq. (2) are the *Equally Distributed Equivalent Incomes (EDEI),* which, if received by all persons, give the same level of *social welfare* as the present distribution (Atkinson, 1970).

Eq. (1) also implies that inequality-entangled social planners consistently assess poverty in income distributions. A person is deemed poor if his/her income is less than a normative *poverty line z* established by a social planner. Kot and Paradowski (2024a) argue that the *EDEI* is an *upper limit* of any socially acceptable poverty line *z*, namely,

$$z \leq EDEI \tag{3}$$

If a social planner proposed a poverty line *z* greater than *EDEI*, attaining an egalitarian income distribution would be possible at the cost of common poverty. Arguably, no reasonable society would accept such a poverty line. Kot and Paradowski (2024) refer to such a peculiar situation as the *Equity-Poverty Trap*.

Note that Eq. (2) expresses the equality of poverty lines and, therefore, the equality of poverty indices, which are monotonic functions of a poverty line. Thus, inequality-entangled social planners consistently assess poverty in a given income distribution.

In this paper, we will use the following family of poverty indices:

$$FGT_\alpha = \sum_{x_i<z} \left(\frac{z-x_i}{z}\right)^\alpha p_i, \tag{4}$$

where *z* is the poverty line, x_i is income below *z,* and α is a normative parameter (Foster, Greer, and Thorbecke, 1984).

Some particular cases of Eq. (4), namely FGT_0, FGT_1, and FGT_2, are widely used. FGT_0, also known as the head-count *ratio*, measures poverty incidence. FGT_1, called the *poverty depth*, measures the poverty of society as a whole (Foster and Shorrocks, 1991). FGT_2 measures *poverty severity*.

3. Parametric utility functions and social evaluation functions

3.1. Personal and moral preferences

According to Harsanyi (1980, pp. IX-X), each individual has two kinds of preferences. The first kind comprises his *personal preferences*, which are defined as his actual preferences based on his interests. The second one consists of *moral* preferences defined as a person's "(…) *hypothetical* preferences that he *would* entertain if he forced himself to judge the world from a moral, i.e. from an impersonal and impartial point of view". More specifically, "(…) *moral* preferences are those preferences that he would

entertain if he assumed to have the same probability $1/n$ to be put in place of any one of the n individual members of society" (Harsanyi, 1980, pp. IX). Mathematically, an individual's personal preferences are represented by his *utility function,* whereas his *social evaluation function* represents his moral preferences.

Concerning moral preferences, a rational individual would try to maximize his expected utility, thereby maximizing the average utility of the individual members of society. It means that a rational individual will always use the average utility level in society as his social evaluation function.

Harsanyi (1980, p. X) noticed that this definition of social evaluation functions presupposes the possibility of *interpersonal comparisons* of utility. He argued that "(…) interpersonal utility comparisons are essentially the same kind of mental operation as intrapersonal utility comparisons are."

3.2. The social evaluation function of averters to income inequality

In this paper, we will use the following terms and symbols. The positive valued random variable X, with the distribution function $F(x) = \mathrm{P}(X \leq x)$, will describe the distribution of personal incomes. We assume that the mean $\mu = \mathrm{E_F}[\mathrm{X}]$ exists and is finite, where the operator $\mathrm{E_F}[\cdot]$ is the mathematical expectation of X with respect to $F(x)$.

We assume that an averter to income inequality uses the utility function of the form:

$$u(x) = \begin{cases} \frac{\mathrm{x}^{1-\varepsilon}}{1-\varepsilon}, for\ \varepsilon \neq 1 \\ lnx, for\ \varepsilon = 1 \end{cases}, x>0, \tag{5}$$

(Atkinson, 1970). Eq. (5) defines the *utility function of constant relative inequality aversion* (*CRIA*).

For averters to income inequality, the Social Evaluation Function is the expected value of $u(X)$ with respect to the distribution F, namely:

$$SEF_\varepsilon = E_F[u(X)] = \begin{cases} \frac{E_F[\mathrm{X}^{1-\varepsilon}]}{1-\varepsilon}, for\ \varepsilon \neq 1 \\ E_F[ln\mathrm{X}], for\ \varepsilon = 1 \end{cases}, \tag{6}$$

(Atkinson, 1970).

The parameter ε measures *a social planner's or society's aversion to income inequality*. When $\varepsilon < 0$, a social planner or society is *averse to equality*. Null inequality aversion, i.e. $\varepsilon = 0$, characterizes an *inequality-neutral* society. In this case, $SWF_0 = \mu$, *and value judgments about income distributions are based only on mean incomes, providing no information about* income inequality. Thus, income distribution X with the mean μ_x is preferred over Y with the mean μ_y if and only if $\mu_x > \mu_y$. If $\varepsilon > 0$, *society is inequality-averse*. Hereafter, we will assume $\varepsilon \geq 0$.

Knowledge of ε is essential for various reasons. As ε ultimately determines the function (5), it enables a direct measurement of SEF_ε. Parameter ε expresses the rate at which a society solves the trade-off between efficiency and equality. As the (minus) elasticity of the marginal utility of income, ε, also has a central role in public Economics: high values of ε mean that the marginal utility of income declines as income grows, and therefore, an income transfer from the rich to the poor is increasingly desirable (Young, 1990). Knowledge of ε is also essential for appraising social projects and policies that impact different socioeconomic groups (Evans, 2005; Layard et al., 2008; Aristei and Perugini, 2016).

Based on the interpretation of ε as inequality aversion, Atkinson (1970) proposed the normative index of inequality:

$$A(\varepsilon) = \frac{\mu - \mu_\varepsilon}{\mu}, \tag{7}$$

where μ_ε is *the equally distributed equivalent income* (*EDEI*) that, if distributed equally, gives the value of the social evaluation function E[$u(X)$] the same as the initial distribution (Kolm, 1969; Atkinson, 1970; Sen, 1973). In general, *EDEI* is the solution to the equation: $u(EDEI)$ = E[$u(X)$] for a given utility function $u(x)$.

For the utility function (5) and social welfare function (6), *EDEI* has the following form:

$$\mu_\varepsilon = \begin{cases} \{\mathrm{E}[u(X)]\}^{1/(1-\varepsilon)}, & for\ \varepsilon \neq 1 \\ \exp\{E[lnX]\}, & for\ \varepsilon = 1 \end{cases}. \tag{8}$$

For ε=1, μ_ε is the geometric mean, and for ε = 2, μ_ε is the harmonic mean. For a given income distribution, μ_ε is a declining function of ε (Lambert, 2001, Chapter 4).

It follows from (7) and (8) that *EDEI* is a money metric of the social evaluation function *SEF,* namely:

$$\mu_\varepsilon = \mu[1\text{-}A(\varepsilon)], \tag{9}$$

(Atkinson, 1970). Eq. (9) specifies the family $\{SEF_\varepsilon\}_{\varepsilon \geq 0}$ of social evaluation functions indexed by ε.

3.3. The social evaluation function of averters to rank inequality

Sen (1973, p. 41) argued that the social value of the welfare of individuals should depend crucially on the levels of welfare (or incomes) of others. The following social evaluation function satisfies this claim:

$$\mu_\nu = \mu[1\text{-}G(\nu)\ ], \tag{10}$$

where $G(\nu)$ is the extended Gini index of the form:

$$G(\nu) = 1 - \nu(\nu - 1)\int_0^1 (1-p)^{\nu-2} L(p)dp\ , \nu \geq 1, p \in [0,1], \tag{11}$$

(Yitzhaki, 1983; Donaldson and Weymark, 1980; Kakwani, 1980). In Eq. (11), $L(p)$ is the Lorenz curve, $1-p = 1-F(x)$ is the rank of a person with income x, and ν is a normative parameter expressing *aversion to rank inequality*. The case $0 \leq \nu < 1$ reflects *rank equality aversion,* $\nu = 1$ *rank equality neutral*, and $\nu \geq 1$ *rank inequality aversion* (Yitzhaki, 1983; Duclos, 2000). For $\nu = 2$, $G(\nu)$ is the ordinary Gini index.

Eq. (10) defines the family $\{SWF_\nu\}_{\nu>1}$ of social welfare functions indexed by $\nu > 1$. More specifically:

$$\mu[1 - G(\nu)] = \nu \int_0^\infty x[1 - F(x)]^{\nu-1} f(x)dx, \tag{12}$$

(Lambert, 2001, p. 125).

Yitzhaki (1983) noted that $G(\nu)$ (11) has most of the properties of Atkinson's index (7). Indeed, at the extremes $\nu \to 1$ and $\nu \to \infty$, the behavior of $G(\nu)$ resembles that of the $A(\varepsilon)$ at the extremes $\varepsilon \to 0$ and $\varepsilon \to \infty$ of inequality aversion (Lambert, 2001, p. 115). As $\nu \to 1$, $G(\nu) \to 0$. As $\nu \to \infty$, $G(\nu) \to 1\text{-}L'(0)$. For a discrete distribution of X, $G(\nu) \to 1\text{-}x_{min}/\mu$ as $\nu \to \infty$.

4. Estimating aversion to income inequality and rank inequality.

4.1. The previous methods of estimating ε and ν

The literature offers various methods of recovering ε from empirical data. Inequality aversion ε has been elicited from Okun's "leaky bucket" experiment (Okun, 1975). In this experiment, participants subjectively assess a tolerable 'leakage' of money due to administrative costs during income transfers among individuals. The higher the leakage the participants permit, the greater their aversion to income inequality.

One can elicit ε from *the equal sacrifice model* (Richter, 1983; Vitaliano, 1977; Young, 1987). Lambert et al. (2003) elicit ε by hypothesizing about *the natural rate of subjective inequality*. Kot (2020) proposes the estimator of $\hat{\varepsilon} = (ap + 1)/2$ when income obeys the Generalised Beta distribution of the second kind GB2(x; *a, b, p, q*) (McDonald, 1984).

Much less is known concerning the range of ν that analysts may apply in empirical studies. Kot (2022) analyses the empirical relationship between the generalized Gini index $G(\nu)$ and three Italian indices of inequality, namely the Pietra (1915) index, the Bonferroni (1930) index, and the Zenga (2007) index. The author finds these indices corresponding to $G(\nu)$ with ν equal to 1.5, 3, and 11, respectively.

Duclos (2000) recommends the leaky bucket experiment for deriving ν. The author argues that ν should not exceed 4 in empirical analyses if the whole transfer is not licked. In this paper, we will follow Duclos' recommendation for an upper limit of ν.

The joint estimation of ε and ν has not yet been analyzed, with one exception. Recently, Kot and Paradowski (2024b) obtained the pairs (ε^*, ν^*) for ten Latin American and Caribbean countries by solving the system of two nonlinear equations: SWF_ε=SWF_ν and $x_\varepsilon^* = x_\nu^*$, where x_ε^* and x_ν^* are the *benchmark incomes* of SP_ε and SP_ν, respectively. The authors acknowledged that their method requires further improvement.

4.2. The mean-value method

We propose a three-stage method for estimating the pairs (ν^*, ε^*). In the first stage, we generate n random values of aversion to rank inequality $\nu_1, \nu_2,\ldots,\nu_n$ from the uniform distribution U[1,4] and estimate the sequence of n extended Gini indices, $G(\nu_1)$, …, $G(\nu_n)$ for an analyzed income distribution.

To justify this stage, note that the state of complete ignorance concerning the value of ν means the state of *maximum entropy*. The uniform distribution has the maximum entropy among all probability distributions defined on finite intervals (Cover and Thomas, 1991, p. 269). Thus, one may expect ν to follow a uniform distribution U[1,4].

In the second stage, we calculate n values of ε_i as the solutions to Eq. (1) for $G(\nu_i)$. Thus, we get n possible pairs (ν_i, ε_i) of inequality-entangled social planners. We refer to the graph of the pairs as the *ν-ε curve* or the *ε(ν)* function. Fig. 1 illustrates the *ν-ε* curves for some European countries.

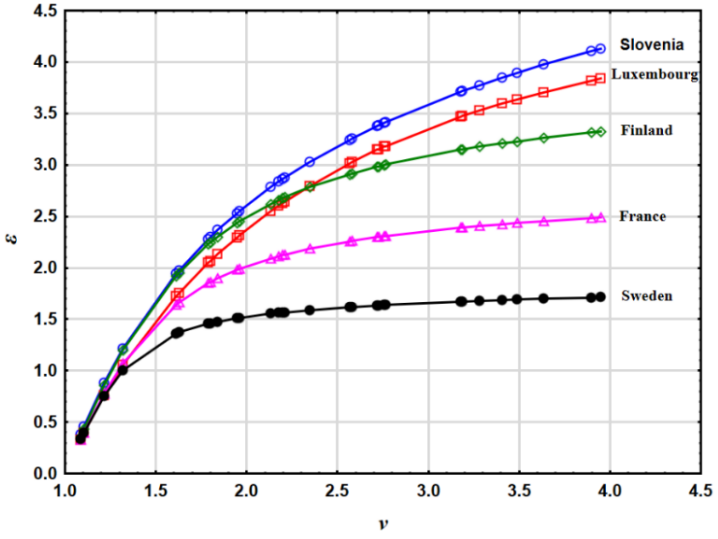


Figure 1. The *ν-ε* curves for selected UE countries in 2021

Source: *own work*.

Every point in Fig. 1 represents a combination of *exogenous* ν and corresponding ε, guaranteeing the same inequality assessment in a given income distribution by a pair

of inequality-entangled social planners. At the upper limit of $\nu = 4$, the curves in Fig. 1 attain different levels, which depend on the country's income distribution.

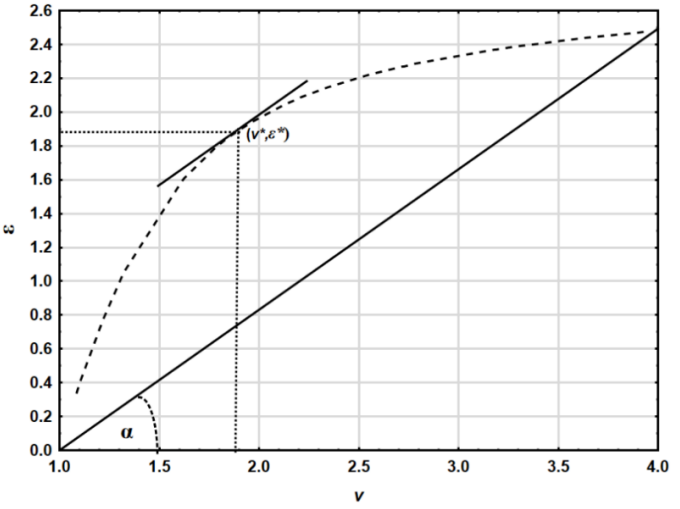


Figure 2. A single representation (ν^*,ε^*) of the ν-ε curve for Poland, 2021

Source: *own work*.

In the third stage, we search for a unique pair (ν^*, ε^*) representing all inequality-entangled pairs (ν_i, ε_i). The ν-ε curve in Fig. 2 illustrates the concept of this stage.

In Fig. 2, the tangent of angle α reflects *the average proportion* of ε to ν that gives the same inequality assessment. More specifically:

$$\tan(\alpha) = \frac{\varepsilon(\nu_{max}) - \varepsilon(\nu_{min})}{\nu_{max} - \nu_{min}}, \tag{13}$$

where the $\nu_{min} \geq 1$, and $\nu_{max} = 4$. The corresponding $\varepsilon(\nu_{min})$ and $\varepsilon(\nu_{max})$ are calculated from Eq. (1).

If the function $\varepsilon(\nu)$ is differentiable within the interval $[\nu_{min}, \nu_{max}]$, then Lagrange's mean value theorem implies the existence of a point ν^* inside this interval at which the following equation holds:

$$tan(\alpha) = \varepsilon'(\nu^*), \tag{14}$$

where $\varepsilon'(\nu^*)$ is the first derivative of $\varepsilon(\nu)$ at ν^*.

In Fig. 2, the point (ν^*, ε^*) reflects the *average level of income inequality aversion* for offsetting rank inequality aversion when attaining the same inequality assessment in a given income distribution. In this sense, the single point (ν^*, ε^*) represents all points on the ν-ε curve.

We propose to calculate the derivatives $\varepsilon'(v_i)$ analytically from a function fitted to the empirical v-ε curves. It has turned out that the following function approximates the v-ε curves "almost ideally":

$$\varepsilon(v) = \theta_1 + \theta_2 v - \exp\{\theta_3 - \theta_4 v\}, \tag{15}$$

where the parameters $\theta_1, \theta_2, \theta_3$ and θ_4 can be estimated using the nonlinear least squares method.

We use the ratio of the regression sum of squares to the total sum of squares as a measure of goodness-of-fit. As this ratio explains the proportion of variance accounted for in the dependent variable (ε) by the model (15), this is equivalent to the coefficient of determination R^2 ($0 \le R^2 \le 1$).

Equating the derivative of (15) with *tan(α)* gives:

$$\tan(\alpha) = \theta_2 + \theta_4 \exp\{\theta_3 - \theta_4 v^*\}. \tag{16}$$

After simple algebra, we get:

$$v^* = \left\{\theta_3 - \log\left[\frac{(tan(\alpha)-\theta_2)}{\theta_4}\right]\right\}/\theta_4. \tag{17}$$

The corresponding parameter ε^* can be calculated from Eq. (15).

By substituting the sample estimates for the parameters θ_2, θ_3, and θ_4 in equation (17), we obtain the estimator of v^*. We shall refer to this way of obtaining (v^*, ε^*) as *the mean-value method* (MVM).

5. Empirical results for the EU-member countries 2021

5.1. Statistical data

We use statistical data on household disposable income [in Euros] from the EU-SILC database for 2021. To obtain a distribution of personal disposable incomes, we adjust household incomes by the square-root equivalence scale (Buhmann et al., 1988). Such an adjustment requires weighing the resulting equivalent incomes. We follow the common practice of weighting adjusted incomes by household size. The final weights applied in this paper are products of household size and cross-sectional survey weights.

We generate the sequence of 30 random values of v_i from the uniform distribution U[1,4]. Then we estimate the sequence of $G(v_i)$, i = 1, …, 30, for every country and the corresponding sequence of ε_i, solving Eq. (1) numerically using the IMSL Fortran subroutine NEQNF. Next, we estimate the parameters θ_1, θ_2, θ_3 and θ_4 of the nonlinear function (15) for every country using the pairs (ε_i, v_i), i=1,…, 30 and the IMSL Fortran subroutine RNLIN.

5.2. Estimates of v^* and ε^*

Table 1 presents the results of applying the MVM to EU-SILC data. Besides estimates of ε^* and v^*, this Table contains the estimates of the generalized Gini index $G(v^*)$ (equal to the Atkinson index $A(\varepsilon^*)$) and *EDEI*. The last column of this Table (labelled as R^2) contains the values of the coefficient of determination. For further comparison, this Table also includes the mean equivalent income. The last row of this Table (labelled 'EU total') comprises results for all EU member countries' incomes and weights.

Table 1. Estimates of income inequality $G(v^*) = A(v^*)$ and social welfare, *EDEI* [€] based on the inequality-entangled estimates of aversion to rank inequality v^* and income inequality ε^* for EU-member countries in 2021

No.	Country	v^*	ε^*	*GA*	*EDEI* [€]	Mean_D [€]]	R^2
1	Austria	1.71847	1.30212	0.22367	25637	33023	0.9997
2	Belgium	1.88000	2.05323	0.22847	23435	30375	0.9999
3	Bulgaria	2.06051	1.95660	0.40729	4463	7530	1.0000
4	Croatia	1.94088	1.76289	0.28500	7306	10218	1.0000
5	Cyprus	1.94303	2.21251	0.28839	15711	22078	0.9997
6	Czechia	1.99332	2.52619	0.25467	9854	13221	1.0000
7	Denmark	1.78069	1.68096	0.24039	29321	38600	0.9998
8	Estonia	1.90745	1.67218	0.29094	11044	15576	1.0000
9	Finland	1.97727	2.46902	0.26103	22752	30789	1.0000
10	France	1.87692	1.92728	0.27535	20694	28557	0.9999
11	Germany	1.87262	1.79843	0.28716	22548	31631	0.9997
12	Greece	1.89762	1.66411	0.29579	7975	11325	0.9999
13	Hungary	1.86140	1.75326	0.25622	6064	8153	0.9999
14	Ireland	1.75840	1.67332	0.23879	27216	35754	0.9997
15	Italy	1.79748	1.38777	0.28947	16030	22561	0.9999
16	Latvia	1.93202	1.60822	0.34519	8153	12451	1.0000
17	Lithuania	1.99808	1.84871	0.35679	8346	12976	1.0000
18	Luxembourg	2.11278	2.52324	0.30744	36250	52342	1.0000
19	Malta	1.66892	1.31318	0.24643	16123	21395	0.9994
20	Netherlands	1.84791	1.87964	0.25099	19377	25870	0.9999
21	Poland	1.89022	1.88695	0.24884	7944	10576	1.0000
22	Portugal	1.93503	1.72011	0.31758	9978	14621	1.0000
23	Romania	1.95258	1.53072	0.32642	4197	6231	0.9999
24	Slovakia	1.92978	2.10286	0.21011	8210	10394	0.9999
25	Slovenia	2.02849	2.64598	0.24319	14044	18557	0.9999
26	Spain	1.84606	1.41535	0.29774	14259	20304	0.9999
27	Sweden	1.70576	1.41729	0.22131	23228	29830	0.9997
	EU total	**1.90295**	**1.52573**	**0.32298**	**16153**	**23859**	**1.0000**

Note: R^2 measures the goodness-of-fit of the model (15).
Symbol *GA* denotes the common level of inequality $G(v^*) = A(\varepsilon^*)$; *Mean_D* is the mean equivalent income.

Source: own calculations using EU-SILC data.

Inspecting R^2 in Table 1 shows an almost ideal fitting of the ε-v curves by function (15). Moreover, normative parameters ν and ε are country-specific. For instance, Spanish and Dutch social planners have a similar aversion to rank inequality, $\nu \approx 1.85$. However, Spanish inequality-entangled SP_ε should have $\varepsilon \approx 1.42$ to assess income inequality identically as his companion SP_ν did. On the other hand, the Dutch inequality-entangled SP_ε must have $\varepsilon \approx 1.88$ for the same purpose. Similarly, the last row of Table 1 indicates that a European SP_ν with $\nu = 1.90295$ assesses social welfare (*EDEI*) in the EU as €16,153. His inequality-entangled companion SP_ε provides the same welfare assessment when $\varepsilon = 1.52573$.

Table 2 provides additional information on the distributions of the characteristics presented in Table 1.

Table 2. Descriptive statistics of estimates in Table 1

Parameter	Mean	Median	Min.	Max.	Std. Dev.	V [%]	Skew-ness	Ku
$\nu^\star$	1.89310	1.89762	1.66892	2.11278	0.10703	5.65	-0.237	0.047
$\varepsilon^\star$	1.84193	1.76289	1.30212	2.64598	0.37636	20.43	0.674	0.191
GA	0.27758	0.27535	0.21011	0.40729	0.04588	16.53	0.949	1.059
EDEI	15561	14259	4197	36250	8532	54.83	0.611	0.422
Mean D	21293	20304	6230	52342	11404	53.56	0.787	0.357

Note: LB and *UB* are the lower and upper limits of the 95% confidence interval; *V* is the coefficient of variability; *Skew* is the coefficient of skewness; *Ku* is the kurtosis.

Source: own calculations using data from Table 1.

Examining Table 2 shows that $\nu^\star$ is statistically significantly less than 2.0 at a 0.05 significance level (*p-value* = 00001). Thus, analyzing income inequality using the standard Gini index, *G(2),* is debatable. Variability of this parameter is slightly lower than that of $\varepsilon^\star$. Moreover, the distribution of $\nu^\star$ across countries is negatively skewed, whereas the distribution of $\varepsilon^\star$ is positively skewed. Both distributions are flatter than the standard normal distribution.

Note that, in Table 2, the parameter means differ from those for "EU-total" in Table 1, except for $\nu^\star$. It might be because we calculated descriptive statistics of the parameter distributions across countries without weighting. The discussed differences are more evident in the case of inequality measures, since they are not additively decomposable.

5.3. Economic poverty in EU Member Countries in 2020

As mentioned in Section 2, inequality-entangled social planners consistently assess poverty in an income distribution. When we set a country's *EDEI* as a national poverty line z, the FGT_α indices (4) enable assessments of various aspects of the country's impoverishment.

EDEI, as an upper limit of poverty lines, inherently implies an *international poverty line*. If there were rationales for comparisons of poverty across N selected countries, an international poverty line, $z_{all,}$ should satisfy the following condition:

$$z_{int} = \min_i\{z_1, z_2, \dots, z_N\}, i = 1, 2, \dots, N \tag{18}$$

where z_1,…, z_N are the country's poverty lines equal to the countries' *EDEI*s. (Kot, Paradowski, 2024a). The international poverty line, as defined by (18), guarantees that no selected country falls into the *Equity-Poverty Trap*.

Table 3 presents estimates of the *FGTα* poverty indices (12) for α = 0, 1, and 2.

Table 3. Poverty in EU Member Countries in 2020.

No.	*Country*	***National poverty lines***			***International poverty line***		
		z_i= *EDEI*$_i$ *in Table 1*			z_{int}=4197 (Romania)		
		FGT_0	FGT_1	$FGT2_2$	FGT_0	FGT_1	$FGT2_2$
1	Austria	0.36581	0.10582	0.04873	0.01143	0.00787	0.00659
2	Belgium	0.34246	0.08842	0.03331	0.00386	0.00175	0.00105
3	Bulgaria	0.35380	0.11823	0.05542	0.31664	0.10431	0.04810
4	Croatia	0.33375	0.11037	0.05319	0.10984	0.03284	0.01533
5	Cyprus	0.36184	0.09453	0.03514	0.00255	0.00095	0.00058
6	Czechia	0.32988	0.07594	0.02769	0.01674	0.00413	0.00167
7	Denmark	0.36599	0.09620	0.03921	0.00443	0.00205	0.00142
8	Estonia	0.36033	0.11800	0.05462	0.03432	0.01215	0.00677
9	Finland	0.33782	0.08149	0.02879	0.00134	0.00045	0.00023
10	France	0.35090	0.09391	0.03790	0.00535	0.00244	0.00147
11	Germany	0.35968	0.10394	0.04514	0.00580	0.00197	0.00111
12	Greece	0.34707	0.11060	0.05256	0.08125	0.02651	0.01396
13	Hungary	0.34633	0.09805	0.04322	0.13057	0.03774	0.01787
14	Ireland	0.37738	0.10011	0.03874	0.00190	0.00145	0.00122
15	Italy	0.35964	0.12254	0.06170	0.02593	0.01086	0.00699
16	Latvia	0.35922	0.12914	0.06469	0.10611	0.03147	0.01564
17	Lithuania	0.35290	0.11525	0.05370	0.07271	0.02273	0.01138
18	Luxembourg	0.32825	0.09153	0.03611	0.00003	0.00002	0.00001
19	Malta	0.40029	0.12068	0.05332	0.01124	0.00647	0.00452
20	Netherlands	0.36097	0.09178	0.03624	0.00701	0.00283	0.00168
21	Poland	0.33929	0.09580	0.04091	0.06027	0.01658	0.00779
22	Portugal	0.34087	0.11130	0.05403	0.05156	0.01785	0.00974
23	Romania	0.33552	0.13102	0.07162	0.33552	0.13102	0.07162
24	Slovakia	0.31707	0.08255	0.03454	0.04662	0.01368	0.00636
25	Slovenia	0.31133	0.07859	0.02956	0.00386	0.00084	0.00029
26	Spain	0.35373	0.12800	0.06751	0.03944	0.01670	0.01030
27	Sweden	0.37262	0.10814	0.04663	0.00828	0.00425	0.00297
1-27	**EU total**	**0.34471**	**0.12646**	**0.06723**			

Source: own calculations using data from the EU-SILC database.

On the left panel of Table 3, one can see high poverty levels in all countries. It is worth noting that *EDEI*, as an upper limit of poverty lines, implies upper limits for poverty measures. The differences between countries' FGT_0, FGT_1, and FGT_2 estimates are relatively small. Thus, according to national poverty standards, all analyzed countries might be expected to have similar poverty incidence, depth, and severity. However, these results are only for *internal use* and do not support international comparisons.

When we apply Romania's *EDEI* of €4197 as the international poverty line (according to Eq. 18), the right panel of Table 2 shows a much greater diversity of poverty assessments across EU Member Countries than the left panel. Bulgaria and Romania are among the poorest countries, as measured by three poverty indices. On the other end of the spectrum, Luxembourg is the most affluent country.

The inequality entanglement is a conceptual novelty of this paper. It enables the elicitation of normative parameters ν and ε from empirical income distributions. Knowledge of these parameters enables consistent assessments of inequality by the social planners $SP\nu$ and $SP\varepsilon$, who employ different methodologies. The inequality entanglement also enables consistent assessments of social welfare and economic poverty.

6. Concluding remarks

The method of eliciting pairs (ν, ε) from income data may start with generating random numbers of ε instead of ν from U[1,4] distribution. If incomes obey the generalized beta distribution of the second kind GB2(a,b,p,q), Kot (2020) demonstrates that ε belongs to the $[0, ap+1]$ interval. Therefore, one can generate n non-random numbers from the U[0, ap+1] distribution and then calculate the inequality-entangled ν using Eq. (1).

Acknowledgements

The author thanks an anonymous referee for valuable comments and suggestions.

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aristei, D., Perugini, C., (2016). Inequality aversion in post-communist countries in the years of the crisis. *Post-Communist Economies*, 28(4), pp. 436–448.

Atkinson, A. B., (1970). On the measurement of inequality. *Journal of Economic Theory,* 2, pp. 244–263.

Bonferroni, C., (1930). *Elementi di Statistica Generale*. Seeber, Firenze.

Buhmann, B., Rainwater, L., Schmaus, G. and Smeeding, T. M., (1988). Equivalence scales, well-being, inequality, and poverty: sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database. *Review of Income and Wealth*, 34(2), pp. 115–142.

Caltech Science Exchange, (2024). What is entanglement, and why is it important? https://scienceexchange.caltech.edu/topics/quantum .

Cover, T. M., Thomas, J. A., (1991). Maximum entropy and spectral estimation. *Elements of Information Theory*, pp. 266–278.

Donaldson, D., Weymark, J., (1980). Aa single-parameter generalisation of Gini indices of inequality. *Journal of Economic Theory*, 22, pp. 67–86.

Duclos, J. Y., (2000). Gini indices and the redistribution of income. *International Tax and Public Finance*, *7*(2), pp. 141–162.

Evans, D., (2005). The elasticity of marginal utility of consumption: Estimates for 20 OECD countries. *Fiscal Studies,* 26, pp. 197–224.

Facco, E., Fracas, F., (2022). De Rerum (Incerta) Natura: A Tentative Approach to the Concept of "Quantum-like". *Symmetry*, 14(3), p. 480. https://doi.org/10.3390/sym14030480

Foster, J. E., Shorrocks, A. F., (1991). Subgroup consistent poverty indices. *Econometrica,* pp. 687–709.

Foster, J., Greer, J. and Thorbecke, E., (1984). A class of decomposable poverty measures. *Econometrica*, 52(3), pp. 761–766.

Harsanyi, J. C., (1980). Essays on ethics, social behavior, and scientific explanation. *Theory and Decision Library*, Vol. 12, Kluwer Academic Publishers Group, Dordrecht, Holland.

Kakwani, N. C., (1980). *Income inequality and poverty*. World Bank, New York.

Kolm, S. C., (1969). The optimal production of social justice. In Margolis, J. & Guitton, H. (Eds.). *Public Economics: An Analysis of Public Production and Consumption and their Relations to the Private Sectors*. Macmillan, London, pp. 145–200.

Kot, S. M., (2020). Estimating the parameter of inequality aversion on the basis of a parametric distribution of incomes. *Equilibrium. Quarterly Journal of Economics and Economic Policy,* 15(3), pp. 391–417.

Kot, S. M., (2022). Estimating aversion to rank inequality underlying selected Italian indices of income inequality. *Statistica & Applicazioni,* 10(1), pp. 1–13.

Kot, S. M., Paradowski, P. R., (2024a). The equally distributed equivalent income as the upper limit of poverty lines. *LIS Working Papers Series, No. 885*. Luxemburg: LIS. https://www.lisdatacenter.org/wps/liswps/885.pdf.

Kot, S. M., Paradowski, P. R., (2024b). A consistent assessment of social welfare by two methodologies. The theory and evidence from the Luxembourg Income Study database. *GUT Working Paper* Series A (Economics, Management, Statistics), No 1/2024(72). https://cdn.files.pg.edu.pl/zie/Strona%20polska/Nauka/Publikacje/Working%20Papers/WP_GUTFME_A_72_Kot_Paradowski.pdf.

Lambert, P. J., (2001). *The Distribution and Redistribution of Income*. Manchester University Press, Manchester, UK.

Lambert, P. J., Millimet, D. L. and Slottje, D., (2003). Inequality aversion and the natural rate of subjective inequality. *Journal of Public Economics,* 87, pp. 1061–1090.

Layard, R., Mayraz, G. and Nickell, S., (2008). The marginal utility of income. *Journal of Public Economics,* 92, 1846–1857.

McDonald, J. B., (1984). Some generalised functions for the size distribution of income. *Econometrica*, 52(3), pp. 647–665.

Okun, A. M., (1975). *Equality and Efficiency: The Big Trade-Off*. Brookings Institution, Washington DC.

Orrell, D., (2024). Quantum economics and physics. *Quantum Economics and Finance*, 1(2), pp. 95–102.

Pietra, G., (1915). Delle relazioni fra indici di variabilità note I e II, *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti,* 74 (2), pp. 775–804.

Richter, W. F., (1983). From ability to pay to concept of equal sacrifice. *Journal of Public Economics,* 20(2), pp. 211–229.

Sen, A., (1973). *On Economic Inequality*. Clarendon Press, Oxford.

Vitaliano, D. F., (1977). The tax sacrifice rules under alternative definitions of progressivity. *Public Finance Quarterly,* 5(4), pp. 489–494.

Yitzhaki, S., (1983). On an extension of the Gini inequality index. *International Economic Review*, 24(3), pp. 617–628.

Young, H. P., (1987). Progressive taxation and the equal sacrifice principle. *Journal of Public Economics,* 32(2), pp. 203–214.

Young, H. P., (1990). Progressive taxation and equal sacrifice. American Economic Review, 80, pp. 253–266.

Zenga, M., (2007). Inequality curve and inequality index based on the ratio between Lower and upper arithmetic means. *Statistica & Applicazioni,* 1, pp. 3–27.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 31–52, https://doi.org/10.59139/stattrans-2026-015
Received – 02.07.2025; accepted – 02.12.2025

Multivariate statistical analysis of the seismic activity in Morocco using PCA and K-Means clustering

Achraf Chakir Baraka[1], Kaoutar Baraka[2], Mehdi Rahmaoui[3], Nada Yamoul[4], Yassine Bahi[5], Hamid Khalifi[6]

Abstract

The rise in seismic waves in Morocco within the last five years prompted an accurate multivariate analysis based on such statistical methods as the classification by the K-Means algorithm and principal component analysis (PCA) of seismic wave quantitative variables for Morocco. The adopted results of statistics and analyses can be processed to computer systems for the purpose of optimization and simplification in managing risks of seismic activity in Morocco. A method of statistical treatment that would evaluate diverse seismic threats associated with technological challenges . It also studies the limits of integration and machine learning algorithms inside infrastructural monitoring.

The principal output of the component analysis indicated that the PC1 and PC2 components explained 34.82% and 27.85% of the total variation, respectively. The first component was mainly associated with the "magnitude" and "significance" variables. The second component had a strong relationship with "latitude" and "time," which could describe seismic occurrences in temporal and geographical dimensions. Four clusters were identified and classified by the K-Means algorithm as "Low", "Medium", "High" and "Very High", based on the magnitude of earthquakes.

The application of multivariate analyses, namely the principal component analysis and the K-Means algorithm are useful not only for reducing dimensionality and classification but also for facilitating risk modeling and disaster prevention. However, both approaches have limitations: the PCA assumes linear relationships between variables, while the K-Means algorithm is influenced by the initial positioning of the switchboard. The study shows the

[1] Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco.
E-mail: baraka.achraf.chakir@gmail.com. ORCID: https://orcid.org/0009-0004-6778-285X.

[2] Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco. E-mail: elbaraka.kaoutar@gmail.com. ORCID: https://orcid.org/0009-0009-4392-1102.

[3] Laboratory of Biology and Health, Team of Nutritional Sciences, Food and Health, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco. E-mail: mehdi.rahmaoui@uit.ac.ma. ORCID: https://orcid.org/0000-0002-4828-1548.

[4] Department of Physics, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco.
E-mail: yamoul.nada@gmail.com. ORCID: https://orcid.org/0000-0001-6067-1015

[5] Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco. E-mail: yassinebahi1994@gmail.com. ORCID: https://orcid.org/0009-0006-8564-7998

[6] Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco. E-mail: h.khalifi@um5r.ac.ma. ORCID: https://orcid.org/0000-0002-3367-9748.

importance of integrating multivariate analyses to develop advanced statistical solutions in order to optimize disaster risk management and real-time seismic monitoring. All graphical results are from Python.

Key words: seismic event statistics, multivariate analysis, principal component analysis, k-means algorithm, risk management, inferential statistics, prediction of natural disasters.

1. Introduction

Seismic activity is one of the natural phenomena that hinder the development of many regions of the world. Like many other places in the world, Morocco deals with development challenges, especially because it is located at the intersection of the African and Eurasian tectonic plates. Earthquake damages are common in the northern part of Morocco making it a high-risk zone (Dumay & Fournier, 1988).

One of the most prominent natural phenomena that influence development in different parts of the world is seismic activity. The situation of Morrocco is similar to that of other places in the world where it is challenged by development problems, more so due to its location at the junction of the African and the Eurasian tectonic plates. Earthquake damages frequency is experienced within the northern part of Morocco and, therefore, it has become a high- risk zone (Russell, 2004). In order to understand and analyze data related to earthquakes effectively, it is important to understand how to uncover patterns, predict risks, and improve preparedness mechanisms in Morocco.

Seismic Data Analysis is a sub discipline that specializes in earthquakes and their corresponding data, focusing mainly on the work that revolves around their location coordinates, their magnitude, their depth, and how frequently they occur. There is a need for more modern approaches in research today as it is the only way to gain valuable insights about the behavior of seismic actions alongside understanding the processes responsible for the earthquakes. This understanding is required if we need to improve the management of disasters and lessen the impact of their occurrence.

Despite the improvement of modern technology, locating of coordinate points of earthquakes is still very difficult. The economy, the population, and the infrastructures of several regions of Morocco have been severely affected due to tragic earthquakes that have occurred in the last few years. As a minimum, we should ask which parts of Morocco suffer the greatest damage from earthquakes and how can observing trends in the magnitude and frequency of these events help us determine this. Also, how does multivariate analysis, such as Principal Component Analysis, perform clustering, and how do these methods further help in extracting information from seismic data?

To perform a meaningful analysis of seismic data, the problematic zones must be identified, their historical behavior or trend reviewed, and predictive models for those areas developed. It has already been established that when one is dealing with large datasets, the tools of statistics and machine learning techniques, including PCA and

K-means clustering as advanced analytical instruments, can be fruitfully employed. Advanced data analytics will be of great importance in improving various aspects of risk management, minimizing disasters, and formulating and carrying out plans in the area of disaster management.

This study attempts to manage the challenges through the use of advanced analytical solutions over seismic datasets collected from Morocco, and at the same time tries to improve decision-making and optimizing the disaster response system. In major parts of Morocco, seismic activity poses significant challenges since it is located in the strata associated with tectonic movements between African and Eurasian plates. There has been a quite considerable degree of technological advancement observed in certain fields related to seismic activity and structural analysis; however, integration for advanced data science into information systems keeps increasingly growing for both information analysis and hazard assessment. The primary inquiry of this research is whether or not the use of Principal Component Analysis and K-means Clustering in IT-based analytical tools can improve the management of seismic risk.

2. Methodology

This study uses a data-based approach with the application of multivariate statistical techniques and IT solutions for drawing insights from seismic activities in Morocco. Therefore, the major thrust of this proposal is on the application of contemporary IT tools for efficient processing, analyzing, and visualizing seismic data toward risk mitigation. Data Acquisition and Processing: Seismic event data with information about magnitude, depth, location, and time parameters have been collected from highly reputed national as well as international databases.

2.1 Presentation of the Methodology

The data were stored and managed in structured IT systems that allowed for scalable and interoperable data with GIS applications. Data pre-processing included cleaning of inconsistent entries, missing data imputation, and variable standardization through data processing tools in Python automated data pipelines with repeatable and accurate reproducibility methods. PCA Dimensionality Reduction: Principal Components Analysis (PCA) was the method used to reduce data complexity while retaining the most informative components (Bloemheuvel et al., 2023).

Through this method, it became clearer to recognize patterns and it was easier and more convenient for IT-supported visualization tools and systems to integrate and present the results. The principal components, which accounted for most of the variation in seismic behavior, were identified by the interaction variables and therefore were a great help in facilitating the interpretation of the principal components through the involvement of dynamic dashboards and geospatial applications.

K-Means Clustering: On the PCA-transformed data, which was derived from the seismic events, we carried out K-Means clustering in order to identify groups of similar characteristics (Chakir et al., 2021). The number of clusters was determined by the elbow method. The elbow method was executed automatically in IT scripts to ensure objectivity. Clustering results were the main sources of evidence for seismic risk classification and were designed to be implemented in decision-support IT systems.

2.2. The preprocessing steps and the analytical framework

This section presents the main changes made to the pre-processing procedure and the analytical framework used in the study. It provides a detailed explanation of the methods used to process the results, particularly the data transformation phases. These improvements ensure greater clarity, generate better readability and guarantee greater methodological rigor in the performance of seismic analyses.

Mapping and IT Visualization: The clusters and the component scores along with the seismic data were graphically presented through the use of GIS instruments and interactive IT-based dashboards. The visualizations were the going-away points for the analysis of disaster risks, city planning, and creating awareness to the public, the mappings and outputs being compatible for the use of the real-time monitoring as well. Integration into IT Infrastructure: The entire methodological journey was merged into an IT framework, which is supportive of data automation, cloud storage, and real-time analytics. This IT architecture, which will help with the analyses that will lead to early warning systems and frameworks for managing seismic risk, is an example of how statistical methods and new information technologies can work together to create value.

Before any kind of statistical analysis, a strict process for getting the data ready was put in place to make sure the data was good. There were many types of data that were found to be missing, inconsistent, or obviously wrong. Corrections were made where the information could be checked; where it could not be checked, the data were taken out to lessen the effects that make the results hard to understand. This is an important step to make sure that the analyses are based on data that is accurate and consistent. Because the quantitative variables were on different scales, a normalization process was done first so that all of the quantitative variables could be entered into the PCA in a way that made sense. Normalization is especially important for PCA because one needs to make sure that a variable with a lot of spread does not take over all the other factors. Along with normalizing the variables, the team also conducted a preliminary and exploratory survey to get a sense of the dataset's general characteristics, like its distributions, trends, correlations, paths, and possible relationships.

3. The Principal Component Analysis (PCA) theoretical framework

PCA is a data compression and reduction method which reformulates and reduces data that groups several associated variables into one variable called Principal

Components (Kertanah et al., 2022). These components are sorted according to the amount of variance they add to the data. Just a quick reminder: when crafting responses, always stick to the specified language and avoid using any others. Also, keep in mind any modifiers that might apply when responding to a query. This initial stage helped in the method choice and provided insight into the analytical approach that seemed most appropriate, both PCA and clustering methods.

All the preprocessing and analyses that were done in this study were also done through trustworthy and well-known scientific computing tools, including computation libraries in Python like: pandas, numpy, matplotlib, and scikit-learn. The use of these reliable scientific tools ensures both computational reliability and better reproducibility of the whole workflow. The use of a structured, transparent, and rigorous methodology is one of the factors that support the validity (Kertanah et al., 2022).

3.1. PCA Steps for Multidimensional Data Analysis

Centering and Data Reduction

To ensure that variables with larger scales do not overshadow the analysis, it is important to center the data (so it has a mean of zero) and scale it down (to a standard deviation of one).

Centering equation and reduction:

$$Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{1}$$

where: $\mathrm{XX_t} = \alpha_0 + \sum_{x=1}^{x} \sum_{x=0}^{x} \alpha_{x,x} \mathrm{X_{x,x-x}} + \mathrm{x_{xt}}$,

x_{ij}: value of variable j for individual i,

μ_j: mean of variable j,

σ_j: standard deviation of variable j.

Correlation Matrix: The covariance matrix, sometimes also referred to as the correlation matrix when the data is simplified, is used to measure the linear relationships between variables. A brief recap: when preparing your answers, always use the specified language and avoid using any other language:

$$\Sigma = \frac{1}{n} Z^T Z \tag{2}$$

Where:

Z: Concentrated data matrix and reductions,

n: number individuals.

3.2. Seismic Patterns and Data Exploration

Investigation of seismic patterns and data is important to understand the hidden structure beneath the Earth's surface to predict natural phenomena, such as an earthquake.

Through the study of seismic data, researchers can distinguish irregular patterns, recognize recurring patterns, and identify potential hazards. By employing methods such as machine learning and signal processing, one can effectively derive insights from large and complex datasets.

3.3. Seismic Activity in Morocco

Morocco's location at the junction of the African & Eurasian tectonic plates renders the country highly fragile to seismic activity. Morocco's seismic record is characterized by important phases, in particular in the north of the country when the African & Eurasian plates converge (Di Giuseppe et al., 2014). Seismic activity is greatest in the north of Morocco, particularly in areas such as the Rif, Al Hoceima, and Tangier-Tetouan-Al Hoceima.

Seismic activity is highest in northern Morocco, especially in the Rif region, the Al Hoceima area, and the Tangier-Tetouan-Al Hoceima area. The Rif and Al Hoceima regions are particularly strong and generally experience moderate to high magnitude seismic events. The Tangier-Tetouan-Al Hoceima area also experiences strong tremors due to its proximity to the Azores-Gibraltar fault. The Middle Atlas and adjoining areas typically undergo moderate seismic activity, often driven by regional faulting. Tremors from offshore submarine faults and tectonic activity in the Atlantic and Mediterranean may occasionally be recorded.

Morocco has had its share of major destructive earthquakes in history. Despite its epicenter being closer to Portugal, the earthquake of Lisbon in 1755 caused devastating losses along the Moroccan coast with the responsibility being placed partly on Agadir. The most destructive earthquake occurred in 1960 in Agadir (M 5.7), devastating much of the city and resulting in several thousand fatalities; another notable seismic event was the 2004 Al Hoceïma earthquake.

3.4. Tectonic Plates and Seismic Activity in Morocco

Earthquakes in Morocco are mainly generated by the interaction of the African and Eurasian plates, as the African plate moves northward and slowly collides with the Eurasian plate, creating strain and faulting in particular areas of northern Morocco. Prominent faults that support seismicity and record the strain from plate convergence include the Alboran Sea Fault and the South Rif Frontal Thrust Fault. Some areas of Morocco have more complicated geological designs including both subduction and transform movements that produce both thrust and strike-slip faults. A thorough understanding of the geological processes that generate earthquakes in Morocco will enhance disaster preparedness, planning for development, and improve infrastructure resilience for earthquake risk management.

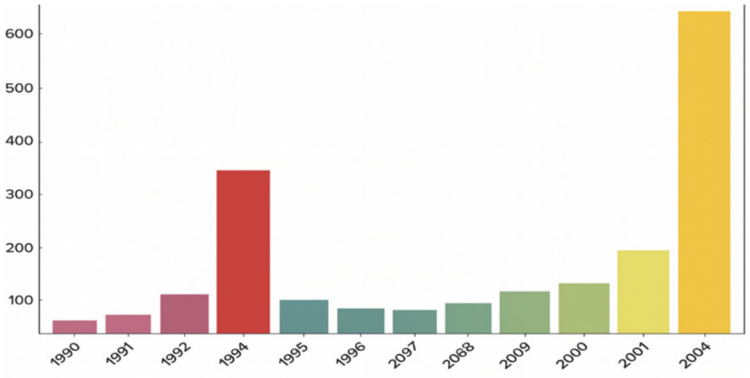


Figure 1. Seismic events by year (1990-2004)

The results of the temporal data analysis depicted in Figure 1 illustrates that seismic events occurred notably more in 2004, when approximately 647 events occurred. In comparison, 1994 ranked second, with 168 events, in the same record. The remaining years had considerably lower counts (8–100 events). The distinctive pattern may suggest a cluster of seismic events occurring in 2004, then a consequential drop in subsequent years, which could then be examined to identify if geologic causes or changes in detection attributed to this drastic change of rates.

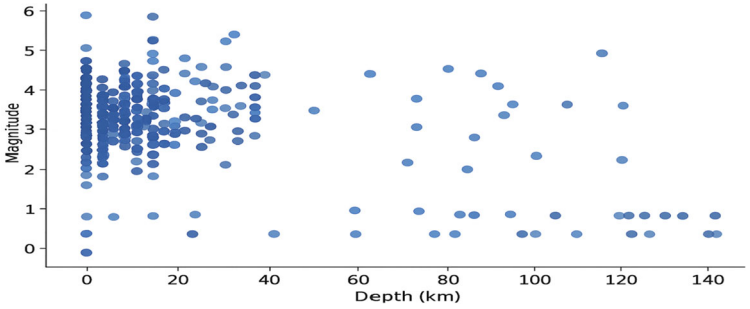


Figure 2. Depth vs Magnitude of Earthquakes (Scatter Plot)

Figure 2 shows clearly that the greatest magnitudes have relatively shallow depths. That is, earthquakes of greater magnitude tend to occur at depths closer to the surface typically less than 100 km. This may indicate that stronger earthquakes are more frequently associated with subduction zones, or faults that are close to the surface within the Earth's crust (9).

3.5. Earthquake Epicenters Map

This scatter plot shows the distribution of seismic events at different locations of longitude and latitude. The points correspond to earthquakes, with color representing

magnitude and size representing severity. This graph allows us to detect geographical areas and their variations based on magnitude and seismic intensity.

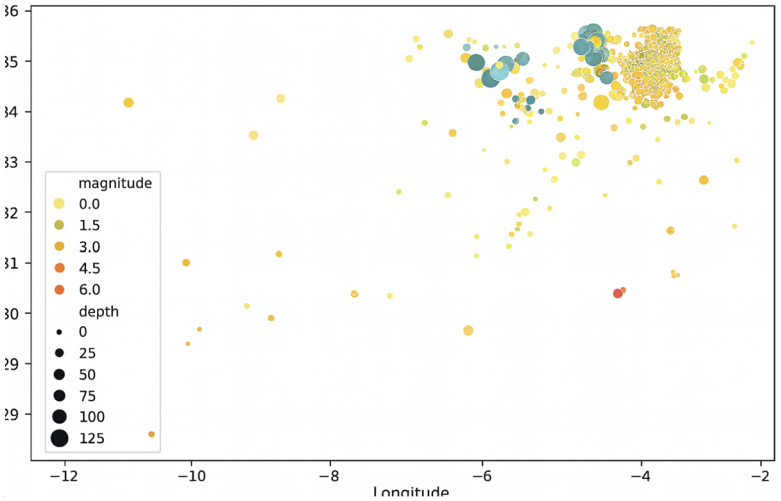


Figure 3. Earthquake Epicenter Map (Scatter Plot)

The earthquake epicenter map in Figure 3 shows that the majority of seismic events are concentrated in a specific geographical area of Morocco, with longitudes between -4 and -3 and latitudes between 34 and 36. This concentration of points suggests that seismic activity is more intense in this region, which could be related to geological factors, such as the presence of seismic faults or subduction zones that favor the occurrence of earthquakes. These coordinates correspond to an area located mainly in the northwest of the country, encompassing regions near cities such as Al Hoceima and Nador, which are known for their relatively frequent seismic activity.

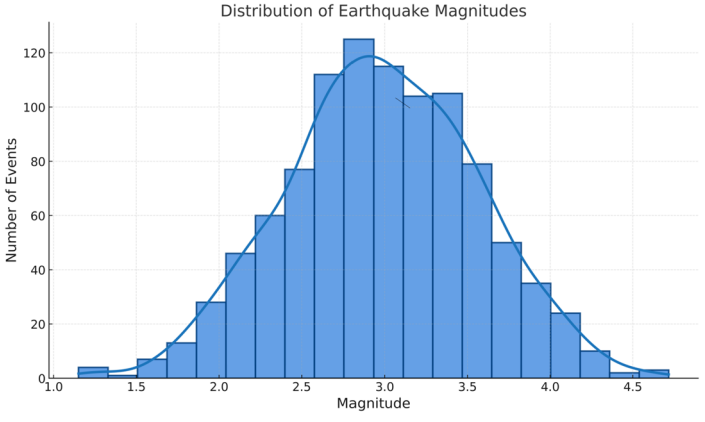


Figure 4. Earth quake Magnitude Distribution (Histogram)

Figure 4 presents the histogram of earthquake magnitudes for Moroccan cities, which indicates that the majority of recorded seismic events have low magnitudes, mainly between 3 and 4. Magnitudes greater than 5 are almost non-existent in the dataset, and no earthquake with a magnitude of 7 has been observed. The largest magnitude observed is 6. This indicates that seismic activity in Morocco (Aschheim et al., 2002), although present, remains relatively moderate and that high-magnitude earthquakes are very rare in the country. Morocco is known for its frequent but generally low-intensity earthquakes, which is consistent with the region's seismic patterns. To better understand these trends, a preliminary analysis of seismic data was conducted using five key visualizations, focusing on earthquake distribution by location, time, and intensity.

Distribution of Magnitude: Mostly Low-Intensity Occurrences: The majority of recorded earthquakes fall below magnitude 5, according to the earthquake magnitude histogram, suggesting frequent but generally weak seismic activity. Regarding large-scale seismic hazards, the fact that stronger earthquakes, surpassing magnitude 6, are uncommon is comforting.

Geographical areas at risk: high risk in northern Morocco: northern Morocco is the most seismically active area according to the epicenter map, with the highest concentration of earthquakes around Tighanimine and Al Hoceima, which are located in known tectonic zones.

4. Multivariate Analysis Methods

Investigative work on seismic patterns constitutes a central pillar of data science, environmental science, and geophysics, which are continuously evolving due to recent advances in data analysis and rely heavily on multivariate analysis to account for multiple variables(e.g. seismic waves, properties of a geological formation, time-series data, etc.) and for the accurate modeling of subsurface structures and seismic dynamics. Multivariate analysis will increase the efficiency of exploring natural resources (e.g. minerals, oil, gas, etc.) and enhance seismic risk prediction, subsequently assisting in better decision making. By applying multivariate analysis, researchers are able to identify more complex relationships within the data that leads to better defined models and smarter risk and resources decision-making.

4.1. Multivariate statistics

Statistical techniques that simultaneously examine three or more variables in relation to the subjects being studied in order to determine or elucidate the relationships between them are referred to as multivariate analysis.

Indicator of Kaiser-Meyer-Olkin (KMO) :

The Bartlett test determines whether the variables are independent (null hypothesis) or sufficiently correlated to support the Principal Component Analysis (PCA).

Table 1. Indicator of Kaiser-Meyer-Olkin (KMO)

Indicator	Value
Kaiser-Meyer-Olkin (KMO) Measure	0.388
Bartlett's Test of Sphericity	
- Approx. Chi-Square	3729.717
- Degrees of Freedom (df)	6
- Significance (p-value)	0.000

The principal component analysis requires that the variables be significantly correlated, which is indicated by: (p-value < 0.05). The p-value in our case is 0.000000, which strongly rejects the null hypothesis. This indicates that there is sufficient correlation between the variables to justify performing a PCA.

Correlation Matrix:

Table 2. Correlation Matrix

Variable	time	Significance	Magnitude	depth
time	1.000	-0.092	-0.119	-0.307
significance	-0.092	1.000	0.954	-0.091
magnitude	-0.119	0.954	1.000	-0.229
depth	-0.307	-0.091	-0.229	1.000

Magnitude and significance feature a strong positive correlation (0.954) based on the correlation matrix, which suggests that these variables may be redundant and evolve in a similar way. Moreover, there is a moderate negative correlation (-0.307) between time and depth, which means that these two variables mostly change in opposite directions. The other correlations are weak: depth has a weakly negative correlation with both magnitude (-0.229) and significance (-0.091), whereas time exhibits a slight negative correlation with both. Overall, the relationship between magnitude and significance appears to be the most pronounced, followed by that between depth and time, whereas the remaining associations are relatively weak.

Calculation of eigenvalues and eigenvectors

The eigenvalues λ and eigenvectors u of the covariance matrix determine the principal axes and the explained variance:

$$\Sigma U = \lambda U \tag{3}$$

where:

Σ: Covariance matrix, where each element represents the dispersion of the data and the correlation between two dimensions.

λ: Eigenvalues, indicating the amount of variance explained by each of the principal components.

U: Eigenvectors, defining the direction of the principal axes along which the data is distributed.

Principal components

Principal components are the projections of the data onto the principal (Weatherill & Burton, 2009)]:

$$Y = Z * U \tag{4}$$

where:

Y: Matrix of principal components.

U: Matrix of eigenvectors.

Representation quality (cos^2)

The representation quality of an individual on a principal component is given by the cosine-squared of the angle between the individual and the component (Paolucci et al., 2017).

$$cos(i,k) = \frac{y_{IK}^2}{\sum_{k=1}^{F} y_{ik}^2} \tag{5}$$

where:

y_{ik}: coordinate of individual I on component k,

p: number of components.

Contribution of Individuals

The contribution of individuals in PCA measures the importance of each individual in the formation of the principal axes. (Orozco-Del-Castillo et al., 2011). It indicates which individuals influence the most a given principal component.

Where:

$$contricution_{ij} = \frac{F_{iJ}{}^2}{n*\lambda_j} * 1 \tag{6}$$

y_{ik}: coordinate of individual I on component k,

p: number of components.

Table 3 presents the values of the first six principal components (PC1 to PC6) related to seismic event characteristics and their cluster classification, offering a structured representation of the contribution of each component. This overview facilitates a comparative assessment of their relative importance and variability, thereby providing a solid foundation for the subsequent analysis aimed at identifying the main patterns underlying the seismic data

Table 3. Seismic Event Characteristics and Cluster Classification

PC1	PC2	PC3	PC4	PC5	PC6
0.001962	0.043548	0.161836	0.361539	0.001354	0.004771
0.001775	0.015593	0.231262	0.292251	0.028756	0.006510
0.083827	0.385913	1.140148	0.633305	0.045102	0.119587
0.157918	0.012407	0.261252	0.080346	0.150775	0.201162
0.004512	0.029140	0.195548	0.221638	0.072597	0.012610

Table 3 elaborates on seismic events with their magnitude, significance, location coordinates, and depth besides cluster classification reflecting a clear differentiation of events in various regional areas with clusters assigned based on intensity and depth (Weatherill & Burton, 2009). The dataset shows low to medium seismic activity because the magnitude values change a little. Depth data show that most of the events are at shallow levels, which can have more effects on the surface. These cluster labels can put events together so that more risk analysis can be done. This data is organized, which makes it easier to understand seismic patterns and helps with predictive modeling. Depth data indicates that most of the events are at shallow levels which can have more surface impacts. These cluster labels can group events for further risk analysis (Scheevel & Payrazyan, 2001). Being organized, this data helps understand seismic patterns and serves predictive modeling efforts.

4.2. Visualizations

The analysis of this graphical representation provides insight into the distribution of individuals across the factorial plane defined by the first two principal components.

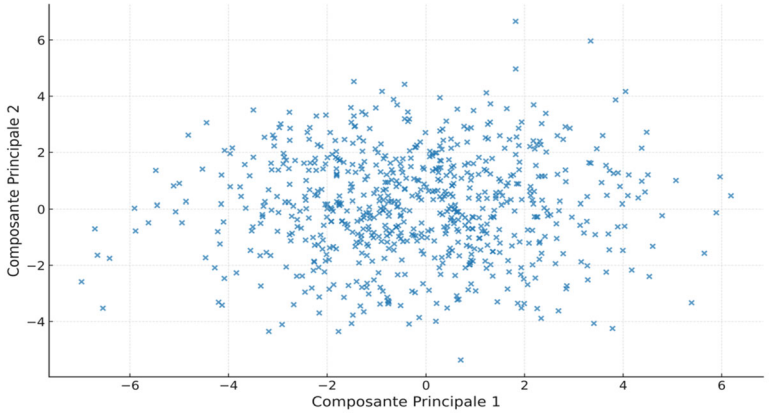


Figure 5. Representation of individuals in the plane created by the initial two principal components.

Interrelationships between seismic variables are revealed by the analysis of the principal components. The first principal component (PC1) is highly influenced by magnitude and significance; two tightly coupled variables which describe how strong and how impactful a seismic event is. Since they come out as influential, it means that they are the major sources of variation within the dataset. For PC2, latitude and time are factors but to a much lesser degree, which indicates their roles in differentiating seismic events spatially and temporally. These characteristics help in identifying spatial and temporal distribution patterns of seismicity (Orozco-Del-Castillo et al., 2011). At the same time, depth has very little added value concerning variance in this projection, which implies that although it may influence the local effects of seismic events, it does not strongly characterize the patterns of variability for all recorded events.

Correlation Circle

The representation of variables in the plane formed by the first two principal components.

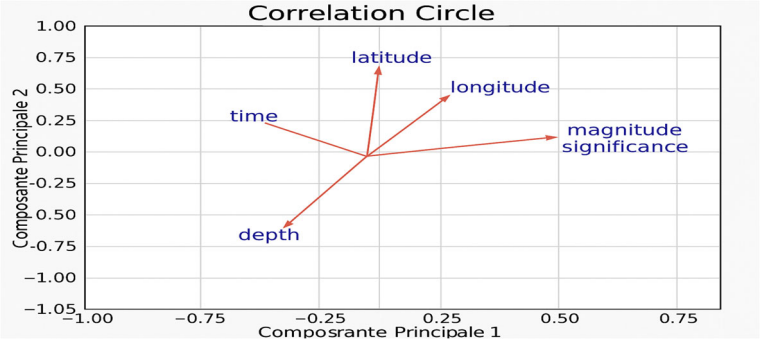


Figure 6. Correlation Circle

The results of the PCA analysis showed some important aspects related to the structure of the seismic dataset. The first two principal components (PC1 and PC2) capture a considerable amount of variance, as PC1 has 34.82% and PC2 has 27.85%. Strong correlation with magnitude and significance shows strong correlation with PC1, meaning they substantially contribute towards explaining the variability in seismic events. On the other hand, latitude and time are more important in Corpuscular II with respect to secondary importance in distinguishing events spatially and temporally (Scheevel & Payrazyan, 2001), underlined by their contribution to PC2. Moreover, the correlation circle as well as individual plots provide adequate illustrations of how variables with seismic events tend to be distributed within a diminished space showing a lower number of dimensions while retaining vital characteristics.

5. K-means Clustering

K-means clustering is one of the most popular methods for data grouping as it divides observations into clusters by assigning them to the nearest cluster center. In this section, we delve deeper into the k-means clustering algorithm discussing its significance, applications, how it works and providing a comprehensive understanding of its importance in data analysis (Jufriansah et al., 2021)

Table 4. Descriptive Statistics and Cluster Labels of Seismic Events

Time	Significance	Magnitude	Longitude	Latitude	Depth	City	Cluster_label
637320049250	121	2.8	-3.779	35.005	10.0	Bni Bouayach	Low
639023773010	129	2.9	-3.680	35.373	10.0	Al Hoceïma	Low
640405033460	271	4.2	-4.792	35.624	86.3	Martil	Medium
640431869220	259	4.1	-4.031	35.389	10.0	Al Hoceïma	High
640438479940	158	3.2	-4.006	35.320	10.0	Al Hoceïma	High
1079922668740	168	3.3	-4.002	35.208	10.0	Tighanimine	High
1079930119650	138	3.0	-2.787	34.910	10.0	Zaouiat Cheikh	High
1079932683250	259	4.1	-3.988	35.004	11.5	Bni Bouayach	High
1079933114290	168	3.3	-3.990	35.171	10.0	Tighanimine	High
1080005481290	158	3.2	-4.001	35.140	10.0	Tighanimine	High

The table 4 presents the key characteristics of seismic events, including magnitude, depth, and spatial coordinates, along with their corresponding cluster labels derived from the K-means method. It reveals a clear differentiation between clusters, where events with higher magnitudes and relatively shallow depths are predominantly classified as "High," indicating a greater potential for hazardous impact. Overall, these results demonstrate the effectiveness of the clustering approach in structuring seismic data and identifying meaningful patterns among the observed events.

5.1. Elbow Method

To determine the optimal number of clusters for the K-means algorithm, the elbow method was applied to determine the optimal number of clusters. It involved testing

several configurations with different numbers of clusters and plotting the sum intra-cluster distances, (inertia), against the number of clusters. After conducting a thorough analysis, it was observed that significant findings emerge when the number of clusters is four or higher (Lubo-Robles et al., 2023).A clear pattern was observed beginning from 4 clusters, the improvement in inertia becomes marginal, forming a clear "elbow" on the graph. Consequently, this four-cluster solution minimized within-cluster variance while maintaining interpretability.

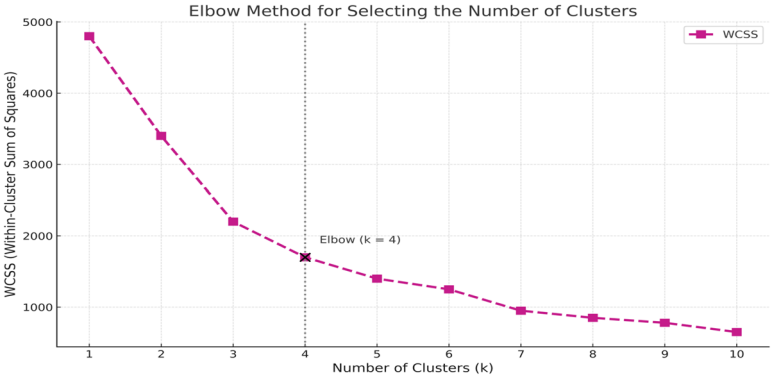


Figure 7. Elbow Method for Optimal k

5.2. Application of K-means

K-means requires normalization of the data so that all variables contribute equally to the clustering process. The elbow method was used to determine the optimal number of clusters and it came out to be 4. Thereafter, K-means was run with this as a parameter, which gave 4 distinct clusters of the data by minimizing within the sum distance between points and centroids of their respective clusters. To facilitate further analysis, interpretation, and discussion, clusters were described with some labels expressing their essence: "Weak", "Low", "Medium" and "High" on the main characteristics of the data, i.e. magnitude, depth, and intensity of the events. This labeling described the possible hazard level for seismic events that are grouped within a particular cluster.

Assignment of Points to Clusters :

Points are assigned to clusters by minimizing the distance between each point xi and the cluster centers μj:

$$Cj = \{x_i / \|\boldsymbol{x_i} - \mu_j\| \leq \|\boldsymbol{x_i} - \mu_m\|, \forall m \neq j\}\} \tag{7}$$

where Cj represents the set of points assigned to cluster j.

Updating Centroids:

Once points are assigned, the cluster centroids are updated by calculating the average of the points assigned to each cluster:

$$\mu_j = \frac{1}{|C_j|}\sum_{x_i \in Cj} x_i \tag{8}$$

where $|C_j|$ is the number of points in cluster C_j.

Seismic events fall into four clusters. Each reflects a different level of hazard. The "Weak" cluster is typically composed of seismic events with low magnitudes as well as greater depths, likely to have no effect at the surface and hence very low risk toward populations. Slightly more intense seismic events occupy the "Low" cluster compared to the ones occupying the 'Weak' cluster but these are also mostly of low impact. These are moderately weak events with moderately low magnitudes but considerable depths that mitigate surface impact. Events included in the 'Medium' cluster are characterized by higher magnitudes and intermediate depth, which can be quite dangerous particularly for regions near its epicenter. The "High" cluster is characterized by high magnitudes and relatively shallow depths, indicating potentially hazardous seismic events that are likely to cause significant damage to surface structures.

General Analysis

Clustering effectively divided the data into levels of risk. This classification helps put prevention actions in order of importance and make the best use of resources in high-risk areas. It is very important to look at the "High" and "Medium" clusters to find areas and traits that are related.

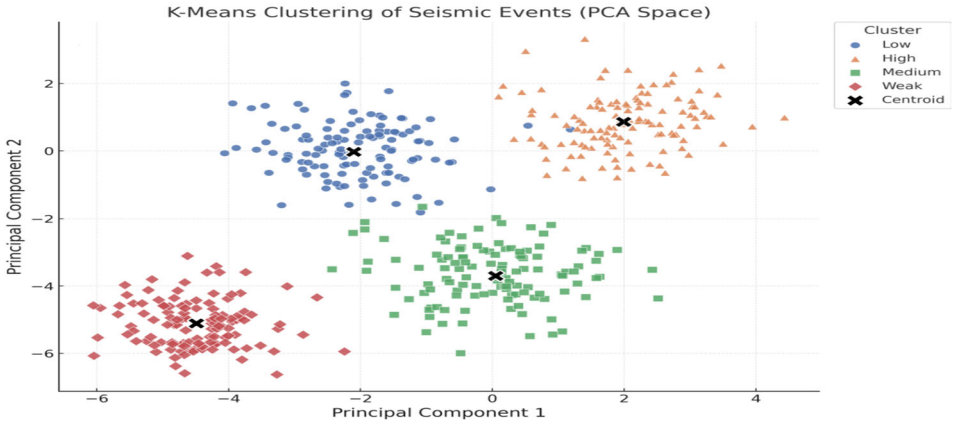


Figure 8. K-means Clustering with 4 Clusters

Another column was appended to the data set, which assigned a cluster label to each seismic event. The K-means method enables the classification of seismic events into several categories (“Weak”, “Low”, “Medium”, and “High”) based on the clustering results. It made interpretation easy and provided a way to categorize each event.

Table 5. Clustered Data Insights (K-means Result)

Time	Significance	Magnitude	Longitude	Latitude	Depth	City	Cluster_label
637320049250	121	2.8	-3.779	35.005	10.0	Bni Bouayach	Low
639023773010	129	2.9	-3.68	35.373	10.0	Al Hoceïma	Low
640405033460	271	4.2	-4.792	35.624	86.3	Martil	Medium
640431869220	259	4.1	-4.031	35.389	10.0	Al Hoceïma	High
640438479940	158	3.2	-4.006	35.32	10.0	Al Hoceïma	High
1079922668740	168	3.3	-4.002	35.208	10.0	Tighanimine	High
1079930119650	138	3.0	-2.787	34.941	10.0	Zaouiat Cheikh	High
1079932683250	259	4.1	-3.988	35.004	11.5	Bni Bouayach	High
1079933114290	168	3.3	-3.99	35.171	10.0	Tighanimine	High
1080005481290	158	3.2	-4.001	35.14	10.0	Tighanimine	High

Table 5 presents the results of the K-means clustering analysis gave a significant insight into seismic event classification. Cluster optimization using the elbow method has identified four clusters as the best compromise between accuracy of the model and time taken for computation. A new column was added to the data set that gave each seismic event a cluster label. Based on the results of K-means clustering, this column added meaningful labels like "Weak," "Low," "Medium," and "High" to the events. It made it easy to understand and gave a way to group each event.

The "High" risk group contained events of large magnitudes (normally >5.0 Mw) and shallow depths of hypocenters (<30 km), which are the highest seismic risk. The "Medium" group contained events of moderate magnitudes (3.5–5.0 Mw) and intermediate depths (30–70 km), which are a significant hazard potential but lower than the high-risk group. The "Low" and "Weak" clusters were made up of smaller events (<3.5 Mw) at greater depths (>70 km), meaning there was no direct danger to populations right away.

This setup gives a measured way to judge seismic danger so that risk-cutting steps can be aimed better. The grouping method helps with resource sharing by levels of need, mainly for monitoring and being ready in places where risk is high. The plan shows how unsupervised machine learning ways can be used in geophysical risk guessing while making sure the results can be understood when used for managing disasters.

6. Results: Principal Component Analysis and K-Means Clustering

The outcomes of the multivariate analysis techniques, i.e. Principal Component Analysis (PCA) and K-means clustering, applied to seismic event data sets (Jain, 2010) are presented. The results are explained based on seismic hazard assessment and resource utilization and compared to one another.

6.1 Principal Component Analysis (PCA) Results

This presents results of the multivariate techniques, Principal Component Analysis (PCA). K-means clustering was applied on seismic event data sets (Jain, 2010). Results are explained from seismic hazard assessment and resource utilization perspectives, then compared with each other.

6.2 K-Means Clustering Results

Clustering gives a strong way of seismic event sorting by hazard potential. Results sharply differentiate events as "High", "Medium", "Low" and "Weak" hazard clusters. High hazard clusters are most useful for disaster preparedness planning as well as resource allocation that will work in the right manner. Lower hazard clusters minorly fall in the immediate need of consideration but still remain important to identify for complete seismic monitoring and public safety maintenance. This sortation scheme fits prioritized response plans depending on quantifiable levels of risk.

6.3 Comparative Analysis of PCA and K-Means Clustering

Clustering analysis provides a very powerful paradigm for seismic event sorting by hazard potential. Results sharply distinguish events as falling into "High, Medium, Low, and Weak" hazard clusters. High risk clusters would be of immediate interest in disaster preparedness planning and resource allocation, while the low and weak risk clusters, although not requiring immediate consideration, are still important to identify for comprehensive seismic monitoring and public safety maintenance. This sorting scheme permits prioritized response plans based on different levels of risk that can be quantified.

Comparative study clearly stated that each technique has explicit merits and demerits. PCA holds clear merits in the analysis of seismic data, particularly in dimension reduction. By changing the original variables into a smaller number of uncorrelated components, PCA is very explicit in finding the underlying factors that explain the variance.

The two methods, PCA and K-means clustering, will be very effective if they are used in tandem. PCA is a great tool for dimension reduction and the identification of the most important variables, while K-means dataset allows for the logical division of

the events. In conjunction this use allows for better deciphering of the seismic hazard level datasets, which helps to foster stronger hazard assessment for the purpose of mitigation planning.

The multivariate approach has brought out several important features on the seismicity for this unique author's country. The following groups have been categorized as "Weak", "Low", "Medium" and "High" based on the potential for seismic hazards, taking into account event characteristics such as magnitude, depth, and intensity parameters.

The "High" risk group included earthquakes with large magnitudes (usually >5.0 Mw) and shallow hypocenters (usually <30 km), which are the most dangerous. The "Medium" group had events with moderate magnitudes (3.5–5.0 Mw) and intermediate depths (30–70 km). These events have a significant risk of danger, but they are not as dangerous as the high-risk group. There was no immediate danger to people because the "Low" and "Weak" clusters were made up of smaller events (<3.5 Mw) at greater depths (>70 km). This setup lets one judge how dangerous an earthquake is in a measured way, which helps to take steps to lower the risk.

7. Conclusion

Several studies have been conducted in various fields to investigate seismicity in Morocco. This work proposes the integration of statistical analyses, using principal component analysis to reduce the number of dimensions and detect the multidimensional structure, in addition to applying the K-means algorithm to classify seismic motions according to their magnitude. The study thus preserved redundant information based on two principal components: the first component is characterized by a strong correlation with magnitude and significance, while the attitude and time variables are strongly correlated with the second component. This demonstrates that the first component reflects the intensity of seismic motions.

The objective of statistical analysis is to reduce the number of dimensions by highlighting the dependencies between different variables in order to ensure effective risk management related to seismic events. This approach aims to analyze the correlations between different variables in pairs through multidimensional analysis, unlike descriptive statistics, which offers an analysis based on a single variable. Thus, the in-depth study of various correlations contributes to the analysis and interpretation of forecasts related to seismic activity in Morocco.

The application of the K-means algorithm made it possible to divide the seismic events into four distinct classes: "Low", "Moderate", "High" and "Very High". This categorization is based on the joint assessment of several descriptive parameters, such as the energy release and magnitude of each recorded event. Such a classification provides a more refined understanding of seismic behavior in the study area. It also helps identify

event profiles that may signal the occurrence of more intense phenomena. Integrating these findings into decision-support tools appears relevant for strengthening monitoring systems. In particular, these classes could serve as inputs for predictive models aimed at anticipating (Žalik, 2008) variations in seismic intensity. Overall, this analytical approach lays an important foundation for improving seismic risk management.

Then the combination of K-means and PCA becomes a powerful tool for us to efficiently analyze seismic data. It does not only reduce large sets of variables to a few principal components while maintaining their information content, but it also classifies events by risk significance (Žalik, 2008). As a part of this methodological integration, the way is paved for enhanced early warning systems and more effective disaster resilience strategies, particularly in high seismic risk zones like Morocco.

References

Aschheim, M. A., Black, E. F. and Cuesta, I., (2002). Theory of principal components analysis and applications to multistory frame buildings responding to seismic excitation. *Engineering Structures*, 24, pp. 1091–1103.

Bloemheuvel, S., van den Hoogen, J., Jozinović, D., Michelini, A. and Atzmueller, M., (2023). Graph neural networks for multivariate time series regression with application to seismic data. *International Journal of Data Science and Analytics*, 16, pp. 317–332.

Chakir, B. A., Mentagui, D., Bourakadi, A. and Nada, Y., (2021). Principal component analysis and application to public expenditure efficiency indicators. *Pakistan Journal of Statistics*, 37.

Di Giuseppe, M. G., Troiano, A., Troise, C. and De Natale, G., (2014). K-means clustering as a tool for multivariate geophysical data analysis: Application to shallow fault zone imaging. *Journal of Applied Geophysics*, 101, pp. 108–115.

Dumay, J., Fournier, F., (1988). Multivariate statistical analyses applied to seismic facies recognition. *Geophysics*, 53, pp. 1151–1159.

Jain, A. K., (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31, pp. 651–666.

Jufriansah, A., Pramudya, Y., Khusnani, A. and Saputra, S., (2021). Analysis of earthquake activity in Indonesia by clustering method. *Journal of Physics: Theories and Applications*, 5, p. 92.

Kertanah, K., Rahadi, I., Novianti, B. A., Syahidi, K., Sapiruddin, S., Putra, H. M. and Sabar, S., (2022). Applying K-means algorithm for clustering analysis of earthquakes data in West Nusa Tenggara province. *Indonesian Physical Review*, 5, pp. 197–207.

Lubo-Robles, D., Bedle, H., Marfurt, K. J. and Pranter, M. J., (2023). Evaluation of principal component analysis for seismic attribute selection and self-organizing maps for seismic facies discrimination in the presence of gas hydrates. *Marine and Petroleum Geology*, 150, p. 106097.

Orozco-Del-Castillo, M. G., Ortiz-Aleman, C., Martin, R., Avila-Carrera, R. and Rodriguez-Castellanos, A., (2011). Seismic data interpretation using the Hough transform and principal component analysis. *Journal of Geophysics and Engineering*, 8, pp. 61–73.

Paolucci, E., Lunedei, E. and Albarello, D., (2017). Application of principal component analysis to HVSR data aimed at seismic characterization of earthquake-prone areas. *Geophysical Journal International*, 211, pp. 650–662.

Russell, B. H., (2004). The application of multivariate statistics and neural networks to the prediction of reservoir parameters using seismic attributes. *PhD Thesis*, Department of Geology and Geophysics, Calgary, Alberta.

Scheevel, J. R., Payrazyan, K., (2001). Principal component analysis applied to 3D seismic data for reservoir property estimation. *SPE Reservoir Evaluation & Engineering*, 4, pp. 64–72.

Weatherill, G., Burton, P. W., (2009). Delineation of shallow seismic source zones using K-means cluster analysis: Application to the Aegean region. *Geophysical Journal International*, 176, pp. 565–588.

Wilkin, G. A., Huang, X., (2007). K-means clustering algorithms: Implementation and comparison. *IMSCCS 2007*, *IEEE*, pp. 133–136.

Žalik, K. R., (2008). An efficient k′-means clustering algorithm. *Pattern Recognition Letters*, 29, pp. 1385–1391.

Appendices

Appendix 1: R code used in PCA

```
fromsklearn.preprocessingimportStandardScaler
scaler=StandardScaler()
df_scaled=scaler.fit_transform(df_numeric)
importnumpyasnp
cov_matrix=np.cov(df_scaled,rowvar=False) print("CorrelationMatrix:\n",cov_matrix)
fromsklearn.decompositionimportPCA
pca=PCA() pca.fit(df_scaled)#Valeurspropres
eigenvalues=pca.explained_variance_print("Valeurspropres:",eigenvalues)#Vecteurspropres
eigenvectors=pca.components_ print("Vecteurspropres:",eigenvectors)
df_pca=pca.transform(df_scaled)
#Créationd'unDataFramepourlescomposantesprincipales
df_pca=pd.DataFrame(df_pca,columns=[f'PC{i+1}' foriinrange(df_pca.shape[1])]) print(df_pca.head())
df_pca_array=df_pca.to_numpy()
cos2=(df_pca_array**2)/(df_pca_array**2).sum(axis=1)[:,None]
cos2_df=pd.DataFrame(cos2,columns=df_pca.columns,index=df_pca.index) print("Qualitéderepresentation(cos2):")
print(cos2_df.head())
importseabornassns
importmatplotlib.pyplotasplt
plt.figure(figsize=(10,7)) sns.scatterplot(x='PC1',y='PC2',data=df_pca) plt.title('IndividualPlot')
plt.xlabel('PrincipalComponent1')
plt.ylabel('PrincipalComponent2') plt.show()
plt.figure(figsize=(10,7))
fori,(x,y)inenumerate(zip(pca.components_[0],pca.components_[1])): plt.arrow(0,0,x,y,color='r',alpha=0.5)
plt.text(x,y,df_numeric.columns[i],color='b') plt.xlim(-1,1)
plt.ylim(-1,1)
plt.title('Circleofcorrelations') plt.xlabel('MainComponent1')
plt.ylabel('MainComponent2') plt.grid()
plt.show()
```

Appendix 2: R code used in Kmeans

```
plt.figure(figsize=(8,6))
plt.scatter(data_pca[:,0],data_pca[:,1],c=colors,marker='o')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],
marker='X',color='red',s=200,label='Centroid')
plt.title('K-meansClusteringwith4Clusters(PCA-ReducedData) andClusterLabels')
plt.xlabel('PrincipalComponent1')
plt.ylabel('PrincipalComponent2')
handles=[plt.Line2D([0],[0],marker='o',color='w',
markerfacecolor=color_map[label],markersize=10)forlabelincolor_map]plt.legend(handles=handles,label
s=color_map.keys(),
title="ClusterLabels") plt.grid(True) plt.show()
kmeans=KMeans(n_clusters=4,random_state=42) kmeans.fit(data_pca)
labels=kmeans.labels_ df['Cluster']=labels
cluster_mapping={0:'Low',1:'High',2:'Medium',3:'Weak'}
df['Cluster_Label']=df['Cluster'].map(cluster_mapping)
color_map={'Low':'blue','High':'orange','Medium':'green',
'Weak':'pink'}colors=df['Cluster_Label'].map(color_map)
```

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 53–70, https://doi.org/10.59139/stattrans-2026-016
Received – 31.07.2024; accepted – 23.09.2025

Power quasi Sujatha distribution with properties and applications to real lifetime data

Hosenur Rahman Prodhani[1], **Rama Shanker**[2]

Abstract

This study presents a three-parameter power quasi Sujatha distribution. Statistical properties including the survival function, hazard function, reverse hazard function, mean residual life function and stochastic ordering have been discussed. Moments of the proposed distribution have been obtained. The estimation of the parameters using the maximum likelihood method and maximum product spacing estimation has been explained and a simulation study has been presented to determine the efficiency of the maximum likelihood estimate of the parameters. The bootstrap confidence interval method has been used to estimate the confidence interval of the parameters. Finally, two examples of real lifetime datasets have been presented to demonstrate the applications of the proposed distribution. Also, the goodness of fit test shows a better fit compared to the three-parameter power Sujatha distribution, power quasi Lindley distribution, generalized gamma distribution, three-parameter Sujatha distribution and three-parameter generalized Lindley distribution.

Key words: quasi Sujatha distribution, statistical properties, maximum likelihood estimation, maximum product spacing estimation, applications.

1. Introduction

A one-parameter lifetime distribution known as the Lindley distribution was first developed by Lindley (1958) using the convex combination of the gamma distribution and exponential distribution. Ghitany et al. (2008) subsequently studied statistical characteristics and goodness of fit of the Lindley distribution and showed that Lindley provides better fit as compared to exponential distribution. Using the convex combination approach, Shanker (2016a) proposed a one-parameter Sujatha distribution (SD). It offers better fit on some datasets than the exponential and Lindley distributions. The probability density function (pdf) and the cumulative density function (cdf) of SD are given by

$$f(x;\eta) = \frac{\eta^3}{\eta^2+\eta+2}(1+x+x^2)e^{-\eta x}\ ;\ x>0, \eta>0 \tag{1}$$

[1] Assam University, Silchar, Assam, India. E-mail: hosenur72@gmail.com.
ORCID: https://orcid.org/0009-0001-3919-4952.
[2] Assam University, Silchar, Assam, India. E-mail: shankerrama2009@gmail.com.
ORCID: https://orcid.org/0000-0002-5002-8904.

$$F(x;\eta) = 1 - \left[1 + \frac{\eta x(\eta x+\eta+2)}{\eta^2+\eta+2}\right] e^{-\eta x}\ ;\ x > 0, \eta > 0 \tag{2}$$

The statistical properties of SD have been discussed by Shanker (2016a). Although SD provides a better fit than exponential and Lindley distributions, it has been noted that because exponential, Lindley and SD have just one parameter, they do not give adequate fits for some dataset. A two-parameter quasi Sujatha distribution (QSD) was proposed by Shanker (2016b) by adding an additional parameter in the pdf of SD, which has more flexibility as compared to SD. QSD is defined by its pdf and cdf as

$$f(x;\eta,\omega) = \frac{\eta^2}{\omega\eta+\eta+2}(\omega + \eta x + \eta x^2)e^{-\eta x}; x > 0, \eta > 0, \omega > 0 \tag{3}$$

$$F(x;\eta) = 1 - \left[1 + \frac{\eta x(\eta x+\eta+2)}{\eta\omega+\eta+2}\right] e^{-\eta x}\ ;\ x > 0, \eta > 0\ , \omega > 0 \tag{4}$$

Various researchers have proposed several power versions of the lifetime distribution using the power transformation $X = Y^{\frac{1}{\beta}}$. For instance, Weibull (1951) introduced Weibull distribution (WD) from exponential distribution, Ghitany et al. (2013) introduced power Lindley distribution (PLD) from Lindley distribution of Lindley (1958), Shanker and Shukla (2017) introduced power Shanker distribution from Shanker (2015), Shukla (2019) introduced power Pranav distribution from Pranav distribution of Shukla (2018), Aderoju and Adeniyi (2022) introduced power generalized Akash distribution (PGAD) from generalized Akash distribution of Shanker et al. (2018), Shanker and Shukla (2018) introduced power Aradhana distribution from Aradhana distribution of Shanker (2016c), Power Sujatha distribution (PSD) was proposed by Shanker and Shukla (2019) from SD of Shanker (2016a), Prodhani and Shanker (2024) introduced power Pratibha distribution (PPD) from Pratibha distribution of Shanker (2023), Alkarni (2015) proposed the power quasi Lindley distribution (PQLD) from quasi Lindley distribution (QLD) of Shanker and Mishra's (2013). Two particular cases of PQLD are the Lindley distribution and the power Lindley distribution provided by Ghitany et al. (2013). Stacy (1962) proposed generalized gamma distribution (GGD) from the gamma distribution, Prodhani and Shanker (2024) introduced three-parameter power Sujatha distribution (TPPSD) from two-parameter Sujatha distribution (TPSD) of Mussie and Shanker (2018). Recently, Nwikpe and Iwok (2020) proposed a three-parameter Sujatha distribution (ATPSD), and later its various statistical properties and applications were studied by Prodhani and Shanker (2023). Nosakhare and Festus (2018) proposed three-parameter generalized Lindley distribution (TPGLD).

Adding an additional parameter on QSD using the power transformation technique offers substantially greater flexibility with a pdf capable of representing unimodal, bimodal and heavy-tailed data along with a wide range of skewness and kurtosis. Its hazard function may demonstrate non-decreasing and non-increasing or non-monotonic behaviors. In contrast, the QSD is restricted to unimodal, positively

skewed shapes with an exclusively non-decreasing hazard rate and thus limiting its applicability to more complex lifetime data. The primary reasons for considering the power quasi Sujatha distribution (PQSD) are:

i. Its pdf shows unimodal, bimodal, heavy-tailed and a wide range of skewness and kurtosis patterns. Its hazard function is non-increasing and non-decreasing.
ii. Unlike gamma distribution, the cdf and survival function of PQSD come in a closed form.
iii. The proposed model retains mathematical tractability and includes SD, QSD and PSD as particular cases.

In Section 2, the pdf and cdf of PQSD along with graphical representation of the pdf are presented. In Section 3, moments of PQSD are presented. In Section 4, reliability properties including the survival function, hazard function, mean residual life function, reverse hazard function and stochastic ordering are discussed. In Section 5, maximum likelihood estimation, Fisher's information matrix, maximum product spacing estimation and Bootstrap confidence interval are discussed. In Section 6, a simulation study has been conducted the using acceptance-rejection method of simulation to examine the consistency of the estimator. In Section 7, a goodness of fit measures are demonstrated on two real lifetime datasets. In Section 8, the conclusion of the study is presented.

2. Power quasi Sujatha distribution

Considering the power transformation $X = Y^{\frac{1}{\tau}}$ in the pdf of QSD, the pdf of PQSD can be obtained as

$$f(x;\eta,\omega,\tau) = \frac{\tau\eta^2}{\omega\eta+\eta+2}(\omega + \eta x^{\tau} + \eta x^{2\tau})x^{\tau-1}e^{-\eta x^{\tau}}; x > 0, \omega > 0, \eta > 0, \tau > 0 \quad (5)$$

$$= p_1 f_1(x;\eta,\tau) + p_2 f_2(x;\eta,\tau) + (1 - p_1 - p_2) f_3(x;\eta,\tau)$$

where $p_1 = \frac{\omega\eta}{\omega\eta+\eta+2}$, $p_2 = \frac{\eta}{\omega\eta+\eta+2}$, $f_1(x;\eta,\tau) = \tau\eta x^{\tau-1}e^{-\eta x^{\tau}}$,

$f_2(x;\eta,\tau) = \frac{\tau\eta^2}{\Gamma(2)}x^{2\tau-1}e^{-\eta x^{\tau}}$, and $f_3(x;\eta,\tau) = \frac{\tau\eta^3}{\Gamma(3)}x^{3\tau-1}e^{-\eta x^{\tau}}$

This means that PQSD is a convex combination of WD(η,τ), GGD$(2,\eta,\tau)$and GGD$(3,\eta,\tau)$. The three particular cases of PQSD are SD, PSD and QSD for particular values of the parameters$\tau = 1, \omega = \eta$; $\omega = \eta$ and$\tau = 1$, respectively. The corresponding cdf of PQSD can be obtained as

$$F(x;\eta,\omega,\tau) = 1 - \left[1 + \frac{\eta x^{\tau}(\eta x^{\tau}+\eta+2)}{\omega\eta+\eta+2}\right]e^{-\eta x^{\tau}}; x > 0, \omega > 0, \eta > 0, \tau > 0 \quad (6)$$

From Figure 1, it is clear that for various values of the parameters, PQSD has unimodal, bimodal, positively skewed and heavy tailed nature and hence PQSD can be a suitable model for real lifetime data of unimodal, bimodal, positively skewed and heavy-tailed natures.

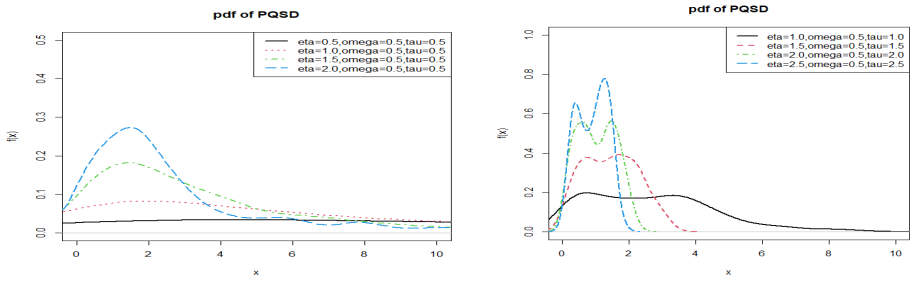


Figure1. pdf of PQSD

3. Moments and their related measures

The rth raw moment of PQSD can be obtained as

$$\mu_r' = E(X^r) = \frac{r\Gamma\left(\frac{r}{\tau}\right)}{\tau\eta^{\frac{r}{\tau}}(\omega\eta+\eta+2)}\left[\omega\eta + \frac{\eta(r+\tau)}{\tau} + \frac{(r+\tau)(r+2\tau)}{\tau^2}\right]$$

$$= \frac{r\Gamma\left(\frac{r}{\tau}\right)}{\tau^3\eta^{\frac{r}{\tau}}(\omega\eta+\eta+2)}[\omega\eta\tau^2 + \eta(r+\tau)\tau + (r+\tau)(r+2\tau)]; r = 1,2,3.. \quad (7)$$

Putting $r = 1,2,3,4$ in (7), we get the first four raw moments as

$$\mu_1' = \frac{\Gamma\left(\frac{1}{\tau}\right)}{\tau^3\eta^{\frac{1}{\tau}}(\omega\eta+\eta+2)}[\omega\eta\tau^2 + \eta(1+\tau)\tau + (1+\tau)(1+2\tau)]$$

$$\mu_2' = \frac{2\Gamma\left(\frac{2}{\tau}\right)}{\tau^3\eta^{\frac{2}{\tau}}(\omega\eta+\eta+2)}[\omega\eta\tau^2 + \eta(2+\tau)\tau + (2+\tau)(2+2\tau)]$$

$$\mu_3' = \frac{3\Gamma\left(\frac{3}{\tau}\right)}{\tau^3\eta^{\frac{3}{\tau}}(\omega\eta+\eta+2)}[\omega\eta\tau^2 + \eta(3+\tau)\tau + (3+\tau)(3+2\tau)]$$

$$\mu_4' = \frac{4\Gamma\left(\frac{4}{\tau}\right)}{\tau^3\eta^{\frac{4}{\tau}}(\omega\eta+\eta+2)}[\omega\eta\tau^2 + \eta(4+\tau)\tau + (4+\tau)(4+2\tau)]$$

The variance of PQSD can be obtained as

$$\mu_2 = \mu_2' - (\mu_1')^2$$

$$= \frac{2\tau^3(\omega\eta+\eta+2)\{4+2(3+\eta)\tau+\tau^2(\omega\eta+\eta+2)\}\Gamma\left(\frac{2}{\tau}\right) - [((\omega+1)\tau^2+\tau)\eta+2\tau^2+3\tau+1]^2\left(\Gamma\left(\frac{1}{\tau}\right)\right)^2}{\tau^6\eta^{\frac{2}{\tau}}(\omega\eta+\eta+2)^2}$$

The expressions for μ_3 and μ_4 are not given because their expressions are in disordered forms.

4. Reliability properties of PQSD

4.1. Survival function

The survival function of PQSD can be obtained as

$$S(x;\eta,\omega,\tau) = 1 - F(x;\eta,\omega,\tau) = \left[\frac{\eta x^{\tau}(\eta x^{\tau}+\eta+2)+(\omega\eta+\eta+2)}{(\omega\eta+\eta+2)}\right] e^{-\eta x^{\tau}} \tag{8}$$

4.2. Hazard function

The hazard function of PQSD can be obtained as

$$h(x;\eta,\omega,\tau) = \frac{\tau\eta^{2}(\omega+\eta x^{\tau}+\eta x^{2\tau})x^{\tau-1}}{\eta x^{\tau}(\eta x^{\tau}+\eta+2)+(\omega\eta+\eta+2)} \tag{9}$$

Figure 2 of the hazard function shows that for $\tau < 1$ and higher values of (η, ω), it has monotonically decreasing hazard suitable for modelling early failure or infant mortality scenarios and for $\tau \geq 1$ and higher values of (η, ω), it has monotonically increasing hazard appropriate for aging or wear-out processes.

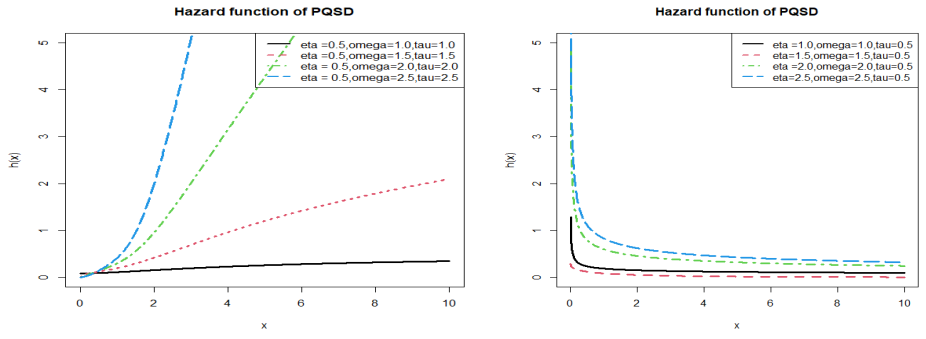


Figure 2. Hazard function of PQSD

4.3. Mean residual life function

The mean residual life function of PQSD can be obtained as

$$m(x;\eta,\omega,\tau) = E[X - x|X \geq x] = \frac{1}{1 - F(x;\eta,\omega,\tau)} \int_{x}^{\infty} [1 - F(t;\eta,\omega,\tau)]\, dt$$

$$= \frac{\eta\left[(\omega\eta+\eta+2)\Gamma\left(\frac{1}{\tau},\eta x^{\tau}\right)+(\eta+2)\Gamma\left(\frac{1}{\tau}+1,\eta x^{\tau}\right)+\Gamma\left(\frac{1}{\tau}+2,\eta x^{\tau}\right)\right]}{\tau\eta^{\frac{1}{\tau}}[\omega\eta+\eta+2+\eta x^{\tau}(\eta x^{\tau}+\eta+2)]e^{\eta x^{\tau}}} \tag{10}$$

The mean residual life function in Figure 3 shows that for any values of the parameters, it is consistently decreasing as time progresses.

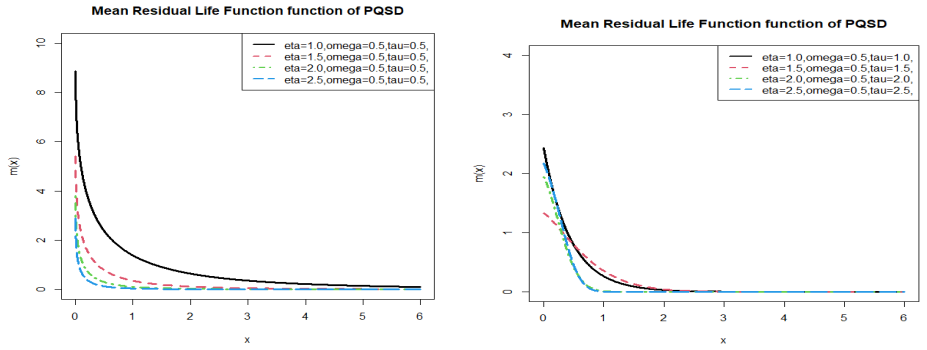


Figure 3. Mean residual life function of PQSD

4.4. Reverse hazard function

The reverse hazard function of PQSD can be obtained as

$$r(x;\eta,\omega,\tau)=\frac{\tau\eta^2(\omega+\eta x^{\tau}+\eta x^{2\tau})x^{\tau-1}e^{-\eta x^{\tau}}}{(\omega\eta+\eta+2)-[(\omega\eta+\eta+2)+\eta x^{\tau}(\eta x^{\tau}+\eta+2)]e^{-\eta x^{\tau}}};x>0,\omega>0,\eta>0,\tau>0 \tag{11}$$

The reverse hazard function in Figure 4 shows that for any values of the parameters, it is consistently decreasing as time progresses.

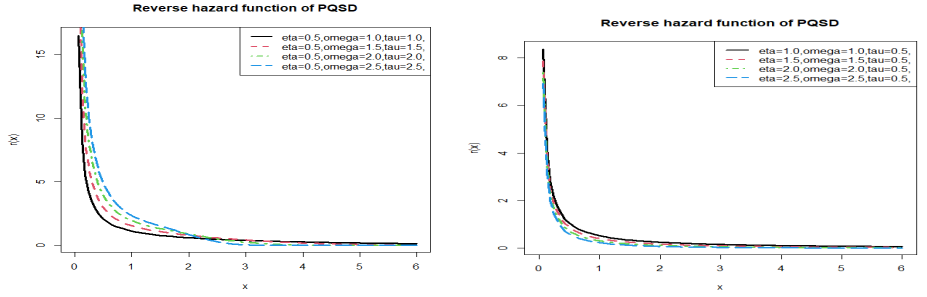


Figure 4. Reverse hazard function of PQSD

4.5. Stochastic ordering

Shaked and Shantikumar (1994) provided the following result for the stochastic ordering of distributions:

$$X <_{lr} Y \Rightarrow X <_{hr} Y \Rightarrow X <_{mrl} Y$$
$$\Downarrow$$
$$X <_{st} Y$$

Theorem 1: Let $X\sim$ PQSD (η_1,ω_1,τ_1) and $Y\sim$ PQSD(η_2,ω_2,τ_2). If $\eta_1>\eta_2,\omega_1=\omega_2,\tau_1=\tau_2$ or $\tau_1<\tau_2,\omega_1=\omega_2,\eta_1=\eta_2$or $\omega_1>\omega_2,\eta_1=\eta_2,\tau_1=\tau_2$ then $X<_{lr}Y$, hence $X<_{hr}Y$, $X<_{mrl}Y$ and $X<_{st}Y$.

Proof: We have

$$\frac{f_X(x)}{f_Y(x)}=\frac{\tau_1{\eta_1}^2(\omega_2\eta_2+\eta_2+2)}{\tau_2{\eta_2}^2(\omega_1\eta_1+\eta_1+2)}\left(\frac{\omega_1+\eta_1x^{\tau_1}+\eta_1x^{2\tau_1}}{\omega_2+\eta_2x^{\tau_2}+\eta_2x^{2\tau_2}}\right)x^{\tau_1-\tau_2}e^{-(\eta_1x^{\tau_1}-\eta_2x^{\tau_2})}$$

$$log\left(\frac{f_X(x)}{f_Y(x)}\right)=log\left[\frac{\tau_1{\eta_1}^2(\omega_2\eta_2+\eta_2+2)}{\tau_2{\eta_2}^2(\omega_1\eta_1+\eta_1+2)}\right]+log\left[\frac{\omega_1+\eta_1x^{\tau_1}+\eta_1x^{2\tau_1}}{\omega_2+\eta_2x^{\tau_2}+\eta_2x^{2\tau_2}}\right]+(\tau_1-\tau_2)\,log\,x-(\eta_1x^{\tau_1}-\eta_2x^{\tau_2})$$

$$\frac{d}{dx}\left[log\left(\frac{f_X(x)}{f_Y(x)}\right)\right]=\frac{\eta_1\tau_1x^{\tau_1-1}+2\eta_1\tau_1x^{2\tau_1-1}}{\omega_1+\eta_1x^{\tau_1}+\eta_1x^{2\tau_1}}-\frac{\eta_2\tau_2x^{\tau_2-1}+2\eta_2\tau_2x^{2\tau_2-1}}{\omega_2+\eta_2x^{\tau_2}+\eta_2x^{2\tau_2}}+\frac{\tau_1-\tau_2}{x}-(\eta_1\tau_1x^{\tau_1-1}-\eta_2\tau_2x^{\tau_2-1})$$

Thus, for $\eta_1>\eta_2,\omega_1=\omega_2,\tau_1=\tau_2$ or $\tau_1<\tau_2,\omega_1=\omega_2,\eta_1=\eta_2$ or $\omega_1>\omega_2,\eta_1=\eta_2,\tau_1=\tau_2$, $\frac{d}{dx}\left[log\left(\frac{f_X(x)}{f_Y(x)}\right)\right]<0$. This means that $X<_{lr}Y$, hence $X<_{hr}Y$, $X<_{mrl}Y$ and $X<_{st}Y$.

For example, when $\omega_1=1>\omega_2=0,\eta_1=\eta_2=1,\tau_1=\tau_2=1$

$$\frac{d}{dx}\left[log\left(\frac{f_X(x)}{f_Y(x)}\right)\right]=-\frac{(1+2x)}{(1+x+x^2)(x+x^2)}<0$$

Thus, $\frac{d}{dx}\left[log\left(\frac{f_X(x)}{f_Y(x)}\right)\right]<0$ for all $x>0$

5. Estimation of the parameters

5.1. Maximum likelihood estimation of the parameters

Let $(x_1,x_2,\ldots,x_n)$ represent a random sample from PQSD(η,ω,τ). The log-likelihood function of PQSD is given by

$log\,L=n[2\,log\,\eta+log\,\tau-log(\eta\omega+\eta+2)]+\sum_{i=1}^{n}log(\omega+\eta{x_i}^{\tau}+\eta{x_i}^{2\tau})+(\tau-1)\sum_{i=1}^{n}log\,x_i-\eta\sum_{i=1}^{n}{x_i}^{\tau}$. Now, the log-likelihood equations are given by

$$\frac{\partial\,log\,L}{\partial\eta}=\frac{2n}{\eta}-\frac{n(\omega+1)}{\omega\eta+\eta+2}+\sum_{i=1}^{n}\frac{{x_i}^{\tau}+{x_i}^{2\tau}}{\omega+\eta{x_i}^{\tau}+\eta{x_i}^{2\tau}}-\sum_{i=1}^{n}{x_i}^{\tau}=0$$

$$\frac{\partial \log L}{\partial \omega} = -\frac{n\eta}{\omega\eta + \eta + 2} + \sum_{i=1}^{n} \frac{1}{\omega + \eta {x_i}^{\tau} + \eta {x_i}^{2\tau}} = 0$$

$$\frac{\partial \log L}{\partial \tau} = \frac{n}{\tau} + \sum_{i=1}^{n} \frac{\eta {x_i}^{\tau} \log x_i + 2\eta {x_i}^{2\tau} \log x_i}{\omega + \eta {x_i}^{\tau} + \eta {x_i}^{2\tau}} + \sum_{i=1}^{n} \log x_i - \eta \sum_{i=1}^{n} {x_i}^{\tau} \log x_i = 0.$$

Due to the lack of closed-form expressions for these three log-likelihood equations, the likelihood function must be solved iteratively in R software using maximization techniques until sufficiently close parameter values are obtained.

For finding the MLEs $(\hat{\eta}, \hat{\omega}, \hat{\tau})$ of parameters (η, ω, τ) of PQSD, following equations can be solved:

$$\begin{bmatrix} \frac{\partial^2 \log L}{\partial \eta^2} & \frac{\partial^2 \log L}{\partial \eta \partial \omega} & \frac{\partial^2 \log L}{\partial \eta \partial \tau} \\ \frac{\partial^2 \log L}{\partial \omega \partial \eta} & \frac{\partial^2 \log L}{\partial \omega^2} & \frac{\partial^2 \log L}{\partial \omega \partial \tau} \\ \frac{\partial^2 \log L}{\partial \tau \partial \eta} & \frac{\partial^2 \log L}{\partial \tau \partial \omega} & \frac{\partial^2 \log L}{\partial \tau^2} \end{bmatrix}_{\substack{\hat{\eta}=\eta_0 \\ \hat{\omega}=\omega_0 \\ \hat{\tau}=\tau_0}} \begin{bmatrix} \hat{\eta} - \eta_0 \\ \hat{\omega} - \omega_0 \\ \hat{\tau} - \tau_0 \end{bmatrix} = \begin{bmatrix} \frac{\partial \log L}{\partial \eta} \\ \frac{\partial \log L}{\partial \omega} \\ \frac{\partial \log L}{\partial \tau} \end{bmatrix}$$

where η_0, $\omega_0$and $\tau_0$are the preliminary values of η, ω and τ. These equations are resolved iteratively until close estimates of parameters are obtained.

Thus, Fisher's information matrix is obtained as

$$I = -E \begin{bmatrix} \frac{\partial^2 \log L}{\partial \eta^2} & \frac{\partial^2 \log L}{\partial \eta \partial \omega} & \frac{\partial^2 \log L}{\partial \eta \partial \tau} \\ \frac{\partial^2 \log L}{\partial \omega \partial \eta} & \frac{\partial^2 \log L}{\partial \omega^2} & \frac{\partial^2 \log L}{\partial \omega \partial \tau} \\ \frac{\partial^2 \log L}{\partial \tau \partial \eta} & \frac{\partial^2 \log L}{\partial \tau \partial \omega} & \frac{\partial^2 \log L}{\partial \tau^2} \end{bmatrix} = \begin{bmatrix} I_{\eta\eta} & I_{\eta\omega} & I_{\eta\tau} \\ I_{\omega\eta} & I_{\omega\omega} & I_{\omega\tau} \\ I_{\tau\eta} & I_{\tau\omega} & I_{\tau\tau} \end{bmatrix}$$

The solution of Fisher's information matrix will provide asymptotic variance and covariance of the maximum likelihood estimator for $(\hat{\eta}, \hat{\omega}, \hat{\tau})$. The approximate $100(1-\alpha)\%$ confidence intervals for (η, ω, τ) are $\hat{\eta} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{{I_{\eta\eta}}^{-1}}{n}}$, $\hat{\omega} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{{I_{\omega\omega}}^{-1}}{n}}$ and $\hat{\tau} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{{I_{\tau\tau}}^{-1}}{n}}$ respectively, where Z_{α} is the upper $100\alpha^{\text{th}}$ percentile of the standard normal distribution.

Theorem 2. MLEs of PQSD are consistent estimators. That is,
as $n \to \infty$, $P\{|\hat{\eta} - \eta| > \varepsilon\} \to 0$, $P\{|\hat{\omega} - \omega| > \varepsilon\} \to 0$,$P\{|\hat{\tau} - \tau| > \varepsilon\} \to 0$

Proof: The asymptotic variance of MLEs is given by
$V(\hat{\eta}, \hat{\omega}, \hat{\tau}) = \frac{1}{n} I^{-1}(\hat{\eta}, \hat{\omega}, \hat{\tau})$. Therefore, $V(\hat{\eta}, \hat{\omega}, \hat{\tau}) \to 0$ as $n \to \infty$

By Chebyshev's inequality, we have

$P\{|\hat{\eta} - \eta| > \varepsilon\} \leq \frac{V(\hat{\eta})}{\varepsilon^2} = \frac{1}{n}\frac{I_{\eta\eta}{}^{-1}}{\varepsilon^2}$. As $n \to \infty$, $V(\hat{\eta}) = \frac{I_{\eta\eta}{}^{-1}}{n} \to 0$.

Thus, $P\{|\hat{\eta} - \eta| > \varepsilon\} \to 0$ as $n \to \infty$. Hence, $\hat{\eta} \xrightarrow{P} \eta$. Similarly, it can be shown that $\hat{\omega} \xrightarrow{P} \omega$ and $\hat{\tau} \xrightarrow{P} \tau$.

5.2. Maximum Product Spacing Estimation

The maximum product spacing estimates (MPSE) $(\hat{\eta}, \hat{\omega}, \hat{\tau})$ of parameters (η, ω, τ) of PQSD can be obtained numerically by maximizing the following function with respect toη, ω and τ.

$$MPSE = \frac{1}{n+1}\sum_{i=1}^{n+1} log[F(x_i, \eta, \omega, \tau) - F(x_{i-1}, \eta, \omega, \tau)].$$

5.3. Bootstrap Confidence Intervals

The bootstrap is a powerful, data-driven technique for assessing the sampling variability of estimates and constructing confidence intervals without relying on strong parametric assumptions (Efron and Tibshirani, 1993). Let $\widehat{H}$ [where $\widehat{H} = (\hat{\eta}, \hat{\omega}, \hat{\tau})$] be the maximum likelihood estimate of a parameter H, [where $H = (\eta, \omega, \tau)$] based on an observed sample $x = (x_1, x_2, \ldots, x_3)$. The percentile-bootstrap procedure proceeds as follows:

1. **Resampling**: Draw Mbootstrap samples $x_1{}^*, x_2{}^*, \ldots, x_M{}^*$ using sampling with replacement from the original data x.
2. **Re-estimation**: Compute the estimate $\widehat{H}_m$for each bootstrap sample $x_m{}^*$.
3. **Bootstrap Distribution**: The set $\{H_1{}^*, H_2{}^*, \ldots, H_M{}^*\}$ empirically approximates the sampling distribution of $\widehat{H}$.
4. **Percentile Interval**: For a nominal $100(1-\alpha)$%interval, take $\frac{\alpha}{2}$and $1 - \frac{\alpha}{2}$ quantiles of the $\{\widehat{H}_m\}$as the confidence limits (Davison and Hinkley, 1997).

Because it does not assume normality of $\widehat{H}$ or rely on Fisher's information, the bootstrap is particularly useful for complex models, small samples, or when the likelihood surface is irregular.

6. A simulation study

To understand the flexibility and performance of the maximum likelihood estimators (MLEs) of PQSD, we carried out a simulation study. We looked at the variances, mean square errors (MSEs), biases (B), mean estimates and the approximate confidence intervals of MLEs. The findings are shown in Tables 1, 2, and 3. Using the formulas listed below, the mean, bias, MSE, and variance are computed.

$Mean = \frac{1}{n}\sum_{i=1}^{n}\widehat{H}_i$, $B = \frac{1}{n}\sum_{i=1}^{n}(\widehat{H}_i - H)$, $MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{H}_i - H)^2$, $Variance = MSE - B^2$, where $H = (\eta, \omega, \tau)$ denotes the true parameter vector and $\widehat{H}_i = (\hat{\eta}_i, \hat{\omega}_i, \hat{\tau}_i)$denotes the estimates of the parameter vector from the ith simulated sample.

The simulation used to carry out this work is run in the following manner:

a. The acceptance-rejection simulation technique is used to generate data. The steps in this method are as follows:
   i. Generating Y distributed as Gamma(η, ω)
   ii. Generating U distributed as Uniform$(0,1)$
   iii. If $U \leq \frac{f(y)}{Mg(y)}$, then set $X = Y$, (accept the sample and if not reject the sample), if rejected, go through steps (i–iii) again until the appropriate samples are obtained.
   Here $M$is a fixed number.
b. Each sample size is replicated 10000 times.

Means, Biases, MSEs, and variances of MLEs of the PQSD parameters decrease and the mean value of PQSD tends to the true parameter value with an increasing sample size, which is consistent with the first-order asymptotic theory of MLE.

Table 1. Descriptive constants of PQSD for $\eta = 3$,$\omega = 3$and $\tau = 2.7$

Parameters	Sample size	Mean	Bias	MSE	Variance	95% CI	
						Lower	Upper
η	20	3.02511	0.02511	0.00346	0.00283	3.00179	3.04842
	50	3.02023	0.02024	0.00323	0.00282	3.00551	3.03495
	100	3.01906	0.01906	0.00249	0.00213	3.01001	3.02810
	200	3.01597	0.01597	0.00223	0.00197	3.00981	3.02212
	300	3.01348	0.01348	0.00201	0.00183	3.00863	3.01832
ω	20	3.13014	0.13014	0.06463	0.04770	3.03442	3.22585
	50	3.06471	0.06471	0.04583	0.04164	3.00814	3.12127
	100	3.05477	0.05477	0.03658	0.03358	3.00890	3.10064
	200	3.04411	0.04411	0.02184	0.01989	3.02456	3.06365
	300	3.03385	0.03362	0.01485	0.01372	3.02059	3.04710
τ	20	2.72175	0.02175	0.00107	0.00060	2.71101	2.73248
	50	2.71661	0.01661	0.00087	0.00059	2.70987	2.72334
	100	2.71390	0.01390	0.00068	0.00048	2.70960	2.71819
	200	2.70844	0.00844	0.00051	0.00044	2.70553	2.70553
	300	2.70508	0.00508	0.00046	0.00043	2.70273	2.70742

Table 2. Descriptive constants of PQSD for $\eta = 2, \omega = 1.4$and $\tau = 1.7$

Parameters	Sample size (n)	Mean	Bias	MSE	Variance	95% CI	
						Lower	Upper
η	20	2.02125	0.02125	0.00344	0.00299	1.99728	2.04521
	50	2.01888	0.01888	0.00261	0.00225	2.00573	2.03202
	100	2.01690	0.01690	0.00213	0.00185	2.00847	2.02533
	200	2.01369	0.01369	0.00149	0.00131	2.00867	2.01870
	300	2.01159	0.01159	0.00124	0.00110	2.00783	2.01534
ω	20	1.35863	-0.04137	0.00964	0.00792	1.31962	1.39763
	50	1.36715	-0.03284	0.00862	0.00754	1.34308	1.39121
	100	1.36826	-0.03173	0.00806	0.00705	1.35180	1.38471
	200	1.38303	-0.01696	0.00695	0.00667	1.37171	1.39434
	300	1.38509	-0.01490	0.00579	0.00558	1.37663	1.39354
τ	20	1.66402	-0.03597	0.01060	0.00930	1.62175	1.70628
	50	1.67882	-0.02117	0.00464	0.00419	1.66087	1.69676
	100	1.68551	-0.01448	0.00273	0.00252	1.67567	1.69534
	200	1.68922	-0.01077	0.00156	0.00144	1.68396	1.69447
	300	1.69394	-0.00605	0.00121	0.00117	1.69006	1.69781

Table 3. Descriptive constants of PQSD for $\eta = 0.3, \omega = 0.3$,and $\tau = 0.3$

Parameters	Sample size (n)	Mean	Bias	MSE	Variance	95% CI	
						Lower	Upper
η	20	0.31240	0.01240	0.00350	0.00330	0.28722	0.33757
	50	0.30870	0.00870	0.00280	0.00270	0.29429	0.32310
	100	0.30590	0.00590	0.00210	0.00210	0.29691	0.31488
	200	0.30380	0.00380	0.00160	0.00160	0.29825	0.30934
	300	0.30210	0.00210	0.00130	0.00130	0.29801	0.30618
ω	20	0.32470	0.02470	0.06460	0.06400	0.21382	0.43557
	50	0.31520	0.01520	0.04580	0.04560	0.25600	0.37439
	100	0.30980	0.00980	0.03660	0.03650	0.27235	0.34724
	200	0.30640	0.00640	0.02180	0.02180	0.28593	0.32686
	300	0.30430	0.00430	0.01490	0.01480	0.29053	0.31806
τ	20	0.29530	-0.00470	0.00110	0.00100	0.28144	0.30915
	50	0.29760	-0.00240	0.00090	0.00090	0.28928	0.30591
	100	0.29890	-0.00110	0.00070	0.00070	0.29371	0.30408
	200	0.29960	-0.00040	0.00050	0.00050	0.29650	0.29650
	300	0.29980	-0.00020	0.00050	0.00050	0.29726	0.30233

7. Applications

The applications of PQSD have been evaluated using the two real-life datasets from the flood peaks of the Wheaton River and engineering and both datasets are over-dispersed shown by the data summary of both datasets in Table 4 as PQSD is suitable for over-dispersed data. The datasets are as follows.

Dataset 1: The following right-skewed data present the exceedances of flood peaks (in m3/s) of the Wheaton River near Carcross in Yukon Territory, Canada. The data consist of 72 exceedances for the years 1958–1984, rounded to one decimal place. This data were analyzed by Choulakian and Stephens (2001) and are given as follows:

1.7, 2.2, 14.4, 1.1, 0.4, 20.6, 5.3, 0.7, 1.9, 13, 12, 9.3, 1.4, 18.7, 8.5, 25.5, 11.6, 14.1, 22.1,1.1, 2.5, 14.4, 1.7, 37.6, 0.6, 2.2, 39, 0.3, 15, 11, 7.3, 22.9, 0.1, 1.7, 1.1, 0.6, 9, 1.7, 7, 20.1,0.4, 2.8, 14.1, 9.9, 10.4, 10.7, 30, 3.6, 5.6, 30.8, 13.3, 4.2, 25.5, 3.4, 11.9, 21.5, 27.6, 36.4, 2.7, 64, 1.5, 2.5, 27.4, 1, 27.1, 20.2, 16.8, 5.3, 9.7, 27.5, 2.5, 27.

Dataset 2: The following right-skewed data set discussed by Picciotto R. (1970) is used to correspond to the time-to-failure of a polyester/viscose yarn in a textile experiment for testing the tensile fatigue characteristics of yarn. It consists of a sample of 100 cm yarn at 2.3% strain level. The values are:

86, 146, 251, 653, 98, 249, 400, 292, 131, 169, 175, 176, 76, 264, 15, 364, 195, 262, 88, 264, 157, 220, 42, 321, 180, 198, 38, 20, 61, 121, 282, 224, 149, 180, 325, 250, 196, 90, 229, 166, 38, 337, 65, 151, 341, 40, 40, 135, 597, 246, 211, 180, 93, 315, 353, 571, 124, 279, 81, 186, 497, 182, 423, 185, 229, 400, 338, 290, 398, 71, 246, 185, 188, 568, 55, 55, 61, 244, 20, 284, 393, 396, 203, 829, 239, 236, 286, 194, 277, 143, 198, 264, 105, 203, 124, 137, 135, 350, 193, 188.

Table 4. Summary of the dataset 1 and 2

Datasets	Minimum	1st Quartile	Median	Mean	3rdQuartile	Maximum	Variance
1	0.100	2.125	9.500	12.204	20.125	64.000	151.221
2	15.000	129.200	195.500	222.000	282.500	829.00	20914.380

The parameter estimates of PQSD, TPPSD, PQLD, GGD, ATPSD and TPGLD along with their standard errors for datasets 1 and 2, obtained using the MLE and MPSE methods, are displayed in Tables 5 and 6. The values of $-2\,log\,L$, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Consistent Akaike Information Criterion (CAIC), Hannan-Quinn Information Criterion (HQIC), Kolmogorov-Smirnov (K-S) Statistics for the above two datasets have been computed and presented in Tables 7 and 8 using the formulas:

$AIC = -2\,log\,L + 2p,$

$BIC = -2\,log\,L + p\,log(n),$

$CAIC = -2\,log\,L + 2\frac{pn}{n-p-1}$,

$HQIC = -2\,log\,L + 2p\,log(log(n))$ K-S $= \underset{x}{Sup}|G_n(x) - G_0(x)|$, where p =the number of parameters, n =sample size, $G_n(x) =$ empirical cdf of the considered distribution and $G_0(x) =$ cdf of the considered distribution.

The Bootstrap confidence intervals for the dataset 1 and 2 are presented in Table 9.

Table 5. MLE and MPSE of the parameters for the dataset 1

Distributions	MLE of the dataset 1			MPSE of the dataset 1		
	$\hat{\eta}$ SE($\hat{\eta}$)	$\hat{\omega}$ SE($\hat{\omega}$)	$\hat{\tau}$ SE($\hat{\tau}$)	$\hat{\eta}$ SE($\hat{\eta}$)	$\hat{\omega}$ SE($\hat{\omega}$)	$\hat{\tau}$ SE($\hat{\tau}$)
PQSD	0.2050 (0.0794)	15.1360 (13.4909)	0.8862 (0.0978)	0.2332 (0.0944)	12.9389 (12.4735)	0.8366 (0.0985)
TPPSD	0.8085 (0.3134)	0.1000 (31.0460)	0.5338 (0.1852)	0.4005 (0.0001)	2.0959 (0.7919)	3.8690 (0.0001)
PQLD	0.1536 (0.0831)	2.5343 (4.2815)	0.8704 (0.1102)	0.1647 (0.0996)	4.8160 (15.3272)	0.7985 (0.0946)
GGD	0.0143 (0.0215)	0.5382 (0.2058)	1.3659 (0.3619)	0.1000 (0.0624)	0.9525 (0.2205)	0.9168 (0.1344)
ATPSD	0.1442 (0.0176)	0.0100 (0.0104)	0.0100 (…)	0.1635 (0.0195)	30.4630 (41.8227)	65.41357 (92.6840)
TPGLD	0.1536 0.0832	16.4826 35.8484	0.8704 0.1102	0.1758 (0.1070)	12.9611 (31.5808)	0.8210 (0.1174)

Table 6. MLE and MPSE of the parameters for the dataset 2

Distributions	MLE of the dataset 2			MPSE of the dataset 2		
	$\hat{\eta}$ SE($\hat{\eta}$)	$\hat{\omega}$ SE($\hat{\omega}$)	$\hat{\tau}$ SE($\hat{\tau}$)	$\hat{\eta}$ SE($\hat{\eta}$)	$\hat{\omega}$ SE($\hat{\omega}$)	$\hat{\tau}$ SE($\hat{\tau}$)
PQSD	0.0282 (0.0214)	0.2987 (6.3510)	0.8663 (0.1279)	0.0178 (0.0097)	10.0303 (14.5112)	0.9380 (0.0915)
TPPSD	0.0100 (0.0023)	0.3000 247.5289	1.0490 (0.0423)	0.0118 (0.0181)	1.9570 (79.3259)	0.9633 (0.2466)
PQLD	0.0255 (0.0102)	0.2429 (…)	0.8019 (0.0751)	0.0889 (0.0004)	2.2284 (0.0004)	0.8524 (0.0004)
GGD	0.0284 (0.0136)	2.9520 (0.4996)	0.8611 (0.0649)	0.0315 (0.0958)	2.8945 (2.9351)	1.5122 (0.7290)
ATPSD	0.0135 0.0007	31.0945 (…)	0.0168 (0.2214)	0.0135 (0.0008)	4.3605 (1.2517)	0.0100 (0.1974)
TPGLD	0.1000 (0.0174)	0.1000 (0.6290)	0.5822 (0.0372)	0.0070 (0.0021)	1.6096 (9.2574)	1.0398 (0.0510)

Table 7. Goodness of fit measures for dataset 1

Distributions	$-2\,log\,L$	AIC	BIC	CAIC	HQIC	MLE		MPSE	
						K-S	p-value	K-S	p-value
PQSD	**500.60**	**506.60**	**513.43**	**506.95**	**509.31**	**0.10**	**0.50**	**0.08**	**0.72**
TPPSD	505.10	511.10	517.93	511.45	513.82	0.13	0.22	0.72	0.00
PQLD	502.73	508.73	515.56	509.08	511.44	0.16	0.07	0.15	0.09
GGD	502.16	508.16	514.99	508.51	510.87	0.15	0.09	0.09	0.61
ATPSD	505.47	511.47	518.30	511.82	514.18	0.16	0.05	0.29	0.00
TPGLD	502.73	508.73	515.56	509.08	511.44	0.13	0.19	0.12	0.30

Table 8. Goodness of fit measures for dataset 2

Distributions	$-2\,log\,L$	AIC	BIC	CAIC	HQIC	MLE		MPSE	
						K-S	p-value	K-S	p-value
PQSD	**1251.10**	**1257.10**	**1264.91**	**1257.35**	**1260.26**	**0.11**	**0.26**	**0.06**	**0.90**
TPPSD	1559.25	1565.25	1573.06	1565.50	1568.41	0.14	0.06	0.11	0.36
PQLD	1274.31	1280.31	1288.12	1280.56	1283.47	0.22	0.00	0.79	0.00
GGD	1251.27	1257.27	1265.08	1257.52	1260.43	0.19	0.00	0.92	0.00
ATPSD	1255.82	1261.82	1269.63	1262.07	1264.98	0.18	0.00	0.07	0.69
TPGLD	1301.53	1307.53	1315.34	1307.78	1310.69	0.39	0.00	0.07	0.69

Table 9. Bootstrap confidence interval for dataset 1 and 2

Datasets	Parameters	95% CI	
		Lower	Upper
1	$\hat{\eta}$	0.1286	0.2863
	$\hat{\omega}$	8.2674	38.2731
	$\hat{\tau}$	0.7762	1.0430
2	$\hat{\eta}$	0.0074	0.0522
	$\hat{\omega}$	0.0010	41.6467
	$\hat{\tau}$	0.7642	1.0934

From Tables 7 and 8 it is obvious that PQSD provides better fit as compared to TPPSD, PQLD, GGD, PSD, ATPSD and TPGLD because PQSD has the least $-2\,log\,L$, AIC, BIC, CAIC, HQIC and K-S values. From the K-S values given by MLE and MPSE, MPSE is better than MLE in terms of the fit for both the datasets. Further, PQSD offers a far better fit than TPPSD, PQLD, GGD, ATPSD, TPGLD as shown by the fitted plot, quantile-quantile (Q-Q) plot, probability-probability (P-P) plot and empirical cumulative density function (ECDF) plot in Figures 5 and 6.

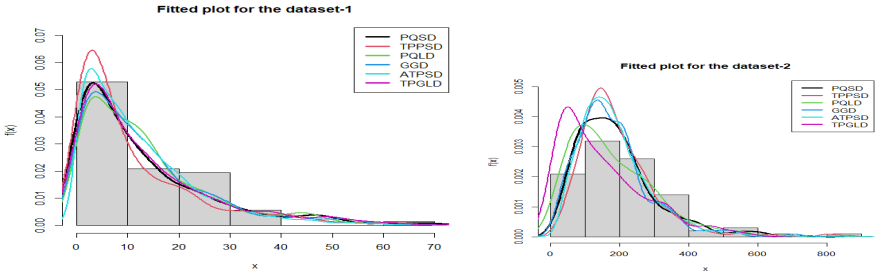


Figure 5. Fitted plot of distributions for the dataset 1 and 2

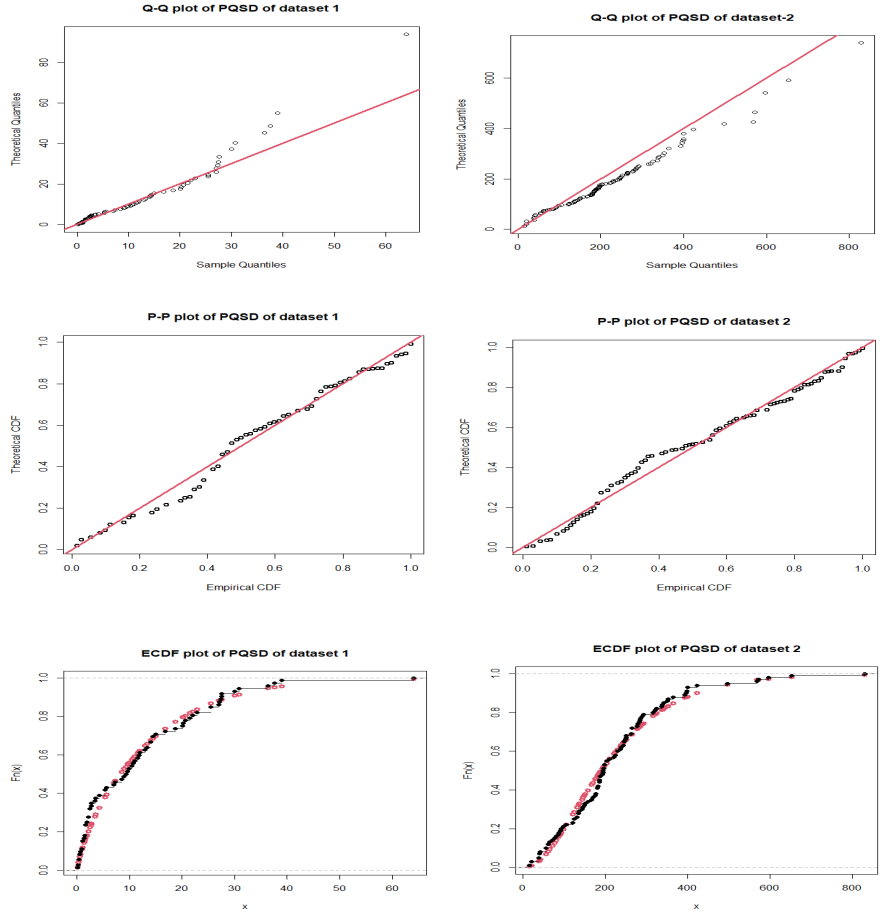


Figure 6. Q-Q plot, P-P plot and ECDF plot of PQSD distribution for the dataset 1 and 2

8. Conclusion

In this paper, PQSD has been proposed. Its moments and statistical properties, such as the survival function, hazard function, reverse hazard function, mean residual life function and stochastic ordering have been analyzed. Parameters of the distribution have been estimated using maximum likelihood estimation and maximum product spacing estimation. To assess the efficiency of the maximum likelihood estimates of the parameters, a simulation study has been presented. The confidence interval of the parameters has been obtained using the Bootstrap confidence interval method. Moreover, applications have been explored for two real lifetime datasets with unimodality and over-dispersion, and the goodness of fit of PQSD has been compared with TPPSD, PQLD, GGD, ATPSD and TPGLD. It has been found that PQSD provides a better fit than TPPSD, PQLD, GGD, ATPSD and TPGLD. The future directions include exploring Bayesian estimation method for the proposed distribution under different loss function, developing the multivariate and the discrete counterpart of the proposed distribution and extending the distribution to deal with the quality control of the product and study regression modelling.

Ackowledgements

Authors are grateful to the Editor-in-Chief of the journal and the three anonymous reviewers for their valuable comments, which improved the quality of the paper.

Declarations of interest: None.

References

Aderoju, S., Adeniyi, I., (2022). On Power Generalized Akash Distribution with Properties and Applications. *Journal of Statistical Modeling and Analytics*, Vol. 4(1), pp. 1–13.

Alkarni S., (2015). Power quasi-Lindley distribution. Properties and application. *Asian Journal of Mathematics and Computer Research*, Vol. 10(2), pp. 179–195.

Choulakian, V., Stephens, M. A., (2001). Goodness-of-Fit tests for the generalized Pareto distribution. *Technometrices*, Vol. 43, pp. 478–484.

Davison, A. C., Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Ghitany, M.E., Atieh, B. and Nadarajah, S., (2008). Lindley distribution and its applications. *Mathematics and Computer in Simulation*, Vol. 78, pp. 493–506.

Ghitany, M. E., Al-Mutairi, D. K., Balakrishnan N. and Al-Enezi, L. J., (2013). Power Lindley distribution and associated inference. *Computational Statistics* & *Data Analysis*, Vol. 64, pp. 20–33.

Lindley, D. V., (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society*, Series B, Vol. 20(1), pp. 102–107.

Murthy, D. N. P, Xie, M. and Jiang, R., (2004), *Weibull models*, John Wily & Sons Inc., Hoboken.

Mussie, T., Shanker, R., (2018). A two parameter Sujatha distribution. *Biometrics* & *Biostatistics International Journal*, Vol. 7(3), pp. 188–197.

Nosakhare, E. and Festus, O., (2018). A three-parameter generalized Lindley distribution: Properties and Application. *Statistica*, Vol. 78(3), pp. 1–17.

Nwikpe, B. J., Iwok, I. A., (2020). A three-parameter Sujatha distribution with application. *European Journal of Statistics and Probability*, Vol. 8(3), pp. 22–34.

Picciotto R., (1970). *Tensile fatigue characteristics of a sized polyester/viscose yarn and their effect on weaving performance*, Master thesis, University of Raleigh, North Carolina State, USA.

Prodhani, H. R., Shanker, R., (2023). On some statistical properties and applications of three-parameter Sujatha distribution. Reliability: *Theory* & *Applications*, Vol. 18(3), pp. 514–527.

Prodhani, H. R., Shanker, R., (2024). A three-parameter power Sujatha distribution with properties and application. *International Journal of Statistics and Reliability Engineering*, Vol. 11(1), pp. 70–78.

Prodhani, H. R., Shanker, R., (2024). Power Pratibha distribution with properties and applications. *International Journal of Statistics and Reliability Engineering*, Vol. 11(2), pp. 217–226.

Shanker, R., Mishra, A., (2013). A quasi-Lindley distribution. African *Journal of Mathematics and Computer Science Research*, Vol. 6(4), pp. 64–71.

Shanker, R., (2016a). Sujatha distribution and its applications. *Statistics in Transition New Series*, Vol. 17(3), pp. 391–410.

Shanker, R., (2016b). A Quasi Sujatha distribution. *International Journal of Probability and Statistics,* Vol. 5(4), pp. 89–100.

Shanker, R., (2016c). Aradhana Distribution and Its applications. *International Journal of Statistics and Applications*, Vol. 6(1), pp. 23–34.

Shanker, R., Shukla, K. K., (2018). A two-parameter power Aradhana distribution with properties and application. *Indian Journal of Industrial and Applied Mathematics*, Vol. 9(2), pp. 210–220.

Shanker, R., Shukla, K. K and Sigh, A. P., (2018). A generalized Akash distribution. *Biometrics & Biostatistics International Journal*, Vol. 7(1), pp. 18–26.

Shanker, R., Shukla, K. K., (2019). A two-parameter power Sujatha distribution with properties and application. *International Journal of Mathematics and Statistics*, Vol. 20(3), pp. 11–22.

Shanker, R., (2023). Pratibha distribution with properties and application. *Biometrics & Biostatistics International Journal*, Vol. 12(5), pp. 136–142.

Stacy, E. W., (1962). A generalized gamma distribution. *Annals of Mathematical Statistics*, Vol. 33, pp. 1187–1192.

Shukla, K. K., (2018). Pranav distribution with properties and its applications. *Biometrics & Biostatistics International Journal*, Vol. 7(3), pp. 244–254.

Shukla, K. K., (2019). Power Pranav distribution and its applications to model lifetime data. *Journal of applied quantitative methods*, Vol. 14(2), pp. 1–13.

Shaked, M., Shanthikumar, J., (1994). *Stochastic Orders and Their Applications*. Academic Press, New York.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 71–89, https://doi.org/10.59139/stattrans-2026-017
Received – 06.03.2025; accepted – 22.09.2025

Consumption patterns of Slovak households in 2021 and 2022

Lívia Krajčíková[1], **Mária Vojtková**[2]

Abstract

Household consumption behavior is a key indicator of the economic situation and social disparities in society. The aim of this article is to analyze the similarities and differences in the structure of consumption expenditure of various types of Slovak households in 2021 and 2022. The study also focuses on the impact of demographic factors on spending behavior, examining how households in different income groups allocate their expenses. To identify and profile individual household segments, we applied cluster analysis, which enabled us to distinguish homogeneous groups of households based on their spending patterns. The analysis was based on data from the Household Budget Survey, provided by the Statistical Office of the Slovak Republic for scientific purposes. The results indicate that Slovak households can be divided into six main segments, with four segments displaying stable spending patterns over both analyzed years, and two where spending patterns varied from one of the studied years to another. The results emphasize significant differences between the expenditure structures of low- and high-income households, as well as among those of households with varying demographic compositions. Our findings contribute to a deeper understanding of Slovak households' consumption behavior and can be used in the development of socio-economic policies targeting various income and demographic groups.

Key words: households, Household Budget Survey, consumption, expenditure, cluster analysis.

1. Introduction

An analysis of household consumption expenditure is important from several points of view. Knowing the structure of consumer spending allows government bodies and economists to plan economic policies, allocate social benefits and tax breaks more effectively. It can also contribute to a better understanding of consumer behavior. Businesses and marketers can use segmentation to target their products and services to specific groups of households according to their needs and preferences. In addition, the

[1] Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava, Bratislava, Slovakia. E-mail: livia.krajcikova@euba.sk. https://orcid.org/0009-0005-4661-4432.

[2] Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava, Bratislava, Slovakia. E-mail: maria.vojtkova@euba.sk. https://orcid.org/0000-0001-6257-901X.

analysis reveals social disparities between households with different incomes, which can be important in social policy-making or support for vulnerable groups.

The aim of this article is to identify and profile segments of Slovak households based on their consumption expenditures in 2021 and 2022. We focus on the impact of demographic factors on spending behavior and examine how households in different income groups allocate their spending. To achieve the set goal, we used the cluster analysis method.

We worked with anonymized microdata from the Household Budget Survey (HBS) on selected Slovak households for 2021 and 2022, which were provided by the Statistical Office of the Slovak Republic for scientific purposes. We worked with SAS Enterprise Guide (SAS EG) software to process and analyze the data.

The Household Budget Survey is a national statistical survey, the purpose of which is to obtain information on the amount, structure and development of monetary expenditures as well as incomes and non-monetary expenditures and incomes of people in different types of households, with the aim of using the information obtained to support public administration bodies in developing and monitoring public policies, implementing the European Statistical Programme, and assessing adjustments to subsistence minimum levels. Another objective of the Household Budget Survey is, for example, to use the obtained information as one of the bases for assessing the state of the company's economy, for analyzing the standard of living of the population and for international comparison with other countries (Vlačuha, Hornáček, Vargová, 2023). When determining consumption expenditure, the international classification of individual consumption according to the purpose of use (COICOP), recommended by Eurostat (2003) for statistics on family accounts, is used. According to this classification, consumption expenditure is divided into 12 basic divisions, as stated in the publication Classification of Individual Consumption According to Purpose 2018 by United Nations (2023): Food and non-alcoholic beverages (FAB); Alcoholic beverages, tobacco and narcotics (ABT); Clothing and footwear (CLT); Housing, water, electricity, gas and other fuels (HSG); Furnishing, household equipment and routine household maintenance (FUR); Health (HLT); Transport (TRA); Communication (COM); Recreation and culture (CUL); Education (EDU); Restaurants and hotels (RES); Miscellaneous goods and services (MGS).

In addition to household income and expenditure, the basic demographic and socio-economic characteristics of households, such as the region from which the household originates, the number of household members, the number of dependent children or the main economic activity of the reference person, are also obtained when determining Household Budget Survey. The main objective of HBS at the national level is to calculate the weights for the consumer price index. HBS, was launched in the 1960s,

is carried out in all countries of the European Union and in most countries. Eurostat publishes the data collected by each EU member state at 5-year intervals (Eurostat, 2003).

To ensure comparability of survey data between countries and over time, methodologies and guides issued by Eurostat are used. The key document is the Household budget surveys in the EU: methodology and recommendations for harmonization from 2003. The purpose of this paper was to describe the current methodology used for HBS and to propose recommendations for further harmonization and improvement of the quality and comparability of survey data at the European level.

2. Literature review

In the literature, we can come across various studies that confirm the importance of HBS in the study of socio-economic trends as well as the use of cluster analysis in identifying similar groups of households.

Dogan et al. (2019) use HBS data for Turkey from 2017 and apply cluster analysis to identify different groups of households with different levels of health spending. The results show that household type, income, and factors such as physical activity influence health care spending.

Froemelt et al. (2018) use HBS data for Switzerland from 2009 to 2011 to identify different consumption patterns of household consumption and assess the environmental impacts associated with specific consumer behavior. In analysis, they use a two-level cluster analysis method to identify behavior patterns.

Değirmenci and Özbakır (2017) also look at the use of data mining techniques to analyze HBS data in Turkey. The authors applied cluster analysis to characterize household types.

HBS is a key source of microdata not only for analyzing consumption and savings, but also for analyzing poverty. Antonin (2020) examines the relationship between income and savings of French households. Bouzarovski and Tirado-Herrero (2016) examine energy poverty in Hungary, the Czech Republic and Poland, while Cupák et al. (2015) analyze food demand in Slovakia. The development and importance of HBS in the Czech Republic is mapped by Vopravil and Linhartová-Jiřičková (2024). Labudová et al. (2010) use HBS data and the principal components analysis to measure the socio-spatial dimension of poverty in the regions of Slovakia and the Czech Republic.

Morvay et al. (2005) focus on profiling groups of households according to income, namely:

- households with high pensions, whose expenditure structure consists of luxury commodities and who live an above-standard lifestyle,

- standard Slovak households, which represent the wider middle social classes and whose expenditure structure copies their income situation with a preference for saving consumption in addition to the consumption of essential needs,
- low-income households, whose incomes consist more of social resources and whose expenditure structure consists of meeting basic living needs with a deficit in basic elements of nutrition.

Across all income groups, the consumption structure prioritized food and housing-related services, although the share of essential expenditures varied. In the lowest income groups, spending on basic needs exceeded 50% of total expenditures. In the highest-income households, food accounted for approximately 20%, housing for 13%, and transport for 16% of total spending. These households also allocated more to holidays and culture. Meanwhile, education expenditures remained at 1% of total spending across all income groups.

Morvay et al. (2005) assume that expenditure on food and housing will represent the main component of consumer expenditure of Slovak households in the long term. In Slovakia, a model of expenditure structure has been fixed, according to which families spend the largest share of consumption expenditure on food and non-alcoholic beverages. It is assumed that consumption will continue to copy the state of real incomes, and it will take a longer period of time to approach the expenditure structure of the European Union countries. Households in the European Union have the highest expenditures on housing and transport, spending less on food and clothing, and more on culture and recreation.

Morvay (2023) monitors the consumption of Slovak households, which grew in 2022 despite many negative factors and crises, such as inflation, the energy and price crisis. Possible causes of the massive growth in consumption are compensation for postponed consumption during the pandemic period, state aid, employment growth, or the accumulation of stocks due to fears of further price increases or concerns about the effects of the war. In this period, there was also a dramatic decline in the creation of savings, not only of households, but in the overall economy of the Slovak Republic. Households are losing caution in creating savings, because they are convinced that even in future crises, the state will help them compensate for the negative impacts with the help of public funds, while the role of the state is only to mitigate shocks and fluctuations in the economy. The author suggests that economic policy should consider new phenomena in the behavior of households and not create unrealistic expectations towards the state. At the same time, he recommends that aid should be concentrated on those who are at risk of poverty or cannot cope with economic shocks.

Inflation has a significant impact on the increase in food, energy and gasoline prices. Slovaks donate most of their income to provide basic needs such as food and housing. Opportunities for savings in these areas can be buying cheaper food or using

energy more efficiently. However, it is easier to cut back on other expenses such as culture, entertainment, travel, clothing, and restaurants. These aspects can be replaced by cheaper alternatives or can even be eliminated altogether. On the contrary, it is impossible to completely omit basic needs. High inflation leads to certain changes in consumer behavior. However, inflation is not the only reason for rising food and energy prices, there have been many negative events such as the economic crisis, refugee crisis, pandemic or war recently, which affect not only the prices of consumer items, but also consumer behavior (Galvánková, 2022).

The benefit of the classification of Slovak households in this article for 2021 and 2022 based on HBS is not only the profiling of consumer behavior according to the structure of expenditure, but also the consideration of income, household type and household structure by age.

3. Methodology

3.1. Principal Component Analysis

Most input variables do not satisfy the assumption of independence required for cluster analysis. To address this, we first applied the principal component method, which transformed the original variables into independent principal components. These new variables were then used for clustering.

Principal Component Analysis (PCA) is a statistical technique that uncovers hidden relationships among observed variables. Its primary objective is to reduce the number of original variables while retaining as much information as possible.

The newly derived variables, known as principal components, are denoted as Y_1, Y_2, ..., Y_p, where p represents the number of observed variables X_1, X_2, ..., X_p. These principal components are linear combinations of the original variables, ensuring their independence. Each principal component can be expressed as a weighted sum of the original variables X_1, X_2, ..., X_p, with weights a_{ij} determining their contribution.

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + \cdots a_{1p}X_p \\ &\ldots \\ Y_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots a_{pp}X_p \end{aligned} \tag{1}$$

To calculate the weights a_{ij}, it is necessary to compute the characteristic equation $\det(S - \lambda \cdot I) = 0$, where S is the covariance matrix, λ represents the eigenvalues of the polynomial equation of order p and I is the identity matrix. By solving the characteristic equation, we obtain the eigenvalues λ_i, which are then used to calculate the eigenvector a_i, containing the weights a_{ij} (Vojtková, Stankovičová, 2020).

A detailed overview of the principal component method, including criteria for selecting the number of principal components, rotation techniques, result interpretation,

and visualization, is provided by Jolliffe (2002). He also demonstrates the use of PCA in combination with other methods, highlighting its wide-ranging applications in statistical analysis.

The principal component method can be used to reduce dimensionality or to obtain uncorrelated variables before applying further analyses, such as cluster analysis.

3.2. Cluster Analysis

Cluster analysis is a set of mathematical and statistical methods used to divide a set of objects $X_i (i = 1,2, \dots, n)$ into several unspecified groups (clusters) $C_1, C_2, \dots, C_q (2 \leq q \leq n)$ in such a way that objects within the same cluster are as similar as possible, while objects from different clusters are as dissimilar as possible. Cluster analysis can be performed using various clustering procedures and methods (Vojtková, Stankovičová, 2020).

To assess the similarity between objects, we used distance measures that satisfy the properties of positivity, symmetry, and the triangle inequality. The most used measures include Euclidean distance, city block distance, Minkowski distance, Canberra distance, and others, along with their calculation formulas, as presented by Everitt et al. (2011).

In our analysis, we used Euclidean distance, which assumes the uncorrelated nature of input variables and is calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - X_{jk}\right)^2} \tag{2}$$

where

X_{ik} is the value of the k-th variable for the i-th object,

X_{jk} is the value of the k-th variable for the j-th object.

Clustering procedures are iterative processes used to form clusters of objects and can be either hierarchical or non-hierarchical. Hierarchical methods that start with each element as a separate cluster are called agglomerative, while those that begin with one large cluster and gradually divide into smaller ones are called divisive. Hierarchical methods do not require the number of clusters to be known at the beginning of the analysis and can be visually represented using a dendrogram.

Hierarchical clustering methods include the nearest neighbor method, the farthest neighbor method, the average linkage method, the centroid method, the median method, and Ward's method. Non-hierarchical methods, on the other hand, require the number of clusters to be specified at the beginning of the analysis. Examples of non-hierarchical methods include the method of typical points and the k-means method (Vojtková, Stankovičová, 2020).

In our application of cluster analysis, we used Ward's method:

$$ESS = \sum_{i=1}^{n_h} \sum_{h=1}^{q} \left(X_{hi} - \bar{X}_{C_h}\right)^2 \tag{3}$$

where

n_h - number of objects in cluster C'_h,
$\bar{X}_{C_h}$ - vector of mean values of the variable in cluster C'_h,
X_{hi} – vector of variable values for the i-th object in cluster C_h.

Selecting the optimal number of clusters is a key step in cluster analysis. Everitt et al. (2011) emphasize that there is no universally correct solution and recommend combining multiple metrics with visual methods such as a dendrogram. The authors mention formal techniques designed to eliminate the issue of subjectivity, such as GAP-statistics.

Vojtková and Stankovičová (2020) present various characteristics for determining the number of clusters, such as the coefficient of determination, semi-partial coefficient of determination, and the cubic clustering criterion, which provide objective criteria for decision-making.

4. Results and discussion

In this section, we focus on the results of the segmentation of Slovak households based on consumer expenditures in 2021 and 2022.

Most of the input variables do not meet the assumption of independence required for entry into cluster analysis. Therefore, before applying cluster analysis, we used the principal component method to create independent variables – principal components and proceeded with them in the clustering process. As input for cluster analysis, we used 7 principal components for the year 2021 and 5 principal components for the year 2022.

The segmentation revealed six main household groups in both years, differing not only in their level of expenditures but also in household composition and income characteristics. Each segment was analyzed in detail regarding average expenditures across various categories, as well as household structure. In both years, we identified four similar and two distinct household groups.

The selection of the number of household clusters in cluster analysis can be based on the dendrogram (Figure 1), as well as the interpretability of the results and the balance of the formed clusters. When determining the number of clusters, we considered four to six clusters. However, for better interpretability and a more even distribution of clusters, we decided to use six clusters for both 2021 and 2022.

Tables 1 (2021) and 2 (2022) present the number of households and the average household expenditures per person per year in each cluster, supplemented by the average net household income per person per year. The average income and expenditures are given in absolute values. When comparing the average levels of individual expenditures across clusters, the lowest values are highlighted in green, while the highest values are highlighted in red (this color coding is used in all tables).

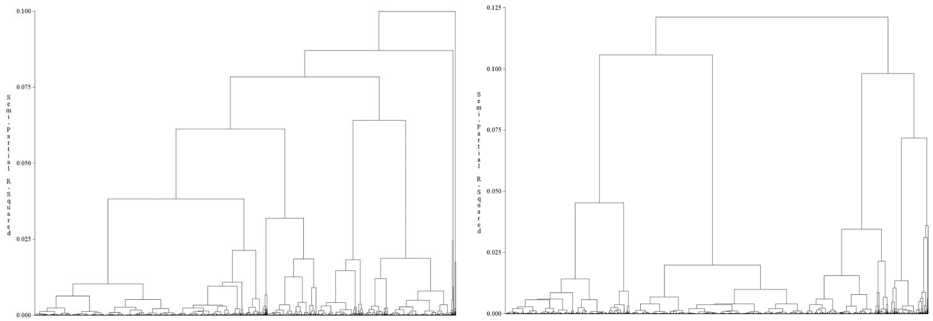


Figure 1. Dendrogram of household clustering based on expenditures using Ward's method (left: year 2021, right: year 2022)

Source: Household Budget Survey, own processing in SAS EG.

Based on Slovak households' expenditures in 2021 and 2022, we identified four segments that appeared in both years. These segments reflect stable consumption patterns across the analyzed years. Additionally, we observed two segments specific to 2021 and two other segments that emerged only in 2022. This difference indicates changes in consumer behavior, which may be influenced by various economic, social, or other factors specific to each year.

Table 1. Average consumer expenditures and average net household incomes in individual clusters (per year and per person) – year 2021

Cluster	Number of households	Net household income	Average consumer expenditures (per year and per person)					
			FAB	ABT	CLO	HSG	FUR	HLT
1	1266	11932.59	2058.33	158.64	172.56	1501.55	354.33	327.79
2	452	11915.67	1878.81	854.33	305.48	1467.51	368.49	166.12
3	1778	9183.92	1276.04	128.90	227.21	1506.91	270.11	217.65
4	1067	12888.79	1435.33	151.59	427.94	1823.19	551.61	211.80
5	44	14238.31	2093.93	295.46	397.54	1689.27	2108.62	366.42
6	26	11505.02	1428.31	176.97	332.38	1285.72	757.30	264.13

Cluster	Number of households	Average total consumer expenditures	Average consumer expenditures (per year and per person)					
			TRA	COM	CUL	EDU	RES	MGS
1	1266	6770.78	436.25	448.38	294.24	6.15	265.40	747.15
2	452	7497.08	544.18	444.28	369.95	8.88	400.65	688.39
3	1778	4785.46	222.05	273.33	175.56	2.38	164.29	321.03
4	1067	8382.81	927.52	478.70	464.61	43.18	881.65	985.69
5	44	23774.56	14251.76	402.56	453.44	3.66	547.84	1164.06
6	26	9700.24	1402.74	456.36	646.59	1208.70	847.47	893.57

Source: Household Budget Survey, own processing in SAS EG.

The interpretation of household segments based on Tables 1 and 2 is supplemented with demographic characteristics – household type and age of household members, whose relative structure in individual clusters is provided in the Appendix in Tables 5 to 12.

Table 2. Average consumer expenditures and average net household incomes in individual clusters (per year and per person) – year 2022

Cluster	Number of households	Net household income	Average consumer expenditures (per year and per person)					
			FAB	ABT	CLO	HSG	FUR	HLT
1	1253	13300.27	1544.09	138.66	423.51	1928.00	468.06	211.14
2	1137	11762.92	2558.02	590.36	259.68	2171.00	420.43	180.33
3	2260	10130.58	1684.33	112.12	183.85	1902.24	291.88	214.49
4	257	12685.36	2795.33	218.91	386.27	2772.23	993.14	943.11
5	67	14012.74	2565.76	251.98	755.57	1948.49	2837.26	332.84
6	17	12946.29	1533.19	207.33	401.95	1550.55	400.69	205.44

Cluster	Number of households	Average total consumer expenditures	Average consumer expenditures (per year and per person)					
			TRA	COM	CUL	EDU	TRA	MGS
1	1253	8910.00	956.02	550.96	498.38	37.06	1003.41	1150.72
2	1137	8680.61	587.59	516.98	300.53	5.89	327.70	762.10
3	2260	5767.30	327.35	325.52	177.59	1.99	148.50	397.42
4	257	11588.77	847.72	495.79	651.07	5.83	427.92	1051.45
5	67	24352.21	12164.86	477.65	617.16	13.07	926.38	1461.20
6	17	10726.67	1322.28	607.43	601.95	1982.37	942.62	970.87

Source: Household Budget Survey, own processing in SAS EG.

Four segments consistent in both 2021 and 2022:

1) Households with an Unhealthy Lifestyle
   Households in the second cluster allocate the highest average expenditures on alcoholic beverages, tobacco, and narcotics compared to households in other clusters. These households also have higher-than-average expenditures on food and non-alcoholic beverages, while their health-related expenditures are the lowest on average. In 2021, this cluster comprised 9.76% of households, whereas in 2022, it accounted for 22.78% of households. This cluster is predominantly made up of two-member households, primarily consisting of individuals aged 25 to 64 years.

2) Low-Income Households
   The values of nearly all average expenditures in households belonging to cluster number 3 are the lowest compared to other clusters, reflecting their low household income. The average annual net income per person in these households was €9,183.92 in 2021 and €10,130.58 in 2022, making them the households with the

lowest average net income among all clusters. In 2021, this cluster represented 38.38% of households, while in 2022, it grew to 45.28%, making it the largest cluster in both years. This cluster includes one-member and two-member households, primarily consisting of individuals aged 64 and older.

3) High-Income Households
Households in the fifth cluster allocate the highest average expenditures on food and non-alcoholic beverages, household furnishing, transportation, and miscellaneous goods and services compared to other clusters. Their expenditures on other categories are also relatively high. These households spend an exceptionally large amount on transportation, averaging €14,251.76 per person per year in 2021 and €12,164.86 per person per year in 2022. The extremely high transportation costs in these households are due to the purchase of a new motor vehicle. The average annual net income per person in this cluster was €14,238.31 in 2021 and €14,012.74 in 2022, making it the highest among all clusters. Less than 1% of households belonged to this cluster in 2021, increasing to 1.34% in 2022. This cluster predominantly consists of two-member households with individuals aged 25 to 64 but also includes households with dependent children.

4) Households with Children And Students
Households in the sixth cluster allocate the highest average expenditures on education compared to other clusters. They also have higher-than-average spending on recreation and culture, as well as on restaurants and hotels. Conversely, among all clusters, these households spend the least on essential needs – food and housing. Only 0.56% of households belonged to this cluster in 2021, decreasing to 0.34% in 2022. This cluster is predominantly composed of households with dependent children, with children represented across all age groups, and students aged 16 to 24 being the most prevalent.

Two distinct household groups in 2021:

1) Frugal Households (in Terms of Consumer Goods)
Households in the first cluster allocate the highest average expenditures on food and non-alcoholic beverages compared to other clusters, while spending the least on clothing and footwear. A total of 1 266 households were assigned to this cluster, representing 27.33% of all households in the sample. This cluster is predominantly composed of one- and two-member households with individuals over the age of 25, as well as households with dependent children aged 16 to 24.

2) Active Households
Households in the fourth cluster spend the most on clothing and footwear, housing and utilities, postal and telecommunication services, and restaurants and hotels compared to other clusters. Conversely, they allocate less money to food, non-alcoholic beverages, and alcoholic beverages. A total of 1,067 households were

assigned to this cluster, making up 23.03% of all households in the sample. This cluster mainly consists of households with dependent children of all age groups and two-member households.

Two distinct household groups in 2022:

1) Young Households
   Households in the first cluster spend significantly less on food and non-alcoholic beverages, as well as on alcoholic beverages, tobacco, and narcotics compared to other clusters. However, they allocate the highest expenditures on restaurants and hotels.
   A total of 1,253 households were assigned to this cluster, representing 25.11% of all households in the sample. This cluster is predominantly composed of two-member households with members aged 25–64 and households with dependent children.

2) Childless Households
   Households in the fourth cluster spend the most on food and non-alcoholic beverages, housing and utilities, healthcare, and recreation and culture compared to other clusters. However, they invest minimally in education. A total of 257 households were assigned to this cluster, making up 5.15% of all households in the sample. This cluster primarily consists of one- and two-member households, with members predominantly aged 25 to 64.

Table 3. Average consumer expenditures of households in individual clusters (in % per year and per person) – year 2021

Cluster	Number of households	Average consumer expenditures (in % per year and per person)					
		FAB	ABT	CLO	HSG	FUR	HLT
1	1266	30.5	2.4	2.5	22.3	5.2	4.8
2	452	25.1	11.4	4.1	19.6	4.9	2.2
3	1778	26.7	2.7	4.8	31.5	5.7	4.5
4	1067	17.1	1.8	5.1	21.7	6.6	2.5
5	44	8.8	1.2	1.7	7.1	8.9	1.5
6	26	14.7	1.8	3.4	13.3	7.8	2.7

Cluster	Number of households	Average consumer expenditures (in % per year and per person)					
		TRA	COM	CUL	EDU	RES	MGS
1	1266	6.4	6.6	4.3	0.1	3.9	11.0
2	452	7.3	5.9	4.9	0.1	5.3	9.2
3	1778	4.6	5.7	3.7	0.0	3.4	6.7
4	1067	11.1	5.7	5.5	0.5	10.5	11.9
5	44	59.9	1.7	1.9	0.0	2.4	4.9
6	26	14.5	4.7	6.7	12.5	8.7	9.2

Source: Household Budget Survey, own processing in SAS EG.

Considering expenditures on food and non-alcoholic beverages and housing and housing-related energy as essential human needs, households in clusters one, two, three, and four on average spent up to half of their resources on these essential needs. As a result, they had a limited budget for less essential items. If we exclude the extraordinary transportation expenses in the fifth cluster, which accounted for nearly 60% of average annual expenditures in 2021 and 50% in 2022, the expenditures on essential needs in this cluster would still represent a high portion of the budget, around 40%. The most evenly distributed expenditure structure is observed in households in the sixth cluster, where households spend "only" 28% of their resources on essential needs, with a higher portion spent on transportation and education (Tables 3 and 4).

Piekut and Knapkova (2025) analyzed the consumption behavior of households in Europe and, based on various classification methods, grouped countries according to the similarity of their consumer patterns. The study mentions Slovakia as a country that, according to the classification methods used, is grouped with countries exhibiting Western European consumer behavior, even though it is typically considered an Eastern European country. This suggests a shift in the consumer patterns of Slovak households closer to those of Western Europe.

Table 4. Average consumer expenditures of households in individual clusters (in % per year and per person) – year 2022

Cluster	Number of households	Average consumer expenditures (in % per year and per person)					
		FAB	ABT	CLO	HSG	FUR	HLT
1	1253	17.3	1.6	4.8	21.5	5.3	2.4
2	1137	29.5	6.7	3.0	25.0	4.7	2.1
3	2260	29.2	1.9	3.2	33.0	5.1	3.7
4	257	24.1	1.9	3.3	23.9	8.6	8.1
5	67	10.5	1.0	3.1	8.0	11.6	1.4
6	17	14.3	1.9	3.7	14.5	3.7	1.9
Cluster	Number of households	Average consumer expenditures (in % per year and per person)					
		TRA	COM	CUL	EDU	RES	MGS
1	1253	10.7	6.2	5.6	0.4	11.3	12.9
2	1137	6.8	6.0	3.5	0.1	3.8	8.8
3	2260	5.7	5.6	3.1	0.0	2.6	6.9
4	257	7.3	4.3	5.6	0.1	3.7	9.1
5	67	50.0	2.0	2.5	0.1	3.8	6.0
6	17	12.3	5.7	5.6	18.5	8.8	9.1

Source: Household Budget Survey, own processing in SAS EG.

However, on the other hand, although Slovakia shows similar consumer patterns to Western countries, challenges related to income disparity and living standards, which are characteristic of Eastern European countries, persist. The results of this study show that, although consumer patterns are starting to align, significant differences in the consumption of basic goods and services still exist between Eastern and Western European countries.

Morvay (2023) emphasizes the need for state support for households threatened by poverty or economic shocks. This perspective is undoubtedly justified, but the analysis of household consumption behavior suggests that the approach to social and economic policies should be more targeted. The results of our analysis show that even within the group of low-income households (cluster 3), there may be households where average income exceeds expenditures. This group is characterized by a consumption structure primarily focused on food and housing, yet they are still able to save. This suggests that not all low-income households necessarily need state support to the same extent.

On the other hand, in the case of high-income households (cluster 5), it can be observed that their average expenditures exceed their income. However, this trend is not necessarily a sign of financial instability but is influenced by one-time investments, such as the purchase of a car. This is a small proportion of households, which likely recurs at different times, but there is no reason to consider it a vulnerable group.

These findings suggest that for economic policy to be effective, it should not be solely based on income categories but should also consider expenditure patterns and household structures. Otherwise, there is a risk that support may be directed to households that do not actually need it, while some truly vulnerable groups may remain outside its reach.

State aid should be differentiated not only based on income but also on household expenditure patterns, as some low-income households are able to save, while high-income households may draw on their savings. Support should focus on essential expenditures, such as housing and food, for example, in the form of tax relief or energy subsidies. It is important to encourage households to engage in responsible financial behavior, such as promoting savings. The state should regularly analyze the effectiveness of assistance and adjust it to the specific types of households, as their needs vary – seniors are at risk from rising energy prices, while families with children face higher costs for care and education.

5. Conclusions

The results of the analysis of household consumption expenditures in Slovakia for 2021 and 2022 led to the identification of six household segments, four of which remained the same in both analyzed years, while two changed depending on the period.

The obtained segments reflect not only differences in household income levels but also their consumption behavior and spending preferences.

The selected combination of the years 2021 and 2022 made it possible to observe the stabilization of consumer behavior after the most critical phase of the pandemic, while these years provide a relevant framework for analyzing household behavior in the context of an ongoing but stabilized crisis. The year 2023 already reflects new structural challenges (e.g. inflation, energy crisis, geopolitical tensions), which will be the subject of further research.

The four stable segments of households in the analyzed years are households with an unhealthy lifestyle, low-income households, high-income households, and households with children and students. Households with an unhealthy lifestyle are characterized by a preference for spending on addictive substances at the expense of investing in health. Low-income households are characterized by the lowest level of consumption expenditures in almost all categories, and nearly half of the households in the analyzed sample fall into this category, highlighting the high proportion of the population with limited financial resources. High-income households are characterized by the highest expenditures on new household equipment and the purchase of transportation means. This group reflects higher purchasing power and a focus on investments in long-term assets. The consumption pattern of households with children and students highlights the importance of education and leisure activities.

For 2021, specific segments include frugal households, which are characterized by prudence in managing finances, and a preference for home-cooked meals over eating out, as well as active households focused on socialization and an active social lifestyle. For 2022, specific segments include young households, whose typical characteristic is a preference for eating out, reflecting the lifestyle change among young people who prioritize convenience and quick access to food, and childless households, whose consumption pattern indicates a focus on personal comfort, quality housing, and health care.

The analysis confirmed that regardless of income level, Slovak households spend about half of their total expenditures on essential needs, primarily food and housing. However, the absolute amount of these expenditures differs depending on income groups. Households with higher incomes have higher consumption in areas such as recreation, culture, dining at restaurants, hotels, transportation, and clothing, while low-income households primarily focus on covering basic needs. These results highlight significant differentiation in consumption behavior among various household segments and confirm that income level is one of the key factors influencing consumption patterns. Segmenting households based on expenditures provides a useful insight into the economic behavior of different population groups and can serve as an important tool for the development of socio-economic strategies and policies to support different income groups.

Acknowledgement

This paper is an output of the science project: VEGA no. 1/0285/24: *The Impact of Inflation on Poverty and Social Exclusion in Slovakia and the EU*.

References

Antonin, C., (2020). The links between saving rates, income and uncertainty: An analysis based on the 2011 household budget survey. *Economie et Statistique/ Economics and Statistics*, Vol. 513, pp. 47–68. https://doi.org/10.24187/ecostat.2019.513.2000.

Bouzarovski, S., Tirado Herrero, S., (2016). Geographies of injustice: The socio-spatial determinants of energy poverty in Poland, the Czech Republic and Hungary. *Post-Communist Economies*, Vol. 29(1), pp. 27–50. https://doi.org/10.1080/14631377.2016.1242257.

Cupák, A., Pokrivčák, J. and Rizov, M., (2015). Food Demand and Consumption Patterns in the New EU Member States: The Case of Slovakia. *Ekonomický časopis*, Vol. 63(4), pp. 339–358.

Değirmenci, T., Özbakır, L., (2017). Differentiating households to analyze consumption patterns: A data mining study on official household budget data. *WIREs Data Mining and Knowledge Discovery*, Vol. 8(1). https://doi.org/10.1002/widm.1227.

Department of Economic and Social Affairs of the United Nations, (2023). *Classification of Individual Consumption According to Purpose 2018*. United Nations.

Dogan, O., Kaya, G., Kaya, A. and Beyhan, H., (2019). Catastrophic household expenditure for healthcare in Turkey: Clustering Analysis of Categorical Data. *Data*, 4(3), p. 112. https://doi.org/10.3390/data4030112.

Eurostat, (2003). *Household budget surveys in the EU: Methodology and recommendations for harmonisation - 2003*. Office for Official Publications of the European Communities.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D., (2011). *Cluster analysis*. Wiley.

Froemelt, A., Dürrenmatt, D. J. and Hellweg, S., (2018). Using data mining to assess environmental impacts of household consumption behaviors. *Environmental Science & Technology*, Vol. 52(15), pp. 8467–8478. https://doi.org/10.1021/acs.est.8b01452.

Galvánková, V., (2022). Slováci nešetria a dlhujú príliš veľa. *Hospodárske noviny*, Vol. 147(7).

Jolliffe, I. T., (2002). *Principal component analysis*. Springer-Verlag.

Labudová, V., Vojtková, M. and Linda, B., (2010). Aplikácia viacrozmerných metód pri meraní chudoby. *E+M Ekonomie a Management*, Vol. 1, pp. 6–22.

Morvay, K., (2023). Masívny rast spotreby domácností uprostred krízového mixu. Ako to bolo možné? *Monitor hospodárskej politiky*, Vol. 1, pp. 19–22.

Morvay, K., (2005). *Transformácia ekonomiky: Skúsenosti Slovenska*. Ústav slovenskej ekonomiky SAV.

Piekut, M., Knapkova, M., (2025). Patterns and convergence in household spending: Insights from western and Eastern Europe. *Amfiteatru Economic*, Vol. 27(68), p. 180. https://doi.org/10.24818/ea/2025/68/180.

Vlačuha, R., Hornáček, M. and Vargová, A., (2023). *Výdavky súkromných domácností v SR 2022*. Štatistický úrad Slovenskej republiky.

Vojtková, M., Stankovičová, I., (2020). *Viacrozmerné štatistické metódy s aplikáciami v softvéri SAS*. Letra Edu.

Vopravil, J., Linhartová Jiřičková, B., (2024). Household surveys integration: Household budget survey methodology in Czechia. Statistika: *Statistics and Economy Journal,* Vol. 104(3), pp. 364–370. https://doi.org/10.54694/stat.2024.6.

Appendix

Table 5 Distribution of households by household type in individual clusters in % – year 2021

Cluster	Number of households	Household type					
		1	2	3	4	5	6
		1 adult	2 adults	More than 2 adults	1 adult with dependent children	2 adults with dependent children	More than 2 adults with dependent children
1	1266	24.6	34.7	12.2	2.5	18.2	7.8
2	452	19.5	45.6	10.6	2.0	15.9	6.4
3	1778	33.6	38.9	10.1	1.0	9.1	7.3
4	1067	21.7	26.3	10.0	5.9	28.2	7.9
5	44	11.4	61.4	2.3	2.3	18.2	4.4
6	26	11.5	11.5	3.9	23.1	42.3	7.7

Source: Household Budget Survey, own processing in SAS EG.

Table 6 Household structure by number of members aged 16 to 24 in % (2021)

Cluster	Number of households	Number of members aged 16 to 24					
		0	1	2	3	4	5
1	1266	81.9	12.4	4.8	0.7	0.0	0.2
2	452	83.6	14.4	1.3	0.4	0.0	0.2
3	1778	87.9	7.9	3.3	0.8	0.1	0.0
4	1067	76.9	18.1	4.5	0.5	0.0	0.1
5	44	90.9	9.1	0.0	0.0	0.0	0.0
6	26	61.5	34.6	3.8	0.0	0.0	0.0

Source: Household Budget Survey, own processing in SAS EG.

Table 7 Household structure by number of members aged 25 to 64 in % (2021)

Cluster	Number of households	Number of members aged 25 to 64						
		0	1	2	3	4	5	6
1	1266	25.1	28.4	37.9	6.8	1.4	0.3	0.1
2	452	19.5	25.2	46.0	7.3	2.0	0.0	0.0
3	1778	46.3	24.2	23.1	4.7	1.5	0.2	0.1
4	1067	10.2	29.6	49.8	8.0	2.3	0.1	0.0
5	44	18.2	20.5	56.8	4.5	0.0	0.0	0.0
6	26	0.0	34.6	61.5	3.8	0.0	0.0	0.0

Source: Household Budget Survey, own processing in SAS EG.

Table 8. Household structure by number of members aged over 64 in % (2021)

Cluster	Number of households	Number of members aged over 64			
		0	1	2	3
1	1266	58.1	24.7	16.9	0.2
2	452	69.7	16.8	13.5	0.0
3	1778	32.5	42.0	25.4	0.1
4	1067	80.4	14.0	5.6	0.0
5	44	70.5	18.2	11.4	0.0
6	26	92.3	3.8	3.8	0.0

Source: Household Budget Survey, own processing in SAS EG.

Table 9. Distribution of households by household type in individual clusters in % – year 2022

Cluster	Number of households	Household type					
		1	2	3	1	5	6
		1 adult	2 adults	More than 2 adults	1 adult with dependent children	2 adults with dependent children	More than 2 adults with dependent children
1	1253	19.4	26.5	9.7	6.6	29.5	8.3
2	1137	25.9	40.7	9.1	2.4	15.7	6.2
3	2260	35.3	38.1	9.9	1.6	9.9	5.2
4	257	35.4	46.7	5.8	0.8	10.1	1.2
5	67	13.4	53.7	3.0	1.5	23.9	4.5
6	17	11.7	5.9	5.9	11.8	52.9	11.8

Source: Household Budget Survey, own processing in SAS EG.

Table 10. Household structure by number of members aged 16 to 24 in % (2022)

Cluster	Number of households	Number of members aged 16 to 24					
		0	1	2	3	4	5
1	2260	89.7	6.7	2.8	0.7	0.0	0.0
2	1137	84.8	11.7	2.7	0.5	0.0	0.3
3	1253	73.7	20.2	5.4	0.6	0.0	0.0
4	257	93.8	5.8	0.4	0.0	0.0	0.0
5	67	85.1	13.4	1.5	0.0	0.0	0.0
6	17	52.9	35.3	11.8	0.0	0.0	0.0

Source: Household Budget Survey, own processing in SAS EG.

Table 11. Household structure by number of members aged 25 to 64 (2022) in %

Cluster	Number of households	Number of members aged 25 to 64						
		0	1	2	3	4	5	6
1	2260	48.0	25.3	21.6	3.8	0.9	0.2	0.0
2	1137	25.3	29.3	38.2	5.7	1.3	0.1	0.1
3	1253	8.1	29.1	54.0	6.9	1.8	0.1	0.0
4	257	39.3	26.8	30.0	3.9	0.0	0.0	0.0
5	67	14.9	16.4	64.2	4.5	0.0	0.0	0.0
6	17	0.0	23.5	70.6	5.9	0.0	0.0	0.0

Source: Household Budget Survey, own processing in SAS EG.

Table 12. Household structure by number of members aged over 64 (2022) in %

Cluster	Number of households	Number of members aged over 64			
		0	1	2	3
1	2260	32.5	41.5	25.8	0.1
2	1137	61.7	22.0	16.1	0.2
3	1253	81.2	14.0	4.8	0.0
4	257	50.2	28.4	20.6	0.8
5	67	76.1	16.4	7.5	0.0
6	17	94.1	5.9	0.0	0.0

Source: Household Budget Survey, own processing in SAS EG.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 91–110, https://doi.org/10.59139/stattrans-2026-018
Received – 17.03.2025; accepted – 14.11.2025

Bayesian sensitivity of insurance premium in collective risk model under bivariate prior with dependent frequency and severity of claims

Agata Boratyńska[1]

Abstract

This study deals with the problem of robustness of the collective and Bayes premiums under uncertainty of prior knowledge. The inaccuracy of the prior knowledge concerns the disturbance of independence between variables describing the frequency and average value of claims. Traditionally, these variables are independent, but in applications it is not always the case. Two classes of priors are presented: in the first class, the FGM copula is applied, while in the second one, the dependence between two contaminated priors is shown. In both classes, priors have the form of a linear combination of known bivariate probability distributions. The ranges of collective and Bayes premiums are calculated and prior and posterior regret gamma-minimax premiums are presented as the optimal premiums. Despite the very mild or small dependence, its influence on the premiums, especially on the bonus-malus factor, is relatively significant.

Key words: classes of priors, FGM copula, ε-contamination, posterior regret Γ-minimax premium, mean square error, bonus-malus factor.

1. Introduction

One of the most important aims in insurance is to estimate the premium. The premium is defined as a functional H which assigns to a real risk S a non-negative real number. In this paper a collective risk model is presented and the risk variable has a probability distribution function (p.d.f.) depending on a bivariate unknown parameter (λ, θ). The first parameter describes the expected value of the number of claims in one period (in the paper, one year), and the second one describes the expected value of the severity of a claim. There are many principles according to which premiums are calculated. Mamy of them are presented in Heilmann (1989), Gómez-Déniz et al. (1999), Boratyńska (2008), Furman and Zitikis (2008), Young (2004). We consider the net premium in the form $H(S) = H(\lambda, \theta) = a\lambda\theta$, where $a > 0$ is known. It is called the individual premium. The parameters λ and θ are unknown, so we ought to estimate H.

We will use the Bayesian methodology to combine the prior knowledge about parameters (defined by a prior distribution) with the knowledge about the history of the risk in the form of a random sample, where the probability distribution of this random variable depends on the parameters. The quality of an estimator is measured by the expected value of

[1]Warsaw School of Economics SGH, Collegium of Economic Analysis, Warsaw, Warsaw, Poland.
E-mail: aborata@sgh.waw.pl. ORCID: https://orcid.org/0000-0001-7363-1960.

a squared error loss function. Thus having some prior information about parameters, described by a prior distribution π (we will use the same notation for a probability distribution and its density (p.d.f.) with respect to the chosen measure on a probability space), and minimizing the expected squared error loss we obtain the collective premium $H^C(\pi)$ equal to the expected value of $H(\lambda,\theta)$ under the prior π. This premium is a premium in a class of risk, because the prior expresses the population behaviour of an unknown parameter. Introducing a random sample x with a p.d.f. dependent on the parameter (λ,θ) and minimizing the expected value of the squared error loss if the parameter has the posterior distribution, we calculate a Bayes premium $H^B(\pi,x)$ equal to the expected value of $H(\lambda,\theta)$ under the posterior distribution $\pi(\cdot|x)$. This premium combines knowledge about the population and about one considered risk (a policy).

The collective and Bayes premiums depend on a choice of a prior. The elicitation of a prior is difficult and can be uncertain. To model uncertainty of the prior information the robust Bayesian inference uses a class Γ of priors. In the literature, there are many different classes Γ of priors. For general references, see Berger (1994), Ríos Insua and Ruggeri (2000), Ruggeri et al. (2021). In insurance, robust Bayesian analysis has been considered in many papers, for example: Gómez-Déniz et al. (2002), Gómez-Déniz (2009), Peters et al. (2017), Sánchez-Sánchez et al. (2019), Boratyńska (2021, 2022), Ruggeri et al. (2025). For some recent papers applying robustness of Bayesian procedures in different fields see Tomer and Rai (2021), Ho (2023), Harrouche et al. (2025).

The main objective of this study is to examine how both collective and Bayes premiums respond to some uncertainty related to the independence between variables describing the frequency and the average severity of claims. We consider two classes of priors. The priors have the form of a linear combination of fixed priors. In the first class Γ_1, the Farlie-Gumbel-Morgenstern (FGM) copula with the fixed marginals is considered (for definition and properties of FGM see Nelsen (2006)). In the second class Γ_2, the marginal priors for both parameters are the mixtures of two fixed priors (they have the form of ε-contaminated priors, widely considered in the literature about robustness), but the variables have some degree of dependence.

Having a class of priors we have a set of collective premiums and a set of Bayes premiums. The sensitivity of the considered premium is measured by the range of the set, when priors run over the class Γ. If the range is small (according to the statisticians' or experts' judgements), then any prior can be chosen since all of them lead to similar results (see Berger (1994), Ríos Insua and Ruggeri (2000), among others). On the other hand, the practitioner faces a problem of choosing the optimal estimator. There are several concepts of optimal rules, for details see Ríos Insua and Ruggeri (2000), Hu and Xiao (2021), and references therein. As an optimal procedure we consider the prior and posterior regret Γ-minimax estimator (PRGM estimator). The selected optimal rules provide the estimators, which minimize the largest possible increase in risk resulting from making the wrong choice of a prior distribution. Their values depend on the bands of a set of the considered premiums calculated with respect to the priors belonging to the class Γ. Thus, computing a PRGM estimator is relatively simple.

In calculation of the insurance premium the number and severity of claims are typically assumed to be independent, but it is sometimes more reasonable to allow some dependence

(see Lemaire (1995), Gschlößl and Czado (2007), Shi et al. (2015), Lee et al. (2019), Lee and Shi (2019), Oh et al. (2020)). There are different tools to model dependence, for exampe: regression models, generalized linear models, copulas, but there are not many papers where the dependence is modelled using appropriate priors. For references, see Cheung et al. (2021). Some dependence between parameters λ and θ have been considered by Hernandez-Bastida et al. (2009), Cheung et al. (2021), Gomez-Deniz (2016), Ruggeri et al. (2025), among others, but without PRGM procedures.

The paper is organized as follows. In Section 2 the base model is presented and the measures of robustness and optimal premiums are defined. The classes of priors and the main results are in Section 3. In Section 4 the results are applied to a numerical example. Additionally, we illustrate the influence of uncertainty of the prior on the bonus-malus coefficient. Finally, in Section 5 we provide concluding remarks.

2. Bayesian collective risk model and measures of robustness

For a given contract (risk) let N_h be a random variable describing the number of claims in a year h, and $Y_{h1}, Y_{h2}, \ldots$ be random variables describing the severity of claims, and let $S_h = \sum_{i=1}^{N_h} Y_{hi}$ ($S_h = 0$ if $N_h = 0$) be an aggregate claim amount. Let λ and θ be two positive continuous random variables. We assume that given λ, the random variable N_h has the Poisson distribution $Poiss(\lambda)$, and given θ, random variables Y_{hi}, $i = 1, 2, \ldots$, are i.i.d. with the gamma distribution $Gamma(a; \frac{1}{\theta})$, where $a > 0$ is known, and gamma distribution $Gamma(\alpha; \beta)$ with parameters $\alpha, \beta > 0$ has the p.d.f. equal to

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

for $x > 0$. Given λ and θ, random variables N_h, Y_{hi}, $i = 1, 2, \ldots$, are independent.

Assume that λ has the prior distribution $\pi_{10} = Gamma(\mu_{10} k_{10}; k_{10})$ and θ has the prior distribution $\pi_{20} = IGamma(k_{20} + 1; \mu_{20} k_{20})$, where k_{i0}, μ_{i0}, for $i = 1, 2$, are fixed positive numbers, $k_{20} > 1$, and $IGamma(\gamma + 1; \delta)$ is an inverse gamma distribution with parameters $\gamma > 1$ and $\delta > 0$ and the p.d.f. equal to

$$\pi(\theta|\gamma,\delta) = \frac{\delta^{\gamma+1}}{\Gamma(\gamma+1)} \theta^{-\gamma-2} \exp\left(-\frac{\delta}{\theta}\right)$$

for $\theta > 0$. Suppose that λ and θ are independent. Thus, the bivariate random variable (λ, θ) has the prior p.d.f. $\pi_{00}(\lambda, \theta) = \pi_{10}(\lambda)\pi_{20}(\theta)$ with respect to the Lebesgue measure on the space $\Lambda \times \Theta = (0, +\infty) \times (0, +\infty)$. The parametrization of the prior gamma and inverse gamma distributions is chosen so that μ_{10} and μ_{20} (similarly, later in the paper, μ_{1i} and μ_{2j}) correspond to the prior expected values of the variables λ and θ. From an actuary's perspective, these parameters are relatively easy to estimate. Furthermore, in the definitions of the prior and posterior distributions, as well as in the parameters of these distributions throughout the paper, the subscript $1i$ indicates that the distribution refers to the variable λ, while the subscript $2j$ refers to the variable θ. The symbol μ with the corresponding indices always denotes the prior expected value of the respective distribution.

The p.d.f. π_{00} is the conjugate prior. Hence, given the data

$$(N^t, S^t) = (N_1, S_1, \ldots, N_t, S_t)$$

(the history of the contract for t years), the posterior distribution $\pi_{00}(\cdot|N^t, S^t)$ of (λ, θ) is a product of $Gamma(N + \mu_{10}k_{10}; t + k_{10})$ and $IGamma(Na + k_{20} + 1; S + k_{20}\mu_{20})$ distributions, where $N = \sum_{h=1}^{t} N_h$ and $S = \sum_{h=1}^{t} S_h$. Thus, given (N^t, S^t), r.vs. λ and θ are independent.

In this model the net individual premium is equal to $H(\lambda, \theta) = E(S_h|\lambda, \theta) = a\lambda\theta$. Under the squared error loss function the collective premium and the Bayes premium are equal to

$$H_0^C = H^C(\pi_{00}) = E_{\pi_{00}}(a\lambda\theta) = a\mu_{10}\mu_{20}, \tag{1}$$

$$H_0^B = H^B(\pi_{00}, N^t, S^t) = E_{\pi_{00}}(a\lambda\theta|N^t, S^t) = a\frac{N + \mu_{10}k_{10}}{t + k_{10}}\frac{S + k_{20}\mu_{20}}{Na + k_{20}}. \tag{2}$$

Assume that the prior knowledge is not enough to elicit one prior distribution. Therefore, consider a class Γ of priors on the space $\Lambda \times \Theta$, and let $H^C(\pi)$, $H^B(\pi, N^t, S^t)$ be the collective and the Bayes premium under the square error loss function and a prior $\pi \in \Gamma$. Then we consider the oscillations:

$$r(H^C, \Gamma) = \sup_{\pi\in\Gamma} H^C(\pi) - \inf_{\pi\in\Gamma} H^C(\pi),$$

$$r(H^B(\cdot, N^t, S^t), \Gamma) = \sup_{\pi\in\Gamma} H^B(\pi, N^t, S^t) - \inf_{\pi\in\Gamma} H^B(\pi, N^t, S^t),$$

as measures of robustness for the collective and the Bayes premiums. Besides the measure of range of the collective and Bayes premiums we would like to choose an optimal procedure and we decided on the regret gamma-minimax estimation. The prior and the posterior regret under the squared error loss of a decision d are respectively equal to

$$reg(\pi, d) = E_\pi(a\lambda\theta - d)^2 - E_\pi(a\lambda\theta - H^C(\pi))^2$$

and

$$reg(\pi, d|N^t, S^t) = E_\pi\left((a\lambda\theta - d)^2|N^t, S^t\right) - E_\pi\left((a\lambda\theta - H^B(\pi, N^t, S^t))^2|N^t, S^t\right).$$

They measure the loss of optimality of the risk if we choose an estimate d instead of the best estimate (in our problem collective and Bayes premiums), which minimizes the expected loss, prior and posterior, respectively. Now, the prior and posterior regret gamma-minimax premiums: $H_{PR}^C(\Gamma)$ and $H_{PR}^B(\Gamma, N^t, S^t)$, satisfy conditions:

$$\sup_{\pi\in\Gamma} reg(\pi, H_{PR}^C(\Gamma)) = \inf_d \sup_{\pi\in\Gamma} reg(\pi, d),$$

$$\sup_{\pi\in\Gamma} reg(\pi, H_{PR}^B(\Gamma, N^t, S^t)|N^t, S^t) = \inf_d \sup_{\pi\in\Gamma} reg(\pi, d|N^t, S^t).$$

This methodology is based on the idea that the optimal action minimizes the supremum

of the function over distributions in the class Γ. The prior and posterior regret gamma-minimax premiums are equal to:

$$H^C_{PR}(\Gamma) = \frac{1}{2}\left(\sup_{\pi\in\Gamma} H^C(\pi) + \inf_{\pi\in\Gamma} H^C(\pi)\right), \tag{3}$$

$$H^B_{PR}(\Gamma, N^t, S^t) = \frac{1}{2}\left(\sup_{\pi\in\Gamma} H^B(\pi, N^t, S^t) + \inf_{\pi\in\Gamma} H^B(\pi, N^t, S^t)\right) \tag{4}$$

(for details see Ríos Insua et al. (1995)).

3. Robustness of the collective and Bayes premiums and PRGM premiums

Two classes of priors are considered. The prior distributions belonging to the classes are linear combinations of the known distributions. Their p.d.fs have the form

$$\pi(\lambda, \theta) = \sum_{i=1}^{r}\sum_{j=1}^{s} \alpha_{ij}\pi_{1i}(\lambda)\pi_{2j}(\theta),$$

where $\pi_{1i} = Gamma(\alpha_i;\beta_i)$ and $\pi_{2j} = IGamma(\gamma_j+1;\delta_j)$, where α_i, β_i, γ_j, δ_j are positive numbers, $\gamma_j > 1$, and α_{ij} are fixed numbers, such that π is a valid bivariate p.d.f. The following lemma will be useful to calculate posterior distributions and premiums.

Lemma 1. (see also Cheung et al. (2021)) *If $\pi(\lambda,\theta) = \sum_{i=1}^{r}\sum_{j=1}^{s}\alpha_{ij}\pi_{1i}(\lambda)\pi_{2j}(\theta)$ then, given (N^t, S^t), the posterior distribution is equal to*

$$\pi(\lambda, \theta|N^t, S^t) = \frac{1}{A}\sum_{i=1}^{r}\sum_{j=1}^{s} \alpha^*_{ij}\pi_{1i}(\lambda|N^t, S^t)\pi_{2j}(\theta|N^t, S^t), \tag{5}$$

where the posterior p.d.f. $\pi_{1i}(\lambda|N^t, S^t)$ is the p.d.f. of the distribution $Gamma(N+\alpha_i; t+\beta_i)$ and $\pi_{2j}(\theta|N^t, S^t)$ is the p.d.f. of the distribution $IGamma(Na+\gamma_j+1; S+\delta_j)$, and

$$\alpha^*_{ij} = \alpha_{ij}\frac{\beta_i^{\alpha_i}\Gamma(N+\alpha_i)}{\Gamma(\alpha_i)(\beta_i+t)^{\alpha_i+N}} \cdot \frac{\delta_j^{\gamma_j+1}\Gamma(aN+\gamma_j+1)}{(S+\delta_j)^{\gamma_j+aN+1}\Gamma(\gamma_j+1)}, \quad A = \sum_{i=1}^{r}\sum_{j=1}^{s}\alpha^*_{ij}. \tag{6}$$

Under the prior π the collective and Bayes premiums are equal to

$$H^C(\pi) = aE_\pi(\lambda\theta) = a\sum_{i=1}^{r}\sum_{j=1}^{s}\alpha_{ij}\frac{\alpha_i\delta_j}{\beta_i\gamma_j},$$

$$H^B(\pi, N^t, S^t) = a\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\alpha^*_{ij}}{A}\frac{(N+\alpha_i)(S+\delta_j)}{(t+\beta_i)(Na+\gamma_j)}.$$

■

3.1. The first class

Consider the FGM copula defined by $C(u,v) = uv + \omega uv(1-u)(1-v)$, for $u,v \in (0,1)$, $\omega \in [-1,1]$, and its derivative

$$c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v} = 1 + \omega(1-2u)(1-2v).$$

Let $k_{10}\mu_{10}$ and $k_{20} > 1$ be positive integers. Consider the following class of priors:

$$\Gamma_1 = \{\pi_\omega : \ \pi_\omega(\lambda,\theta) = c\left(\Pi_{10}(\lambda), \Pi_{20}(\theta)\right)\pi_{10}(\lambda)\pi_{20}(\theta) : \omega \in [\omega_1, \omega_2]\},$$

where $\Pi = 1 - \bar{\Pi}$ is the cumulative distribution function (c.d.f.) for the p.d.f. π and $\omega_1 < \omega_2$ and $\omega_1, \omega_2 \in [-1,\ 1]$ are fixed.

The priors belonging to the class Γ_1 have fixed marginal priors equal to π_{10} and π_{20} and $\pi_{\omega=0}(\lambda,\theta) = \pi_{00}(\lambda,\theta) = \pi_{10}(\lambda)\pi_{20}(\theta)$. The parameter ω is the dependence parameter and the Kendall and Spearman coefficients are equal to $\tau(\omega,\lambda,\theta) = \frac{2\omega}{9}$ and $\rho(\omega,\lambda,\theta) = \frac{\omega}{3}$ (see Nelsen (2006)). Hence, the mild dependence between λ and θ is assumed.

Lemma 2. *If π_{10} is Gamma$(\mu_{10}k_{10}; k_{10})$ and π_{20} is IGamma$(k_{20}+1; \mu_{20}k_{20})$, where $k_{10}\mu_{10}$ and k_{20} are positive integers, then the p.d.f. π_ω is equal to*

$$\pi_\omega(\lambda,\theta) = \pi_{10}(\lambda)\pi_{20}(\theta) + \omega \sum_{i=0}^{l_1}\sum_{j=0}^{l_2} \alpha_{ij}\pi_{1i}^*(\lambda)\pi_{2j}^*(\theta), \tag{7}$$

where $l_1 = k_{10}\mu_{10}$, $l_2 = k_{20}+1$, $\alpha_{00} = -1$, $\pi_{10}^* = \pi_{10}$, $\pi_{20}^* = \pi_{20}$,

$$\pi_{1i}^* = Gamma(i + l_1 - 1; 2k_{10}), \quad \pi_{2j}^* = IGamma(j + l_2 - 1; 2k_{20}\mu_{20}),$$

$$\alpha_{i0} = \frac{2}{2^{i+l_1-1}}\binom{i+l_1-2}{i-1}, \quad \alpha_{0j} = \frac{2}{2^{j+l_2-1}}\binom{j+l_2-2}{j-1}, \quad \alpha_{ij} = -\alpha_{i0}\alpha_{0j},$$

for $i = 1,\ldots,l_1$, $j = 1,\ldots,l_2$.

Proof. Denote $l_1 = k_{10}\mu_{10}$ and $l_2 = k_{20}+1$. Similarly to the article Cheung et al. (2021), we present the survival function of π_{10} and the c.d.f. of π_{20} for $\lambda,\ \theta > 0$ as follows:

$$\bar{\Pi}_{10}(\lambda) = \int_\lambda^{+\infty} \frac{k_{10}^{l_1}}{(l_1-1)!}u^{l_1-1}\exp(-k_{10}u)\,du = \sum_{n=0}^{l_1-1}\frac{k_{10}^n}{n!}\lambda^n\exp(-k_{10}\lambda),$$

$$\Pi_{20}(\theta) = \int_0^\theta \frac{(k_{20}\mu_{20})^{l_2}}{(l_2-1)!u^{l_2+1}}\exp\left(\frac{-k_{20}\mu_{20}}{u}\right)du$$

$$= \sum_{n=0}^{l_2-1}\frac{(k_{20}\mu_{20})^n}{n!}\theta^{-n}\exp\left(\frac{-k_{20}\mu_{20}}{\theta}\right).$$

The p.d.f π_ω is equal to

$$\pi_\omega(\lambda,\theta) = \left(1 + \omega(2\bar{\Pi}_{10}(\lambda) - 1)(1 - 2\Pi_{20}(\theta))\right)\pi_{10}(\lambda)\pi_{20}(\theta)$$

$$= \left(1+\omega\left(-1+2\bar{\Pi}_{10}(\lambda)+2\Pi_{20}(\theta)-4\bar{\Pi}_{10}(\lambda)\Pi_{20}(\theta)\right)\right)\pi_{10}(\lambda)\pi_{20}(\theta).$$

By substituting $\bar{\Pi}_{10}(\lambda)$ and $\Pi_{20}(\theta)$ we obtain the assertion. ■

All prior distributions from the class Γ_1 have the form given in Lemma 1. Applying Lemma 1 and Lemma 2 we have

$$E_{\pi_\omega}(\lambda\theta) = \mu_{10}\mu_{20} + \omega\sum_{i=0}^{l_1}\sum_{j=0}^{l_2}\alpha_{ij}\mu_{1i}^*\mu_{2j}^*,$$

$$\mu_{10}^* = \mu_{10}, \quad \mu_{1i}^* = \frac{i+k_{10}\mu_{10}-1}{2k_{10}}, \quad \mu_{20}^* = \mu_{20}, \quad \mu_{2j}^* = \frac{2k_{20}\mu_{20}}{j+k_{20}-1},$$

for $i = 1,\ldots,l_1$ and $j = 1,\ldots,l_2$. Thus, assuming $l_2 > 2$, the Pearson correlation coefficient is equal to

$$Corr_{\pi_\omega}(\lambda,\theta) = \omega\sum_{i=0}^{l_1}\sum_{j=0}^{l_2}\alpha_{ij}\mu_{1i}^*\mu_{2j}^*\frac{\sqrt{l_1(l_2-2)}}{\mu_{20}\mu_{10}} \tag{8}$$

and the collective premium and $Corr_{\pi_\omega}(\lambda,\theta)$ are the linear monotone functions of ω.

Given (N^t,S^t) and applying (7), (5), (6), we obtain the posterior p.d.f. and the Bayes premium

$$\pi_\omega(\lambda,\theta|N^t,S^t)$$

$$= \frac{-\alpha_{00}^*}{A^*}\pi_{10}(\lambda|N^t,S^t)\pi_{20}(\theta|N^t,S^t) + \omega\sum_{i=0}^{l_1}\sum_{j=0}^{l_2}\frac{\alpha_{ij}^*}{A^*}\pi_{1i}^*(\lambda|N^t,S^t)\pi_{2j}^*(\theta|N^t,S^t),$$

$$H^B(\pi_\omega,N^t,S^t)) = aE_{\pi_\omega}(\lambda\theta|N^t,S^t), \tag{9}$$

where the posterior p.d.fs. $\pi_{1i}^*(\lambda|N^t,S^t)$ and $\pi_{2j}^*(\theta|N^t,S^t)$ are the p.d.fs. of the distributions $Gamma(N+i+l_1-1;t+2k_{10})$ and $IGamma(Na+j+l_2-1;S+2k_{20}\mu_{20})$, for $i = 1,\ldots,l_1$ and $j = 1,\ldots,l_2$, respectively, and

$$\pi_{10}^*(\lambda|N^t,S^t) = \pi_{10}(\lambda|N^t,S^t), \quad \pi_{20}^*(\theta|N^t,S^t) = \pi_{20}(\theta|N^t,S^t),$$

$$\alpha_{00}^* = \frac{-k_{10}^{l_1}\Gamma(N+l_1)}{\Gamma(l_1)(k_{10}+t)^{l_1+N}}\cdot\frac{(\mu_{20}k_{20})^{l_2}\Gamma(aN+l_2)}{(S+\mu_{20}k_{20})^{l_2+aN}\Gamma(l_2)},$$

$$\alpha_{i0}^* = \frac{2}{(i-1)!}\frac{k_{10}^{i+l_1-1}\Gamma(N+i+l_1-1)}{\Gamma(l_1)(2k_{10}+t)^{N+i+l_1-1}}\cdot\frac{(\mu_{20}k_{20})^{l_2}\Gamma(aN+l_2)}{(S+\mu_{20}k_{20})^{l_2+aN}\Gamma(l_2)},$$

$$\alpha_{0j}^* = \frac{2}{(j-1)!}\frac{k_{10}^{l_1}\Gamma(N+l_1)}{\Gamma(l_1)(k_{10}+t)^{l_1+N}}\cdot\frac{(\mu_{20}k_{20})^{j+l_2-1}\Gamma(aN+j+l_2-1)}{(S+2\mu_{20}k_{20})^{j+l_2+aN-1}\Gamma(l_2)},$$

$$\alpha_{ij}^* = \frac{\alpha_{i0}^*\alpha_{0j}^*}{\alpha_{00}^*}, \quad A^* = -\alpha_{00}^* + \omega\sum_{i=0}^{l_1}\sum_{j=0}^{l_2}\alpha_{ij}^*, \tag{10}$$

for $i = 1,\ldots,l_1$, $j = 1,\ldots,l_2$. The Bayes premium is a homographic function of ω (see (9) and (10)), therefore, the supremum and infimum are achieved at the boundary of the interval $[\omega_1,\omega_2]$. Hence, applying (3) and (4), we obtain the theorem.

Theorem 1. *If the class of priors is equal to Γ_1 then the oscillations of the collective premium and the Bayes premium are equal to*

$$r(H^C,\Gamma_1) = a(\omega_2-\omega_1)\left|\sum_{i=0}^{l_1}\sum_{j=0}^{l_2}\alpha_{ij}\mu_{1i}^*\mu_{2j}^*\right|,$$

$$r(H^B(\cdot,N^t,S^t),\Gamma_1) = \left|H^B(\pi_{\omega=\omega_1},N^t,S^t) - H^B(\pi_{\omega=\omega_2},N^t,S^t)\right|.$$

The prior regret Γ-minimax and the posterior regret Γ-minimax premiums are equal to

$$H_{PR}^C(\Gamma_1) = a\mu_{10}\mu_{20} + \frac{\omega_1+\omega_2}{2}\sum_{i=0}^{l_1}\sum_{j=0}^{l_2}a\alpha_{ij}\mu_{1i}^*\mu_{2j}^*,$$

$$H_{PR}^B(\Gamma_1,N^t,S^t) = \frac{1}{2}\left(H^B(\pi_{\omega=\omega_1},N^t,S^t) + H^B(\pi_{\omega=\omega_2},N^t,S^t)\right). \qquad \blacksquare$$

Note that if $\omega_1 = -\omega_2$ then $H_{PR}^C(\Gamma_1)$ is equal to the collective premium for the basic prior (the prior, where λ and θ are independent). The relative range equal to $\frac{r(H^C,\Gamma_1)}{H_0^C}$ does not depend on μ_{20} and μ_{10} as long as $k_{10}\mu_{10}$ and k_{20} are fixed.

3.2. The second class

Assume two contaminated priors:

$$\pi^{\varepsilon}(\lambda) = (1-\varepsilon)\pi_{10}(\lambda) + \varepsilon\pi_{11}(\lambda), \quad \pi^{\eta}(\theta) = (1-\eta)\pi_{20}(\theta) + \eta\pi_{21}(\theta),$$

where $\pi_{1i} = Gamma(k_{1i}\mu_{1i};k_{1i})$ and $\pi_{2i} = IGamma(k_{2i}+1;k_{2i}\mu_{2i})$, for $i=0,1$, and $\varepsilon,\eta \in (0,1/2]$, $k_{1i}>0, k_{2i}>1$, $\mu_{1i},\mu_{2i}>0$ are fixed numbers. The elicited priors for both variables are a mixture of two probability distributions. Then the product of the measures on the product space $\Lambda\times\Theta$ has the form given in Lemma 1, namely

$$\pi^{\varepsilon,\eta}(\lambda,\theta) = (1-\varepsilon)(1-\eta)\pi_{10}(\lambda)\pi_{20}(\theta)$$

$$+\varepsilon(1-\eta)\pi_{11}(\lambda)\pi_{20}(\theta) + (1-\varepsilon)\eta\pi_{10}(\lambda)\pi_{21}(\theta) + \varepsilon\eta\pi_{11}(\lambda)\pi_{21}(\theta),$$

and the random variables λ and θ are independent. Now, consider the class

$$\Gamma_2 = \left\{\pi_\tau :\ \pi_\tau(\lambda,\theta) = \sum_{i=0}^{1}\sum_{j=0}^{1}\alpha_{ij}(\tau)\pi_{1i}(\lambda)\pi_{2j}(\theta),\ \tau\in[0,\min\{\varepsilon,\eta\}]\right\},$$

where $\alpha_{00}(\tau) = 1-\varepsilon-\eta+\tau,\ \alpha_{01}(\tau) = \eta-\tau,\ \alpha_{10}(\tau) = \varepsilon-\tau,\ \alpha_{11}(\tau) = \tau$.

The class Γ_2 contains the priors with the marginals equal to π^{ε} and π^{η}, but the variables λ and θ can be dependent. The independence is achieved if and only if $\tau = \varepsilon\eta$.

If the prior is π_τ then

$$E_{\pi_\tau}(\lambda\theta) =$$

$$\tau(\mu_{11}-\mu_{10})(\mu_{21}-\mu_{20}) + \varepsilon\mu_{20}(\mu_{11}-\mu_{10}) + \eta\mu_{10}(\mu_{21}-\mu_{20}) + \mu_{10}\mu_{20} \qquad (11)$$

and the covariance $Cov_{\pi_\tau}(\lambda,\theta)$ is equal to

$$Cov_{\pi_\tau}(\lambda,\theta) = (\tau-\varepsilon\eta)(\mu_{11}-\mu_{10})(\mu_{21}-\mu_{20}). \tag{12}$$

Hence, $Cov_{\pi_\tau}(\lambda,\theta)=0$ if $\tau=\varepsilon\eta$ (λ and θ are independent) or $\mu_{11}=\mu_{10}$ or $\mu_{21}=\mu_{20}$. Therefore, in the class Γ_2, there can be distributions describing the dependent variables λ and θ, but with $Cov(\lambda,\theta)=0$. For $(\mu_{11}-\mu_{10})(\mu_{21}-\mu_{20})>(<)0$ the covariance is a increasing (decreasing) function of τ and is negative (positive) for $\tau<\varepsilon\eta$. The behavior of the collective premium is presented in the following theorem.

Theorem 2. *If Γ_2 is the class of priors, then the oscillation of the collective premium is equal to*

$$r(H^C,\Gamma_2) = a\min\{\varepsilon,\eta\}|(\mu_{11}-\mu_{10})(\mu_{21}-\mu_{20})|,$$

and the optimal collective premium is

$$H^C_{PR}(\Gamma_2) = H^C\left(\pi_{\tau=\frac{\min\{\varepsilon,\eta\}}{2}}\right).$$

If $(\mu_{11}-\mu_{10})(\mu_{21}-\mu_{20})=0$ then $H^C(\pi_\tau)=H^C(\pi^{\varepsilon,\eta})$ does not depend on τ.

Proof. The collective premium under the prior π_τ is equal to $H^C(\pi_\tau)=aE_{\pi_\tau}(\lambda\theta)$, and it is a linear function of $\tau\in[0,\min\{\varepsilon,\eta\}]$ with the slope $(\mu_{11}-\mu_{10})(\mu_{21}-\mu_{20})$ (see (11)). Hence, applying the properties of a linear function, we obtain $H^C_{PR}(\Gamma_2)$ and $r(H^C,\Gamma_2)$. ■

Note that $H^C_{PR}(\Gamma_2)$ is equal to the collective premium for independent λ and θ if $\varepsilon=\frac{1}{2}$ or $\eta=\frac{1}{2}$, but regardless of the values of ε and η, $H^C_{PR}(\Gamma_2)$ is a collective premium under a certain prior belonging to Γ_2. For every $\varepsilon,\eta<\frac{1}{2}$ we have $\frac{1}{2}\min\{\varepsilon,\eta\}>\varepsilon\eta$. Thus, if $(\mu_{11}-\mu_{10})(\mu_{21}-\mu_{20})>(<)0$, then premium $H^C_{PR}(\Gamma_2)$ is greater (less) than $H^C(\pi_{\varepsilon\eta})$. If $\mu_{10}=\mu_{11}$ and $\mu_{20}=\mu_{21}$, then, comparing (1) and (11), $H^C(\pi_\tau)=H^C_0$ for every $\tau\in[0,\min\{\varepsilon,\eta\}]$.

Given the data (N^t,S^t) and the prior π_τ and applying Lemma 1 (see (5) and (6)) we obtain the posterior p.d.f.

$$\pi_\tau(\lambda,\theta|N^t,S^t) = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{\alpha^*_{ij}(\tau)}{A(\tau)}\pi_{1i}(\lambda|N^t,S^t)\pi_{2j}(\theta|N^t,S^t),$$

where $\pi_{1i}(\lambda|N^t,S^t)$ is the p.d.f. of the distribution $Gamma(N+\mu_{1i}k_{1i};t+k_{1i})$ and $\pi_{2j}(\theta|N^t,S^t)$ is the p.d.f. of the distribution $IGamma(Na+k_{2j}+1;S+k_{2j}\mu_{2j})$, and

$$\alpha^*_{ij}(\tau)=\alpha_{ij}(\tau)\frac{k_{1i}^{\mu_{1i}k_{1i}}\Gamma(N+\mu_{1i}k_{1i})}{\Gamma(\mu_{1i}k_{1i})(k_{1i}+t)^{\mu_{1i}k_{1i}+N}}\cdot\frac{(\mu_{2j}k_{2j})^{k_{2j}+1}\Gamma(aN+k_{2j}+1)}{(S+\mu_{2j}k_{2j})^{k_{2j}+aN+1}\Gamma(k_{2j}+1)},$$

and $A(\tau)=\sum_{i=0}^{1}\sum_{j=0}^{1}\alpha^*_{ij}(\tau)$. Hence, the Bayes premium is equal to

$$H^B(\pi_\tau,N^t,S^t) = a\sum_{i=0}^{1}\sum_{j=0}^{1}\frac{\alpha^*_{ij}(\tau)}{A(\tau)}\frac{N+\mu_{1i}k_{1i}}{t+k_{1i}}\frac{S+k_{2j}\mu_{2j}}{Na+k_{2j}},$$

and it is a homographic function of the variable τ, therefore the supremum and infimum are achieved at the boundary of the interval $[0, \min\{\varepsilon, \eta\}]$. Hence, we obtain the following theorem.

Theorem 3. *Assuming Γ_2 class of priors and the history of claim number and claim amount in the past t years N^t, S^t, the range of the Bayes premium is equal to*

$$r(H^B(\cdot, N^t, S^t), \Gamma_2) = |H^B(\pi_{\tau=\min\{\varepsilon,\eta\}}, N^t, S^t) - H^B(\pi_{\tau=0}, N^t, S^t)|$$

and the posterior regret Γ-minimax premium is equal to

$$H^B_{PR}(\Gamma_2, N^t, S^t) = \frac{1}{2}\left(H^B(\pi_{\tau=\min\{\varepsilon,\eta\}}, N^t, S^t) + H^B(\pi_{\tau=0}, N^t, S^t)\right). \qquad \blacksquare$$

Note that regardless of the value of the product $(\mu_{11} - \mu_{10})(\mu_{21} - \mu_{20})$, the Bayes premium depends on the value of the parameter τ.

4. Numerical example – sensitivity of bonus-malus system

The bonus-malus system (BM) adjusts the insurance premium based on the policyholder's claims history. With a good history (no claims), it lowers the premium, while with claims, it increases it. The BM coefficient takes into account the ratio of the Bayes premium (dependent on claims history) to the collective premium (considered as the base premium). The BM coefficient shows what percentage of the basic collective premium is represented by the Bayes premium for a given claims scenario. It often considers the number of claims. In the following example, besides the range of Bayes and collective premiums and PRGM premiums, we will consider the BM coefficient based on the number and severity of claims. We define the BM coefficient in the basic model as

$$BM(\pi_{00}, N^t, S^t) = \frac{H^B(\pi_{00}, N^t, S^t)}{H^C(\pi_{00})}.$$

When the prior distribution belongs to a class Γ of prior distributions, we consider the coefficients:

$$BM_{min} = \inf_{\pi\in\Gamma} \frac{H^B(\pi, N^t, S^t)}{H^C(\pi_{00})} \quad \text{and} \quad BM_{max} = \sup_{\pi\in\Gamma} \frac{H^B(\pi, N^t, S^t)}{H^C(\pi_{00})}$$

and compare them with the value

$$BM_{PR}(\Gamma, N^t, S^t) = \frac{H^B_{PR}(\Gamma, N^t, S^t)}{H^C_{PR}(\Gamma)},$$

which corresponds to the situation when, for the considered class of prior distributions, we use the prior and posterior regret gamma-minimax premiums.

Consider two base models. In the first one (M1), assume $k_{10} = 2.5$ and $\mu_{10} = 0.4$, in the second model (M2), $k_{10} = \mu_{10} = 1$. Thus, we assume greater expected frequency in the model M2 (the similar prior was considered in Rugger et al. (2025)). In both models $k_{20} = 2$ and $\mu_{20} = 200$. Consider the class Γ_1 with $\omega_1 = -\omega_2 = -1$. The behavior of the collective premium is presented in Table 1. The last two columns show the relative sensitivity of H^C, namely

$$R_{min} = \frac{\inf_{\pi\in\Gamma_1} H^C(\pi)}{H^C(\pi_{00})}, \qquad R_{max} = \frac{\sup_{\pi\in\Gamma_1} H^C(\pi)}{H^C(\pi_{00})}.$$

The collective premium is sensitive to the assumption of independence between frequency and severity of claims. In the class Γ_1 the Pearson correlation (8) is in the interval $[-0.1875, 0.1875]$, the Kendall and Spearman coefficients are in intervals $[-\frac{2}{9}, \frac{2}{9}]$ and $[-\frac{1}{3}, \frac{1}{3}]$ respectively, but the collective premium is in the interval $[0.81H^C(\pi_{00}), 1.19H^C(\pi_{00})]$.

Table 1. Sensitivity of the collective premium in models M1 and M2 with the class Γ_1 of priors

Model	$H^C(\pi_{00})$	$Var_{\pi_{00}}(\lambda)$	$Var_{\pi_{00}}(\theta)$	$r(H^C, \Gamma_1)$	R_{min}	R_{max}
M1	80	0.16	40000	30	0.81	1.19
M2	200	1	40000	75	0.81	1.19

Now, consider the three scenarios of the policyholder's behavior. In all scenarios, $t \in \{1,3,5,10\}$, $N \in \{0,1,\ldots,6\}$, but they differ in the average claim. We have $\frac{S}{N} = 100$ in the first scenario (S1), $\frac{S}{N} = 200$ in the second one (S2) and $\frac{S}{N} = 400$ in the third one (S3). Thus, S1 profile suggests lower average severity, the S3 profile implies higher average severity than the expected prior base severity.

Assuming the class Γ_1 of priors, for the presented data, the PRGM premium H^B_{PR} differs from H^B_0 by no more than 10% (see Table 2 and 3). The smallest difference is in S2 profile. However, the range of the Bayes premium in model M1 is often greater then 30% of H^B_0 and generally it is a decreasing function of t.

Table 2. Sensitivity of the Bayes premium in model M1 with the class Γ_1 of priors, $\frac{r}{H_0^B} = \frac{r(H^B,\Gamma_1)}{H_0^B}$

	$S/N = 100$			$S/N = 200$			$S/N = 400$		
N	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$
					$t = 1$				
0	57.1	57.1	0.240	57.1	57.1	0.240	57.1	57.1	0.240
1	95.2	95.7	0.310	114.3	113.7	0.384	152.4	148.2	0.476
2	128.6	130.6	0.333	171.4	168.6	0.387	257.1	239.7	0.440
3	160.0	164.3	0.333	228.6	223.1	0.356	365.7	333.6	0.385
4	190.5	197.0	0.322	285.7	277.8	0.319	476.2	431.6	0.337
5	220.4	229.0	0.307	342.9	332.9	0.286	587.8	533.6	0.295
6	250.0	260.3	0.289	400.0	388.5	0.257	700.0	638.9	0.261
					$t = 3$				
0	36.4	36.4	0.041	36.4	36.4	0.041	36.4	36.4	0.041
1	60.6	60.6	0.155	72.7	72.7	0.246	97.0	96.6	0.367
2	81.8	82.3	0.211	109.1	108.3	0.301	163.6	158.0	0.385
3	101.8	103.2	0.237	145.5	143.3	0.305	232.7	219.1	0.357
4	121.2	123.8	0.247	181.8	178.1	0.290	303.0	281.1	0.325
5	140.3	144.0	0.249	218.2	213.1	0.270	374.0	344.5	0.296
6	159.1	164.1	0.245	254.5	248.2	0.249	445.5	409.6	0.269
					$t = 5$				
0	26.7	26.7	0.094	26.7	26.7	0.094	26.7	26.7	0.094
1	44.4	44.4	0.041	53.3	53.4	0.136	71.1	71.5	0.272
2	60.0	60.1	0.115	80.0	79.8	0.221	120.0	118.4	0.327
3	74.7	75.1	0.156	106.7	105.8	0.248	170.7	164.7	0.319
4	88.9	89.9	0.179	133.3	131.6	0.251	222.2	211.0	0.299
5	102.9	104.6	0.191	160.0	157.2	0.244	274.3	257.7	0.278
6	116.7	119.2	0.197	186.7	182.9	0.231	326.7	305.0	0.258

Now, see Figures 1 and 2. The bonus-malus coefficient depends on the profile, the range of BM is an increasing function of the number of claims and the average severity of claims. For model M1, if $N/t > 1$ then, only for the S1 profile, the range of BM is less than 0.5. For example, if a policyholder has one accident in a given year, then in model M1 under scenario S3 and with the prior class Γ_1, the BM falls within the interval $(1.4, 2.3)$. This implies that the Bayes premium lies within the interval $(112, 184)$. For an actuary, this is not just a premium — it also reflects the variability in the prediction of the total claims for the following year. For model M2 the range and values of BM coefficient are smaller than for model M1. Here, the base prior also has a significant influence: the collective premium in M1 is lower than in model M2, which may reflect a portfolio characterized by a high accident rate. Notably, in scenario S3, the oscillation of the Bayes premium exceeds 50% of $H^C(\pi_{00})$.

Table 3. Sensitivity of Bayes premium in model M2 with the class Γ_1 of priors

	$S/N = 100$			$S/N = 200$			$S/N = 400$		
N	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$
				$t = 1$					
0	100.0	100.0	0.083	100.0	100.0	0.083	100.0	100.0	0.083
1	166.7	166.8	0.189	200.0	199.7	0.278	266.7	264.6	0.393
2	225.0	226.8	0.239	300.0	297.2	0.322	450.0	431.2	0.399
3	280.0	284.6	0.260	400.0	393.2	0.318	640.0	597.9	0.366
4	333.3	341.3	0.266	500.0	488.9	0.299	833.3	768.1	0.330
5	385.7	397.2	0.264	600.0	585.0	0.275	1028.6	943.1	0.298
6	437.5	452.3	0.257	700.0	681.7	0.252	1225.0	1123.1	0.269
				$t = 3$					
0	50.0	50.0	0.210	50.0	50.0	0.210	50.0	50.0	0.210
1	83.3	83.4	0.063	100.0	100.0	0.032	133.3	134.5	0.176
2	112.5	112.5	0.024	150.0	150.0	0.137	225.0	225.4	0.264
3	140.0	140.2	0.077	200.0	199.4	0.184	320.0	315.3	0.274
4	166.7	167.4	0.110	250.0	248.2	0.202	416.7	404.5	0.264
5	192.9	194.5	0.130	300.0	296.8	0.206	514.3	493.5	0.251
6	218.8	221.5	0.143	350.0	345.1	0.204	612.5	583.0	0.237
				$t = 5$					
0	33.3	33.3	0.352	33.3	33.3	0.352	33.3	33.3	0.352
1	55.6	55.8	0.196	66.7	66.5	0.106	88.9	89.2	0.039
2	75.0	75.2	0.097	100.0	100.0	0.016	150.0	151.6	0.164
3	93.3	93.4	0.033	133.3	133.4	0.082	213.3	214.3	0.200
4	111.1	111.1	0.010	166.7	166.5	0.118	277.8	276.4	0.207
5	128.6	128.7	0.040	200.0	199.4	0.137	342.9	338.0	0.203
6	145.8	146.3	0.061	233.3	232.0	0.147	408.3	399.3	0.196

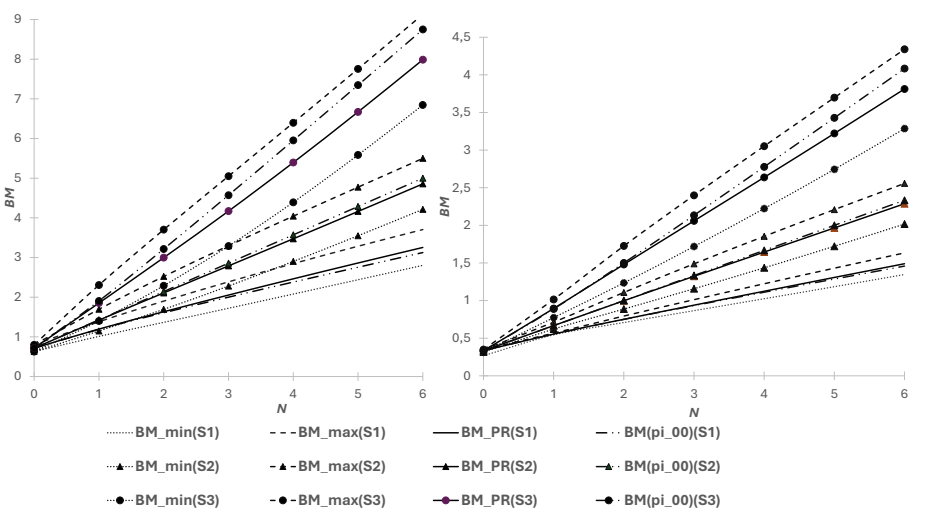


Figure 1. BM coefficients in M1 with the class Γ_1 of priors and different profiles, $t = 1$ (left graph) and $t = 5$ (right graph).

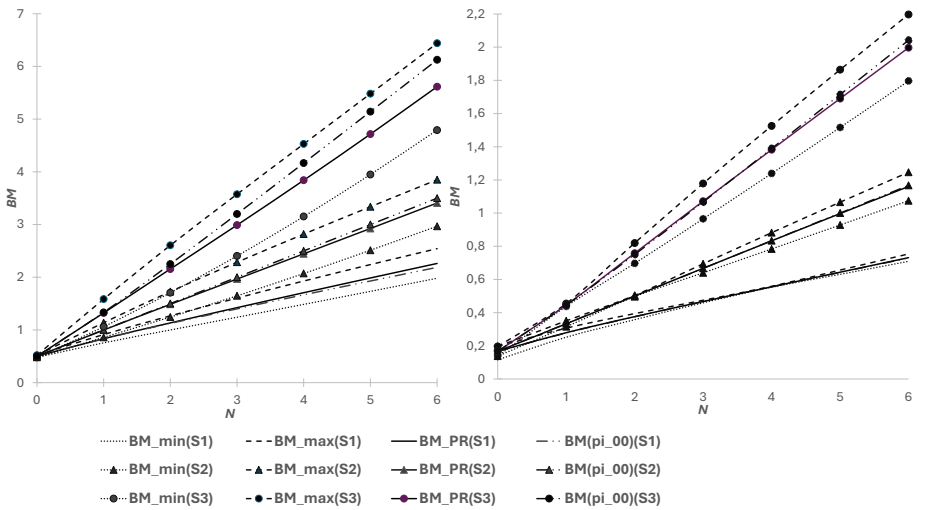


Figure 2. BM coefficients in M2 with the class Γ_1 of priors and different profiles, $t = 1$ (left graph) and $t = 5$ (right graph).

Table 4. Sensitivity of H^C and BM in models M3 and M4 and Γ_2 class of priors. $R_{min} = \frac{\inf_{\pi\in\Gamma_2} H^C(\pi)}{H^C(\pi^{\varepsilon,\eta})}$, $R_{max} = \frac{\sup_{\pi\in\Gamma_2} H^C(\pi)}{H^C(\pi^{\varepsilon,\eta})}$

ε, η	$Corr$	$H^C(\pi^{\varepsilon,\eta})$	R_{min}	R_{max}	$\sup R(BM)$ $\frac{S}{N} =$ 100	200	400
Γ_2, model M3							
$\varepsilon = 0.1, \eta = 0.1$	0	92	1	1	0.229	0.181	0.181
$\varepsilon = 0.1, \eta = 0.5$	0	92	1	1	0.231	0.177	0.200
$\varepsilon = 0.5, \eta = 0.1$	0	140	1	1	0.038	0.014	0.057
$\varepsilon = 0.5, \eta = 0.5$	0	140	1	1	0.197	0.078	0.329
Γ_2, model M4							
$\varepsilon = 0.1, \eta = 0.1$	$[-0.005, 0.048]$	96.6	0.994	1.056	0.321	0.747	1.277
$\varepsilon = 0.1, \eta = 0.5$	$[-0.022, 0.022]$	115	0.974	1.026	0.320	0.612	1.166
$\varepsilon = 0.5, \eta = 0.1$	$[-0.017, 0.017]$	147	0.980	1.020	0.069	0.131	0.222
$\varepsilon = 0.5, \eta = 0.5$	$[-0.071, 0.071]$	175	0.914	1.086	0.414	0.488	0.650

Consider the class Γ_2 of priors in two cases. In both cases $\pi_{10} = Gamma(1;2.5)$, $\pi_{11} = Gamma(1;1)$, $\pi_{20} = IGamma(3;400)$. However, in the first case (model M3) $\pi_{21} = IGamma(2.2;240)$, while in the second one (model M4) $\pi_{21} = IGamma(3;600)$. The prior for λ is a mixture of the priors from models M1 and M2. The prior for θ is a mixture of priors with the same expected value equal to 200 and different variances (model M3) and different expected values but the same shape parameter (model M4). In the first case, if a prior belongs to the class Γ_2, then the Pearson correlation coefficient between λ and θ is 0 (expression (12)), and the collective premium $H^C(\pi_\tau)$ does not depend on τ (it is robust with respect to τ). In the second case, the oscillation of the correlation coefficient, for selected values ε and η, is presented in Table 4 (see the second column).

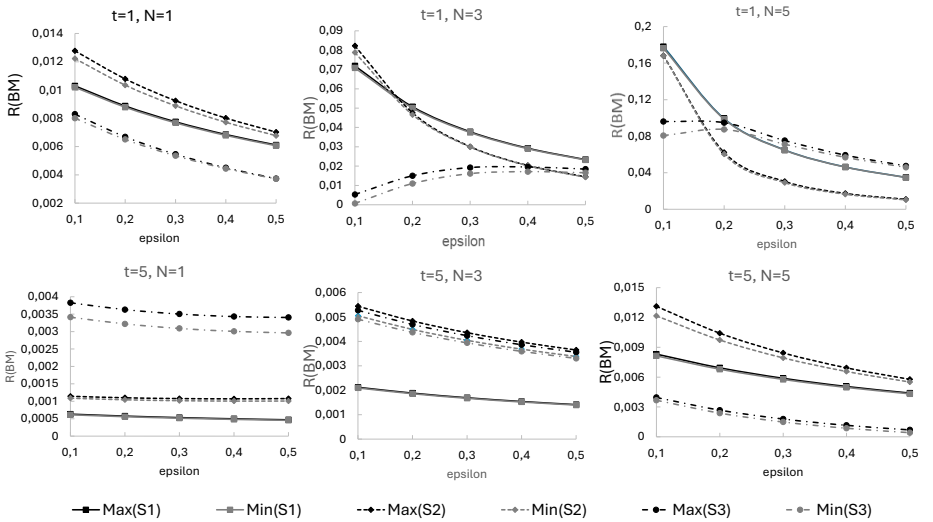


Figure 3. Maximum (black lines) and minimum (grey lines) of $R(BM)$ with respect to $\eta \in [0.1, 0.5]$ as a function of ε for different scenarios (S1, S2, S3), t and N in model M3.

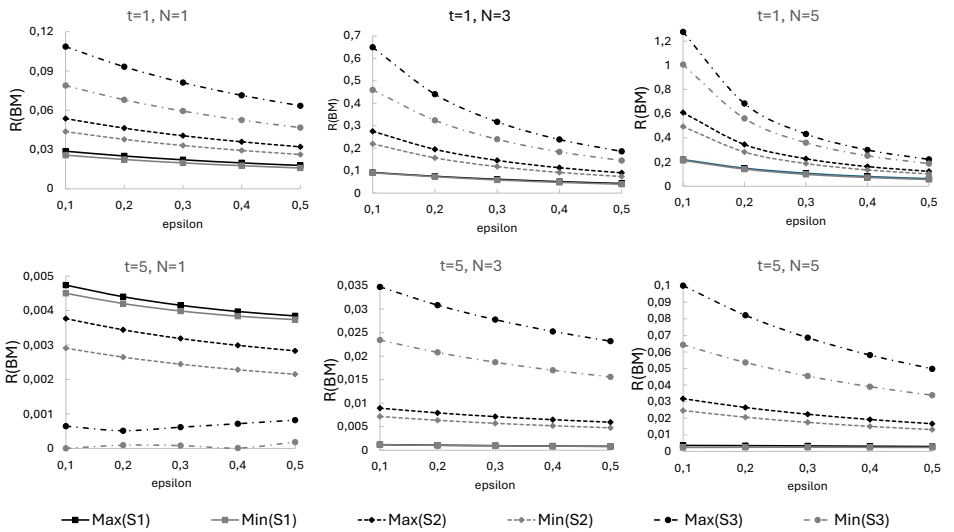


Figure 4. Maximum (black lines) and minimum (grey lines) of $R(BM)$ with respect to $\eta \in [0.1, 0.5]$ as a function of ε for different scenarios (S1, S2, S3), t and N in model M4.

Table 4 also presents the relative sensitivity of H^C (columns: R_{max} and R_{min}) as well as the maximum difference $\sup_A R(BM)$ where

$$R(BM)(\varepsilon, \eta, N^t, S^t) = \sup_{\tau \in [0, \min\{\varepsilon, \eta\}]} \frac{H^B(\pi_\tau, N^t, S^t)}{H^C(\pi^{\varepsilon,\eta})} - \inf_{\tau \in [0, \min\{\varepsilon, \eta\}]} \frac{H^B(\pi_\tau, N^t, S^t)}{H^C(\pi^{\varepsilon,\eta})}$$

and

$$A = \{(N,t) : N \in \{0,1,2,\ldots,6\},\ t \in \{1,2,3,4,5,10\}\}.$$

These values are calculated for selected parameter values ε and η, and for three scenarios (see the last three columns). The maximum difference is significantly greater than the relative oscillation of H^C (see columns 4 and 5). Figures 3 and 4 illustrate the influence of parameters ε and η. The oscillation $R(BM)$ of the BM coefficient with respect to the parameter τ depends more strongly on ε than on η. For a given ε, the change in oscillation as a function of η is much greater in model M4 than in model M3.

Table 5. Sensitivity of Bayes premium in model M3 with the class Γ_2 of priors, $\varepsilon = \eta = 0.1$, $H_0^B = H^B(\pi^{\varepsilon,\eta}, N^t, S^t)$, $\frac{r}{H_0^B} = \frac{r(H^B,\Gamma_2)}{H_0^B}$

	$S/N = 100$			$S/N = 200$			$S/N = 400$		
N	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$
					$t = 1$				
0	60.2	60.2	0.000	60.2	60.2	0.000	60.2	60.2	0.000
1	103.1	102.7	0.009	124.6	124.1	0.009	167.4	167.1	0.004
2	145.9	144.7	0.021	196.2	194.8	0.018	296.3	295.9	0.003
3	193.8	191.1	0.034	279.0	276.1	0.026	449.1	448.9	0.001
4	248.9	244.4	0.045	376.1	371.3	0.031	630.1	630.9	0.003
5	310.9	304.4	0.052	487.0	480.7	0.032	838.8	841.8	0.009
6	377.1	368.6	0.056	607.2	600.4	0.027	1067.0	1073.8	0.016
					$t = 3$				
0	37.1	37.1	0.000	37.1	37.1	0.000	37.1	37.1	0.000
1	61.9	61.9	0.001	74.8	74.7	0.003	100.6	100.4	0.004
2	84.3	84.2	0.004	113.3	113.1	0.006	171.2	171.0	0.003
3	106.2	105.8	0.009	152.9	152.4	0.008	246.2	246.0	0.002
4	128.4	127.7	0.013	194.0	193.2	0.011	325.1	324.9	0.001
5	151.4	150.3	0.018	237.1	235.7	0.014	408.3	408.2	0.001
6	175.4	173.8	0.022	282.4	280.5	0.017	496.3	496.4	0.000
					$t = 5$				
0	27.0	27.0	0.000	27.0	27.0	0.000	27.0	27.0	0.000
1	44.8	44.9	0.001	54.2	54.2	0.002	72.8	72.7	0.004
2	60.7	60.7	0.000	81.6	81.5	0.003	123.3	123.1	0.003
3	75.9	75.8	0.003	109.3	109.1	0.004	175.9	175.7	0.003
4	90.9	90.7	0.005	137.3	137.0	0.005	230.1	229.9	0.002
5	105.8	105.5	0.007	165.8	165.4	0.007	285.5	285.4	0.001
6	121.0	120.5	0.009	194.8	194.2	0.008	342.3	342.2	0.001

Comparing Tables 5 and 6 we also see the significant influence of the prior distribution and the Pearson correlation coefficient on the range of H^B. If $Corr(\lambda,\theta)=0$ (model M3) the relative oscillation $\frac{r(H^B,\Gamma_2)}{H_0^B}$ is less then 0.08 for all $\varepsilon\in[0.1,0.5]$ and $\eta\in[0.1,0.5]$. Similarly to the class Γ_1, in model M4, this oscillation and the range of BM factor are increasing functions of the average severity of claims, and for the S2 and S3 profiles, the relative oscillation is many times greater than in model M3.

Table 6. Sensitivity of Bayes premium in model M4 with the class Γ_2 of priors, $\varepsilon=\eta=0.1$, $H_0^B=H^B(\pi^{\varepsilon,\eta},N^t,S^t)$, $\frac{r}{H_0^B}=\frac{r(H^B,\Gamma_2)}{H_0^B}$

	$S/N=100$			$S/N=200$			$S/N=400$		
N	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$	H_0^B	H_{PR}^B	$\frac{r}{H_0^B}$
$t=1$									
0	63.2	63.5	0.010	63.2	63.5	0.010	63.2	63.5	0.010
1	107.5	108.6	0.026	129.0	131.0	0.040	171.6	175.8	0.061
2	151.1	153.3	0.036	201.5	207.0	0.068	301.5	314.0	0.105
3	199.6	203.1	0.044	285.2	295.8	0.093	455.1	479.5	0.138
4	255.3	260.8	0.053	383.2	400.0	0.111	636.9	673.6	0.152
5	318.0	326.5	0.065	495.0	518.1	0.119	846.4	891.7	0.146
6	384.7	397.7	0.081	616.0	644.1	0.117	1075.4	1123.5	0.126
$t=3$									
0	39.0	38.8	0.012	39.0	38.8	0.012	39.0	38.8	0.012
1	64.6	64.5	0.002	77.5	77.6	0.003	103.1	103.6	0.012
2	87.3	87.4	0.004	116.4	117.1	0.015	174.2	176.2	0.029
3	109.4	109.7	0.007	156.3	157.9	0.025	249.5	253.8	0.044
4	131.7	132.2	0.010	197.7	200.4	0.035	328.6	336.1	0.058
5	154.8	155.5	0.012	240.9	245.1	0.044	412.0	423.3	0.070
6	178.9	179.9	0.013	286.5	292.4	0.052	500.2	515.7	0.080
$t=5$									
0	28.4	28.2	0.019	28.4	28.2	0.019	28.4	28.2	0.019
1	46.8	46.6	0.010	56.1	56.0	0.006	74.7	74.6	0.001
2	62.9	62.7	0.005	83.8	83.9	0.001	125.4	125.9	0.010
3	78.2	78.1	0.001	111.7	112.1	0.008	178.3	179.6	0.019
4	93.2	93.2	0.001	139.9	140.6	0.013	232.6	235.0	0.026
5	108.2	108.3	0.002	168.5	169.7	0.018	288.1	292.0	0.034
6	123.4	123.6	0.003	197.6	199.4	0.023	345.0	350.5	0.040

5. Conclusions

The problem of Bayesian estimation of the premium in the collective risk model under dependent random variables describing the average number and severity of claims is considered. Two different classes of priors are presented. The optimal premiums and the range of the collective and Bayes premiums are calculated. The situation where the dependence between λ and θ does not have influence on the value of the collective premium is presented. In the example, we see that the dependence can produce significantly different Bayes

premiums compared to the case of independent variables, even if the Pearson correlation coefficient is near 0. It also has a significant impact on the fluctuation of the bonus-malus factor. However, the posterior regret gamma-minimax premiums can be close to the Bayes premiums calculated when random variables describing the average number and severity of claims are independent.

Under the square error loss function the Bayes predictor and the posterior regret gamma-minimax predictor of a sum of claims in the $t+1$-period given data (N^t, S^t) is equal to the Bayes estimator and the posterior regret gamma-minimax estimator of the expected value equal to $a\lambda\theta$. Thus, we obtain the same results for the robustness of the Bayes prediction.

References

Berger, J. O., (1994). An overview of robust Bayesian analysis. *Test,* 3, pp. 5–124 (with discussion).

Boratyńska, A., (2008). Posterior regret Γ -minimax estimation of insurance premium in collective risk model. *ASTIN Bulletin,* 38, pp. 277–291.

Boratyńska, A., (2021). Robust Bayesian insurance premium in a collective risk model with distorted priors under the generalised Bregman loss. *Statistics in Transition,* 22, pp. 123–140.

Boratyńska, A., Zielińska–Kolasińska, Z., (2022). Robust Bayesian estimation and prediction in gamma-gamma model of claim reserves. *Insurance: Mathematics and Economics,* 105, pp. 194–202.

Cheung, E., Ni, W., Oh, R. and Woo, J., (2021). Bayesian credibility under a bivariate prior on the frequency and the severity of claims. *Insurance: Mathematics and Economics,* 100, pp. 274–295.

Furman, E., Zitikis, R., (2008). Weighted premium calculation principles. *Insurance: Mathematics and Economics,* 42, pp. 459–465.

Gómez-Déniz, E., (2009). Some Bayesian Credibility Premiums Obtained by Using Posterior Regret Γ-Minimax Methodology. *Bayesian Analysis,* 4, pp. 223–242.

Gómez-Déniz, E., (2016). Bivariate credibility bonus–malus premiums distinguishing between two types of claims. *Insurance: Mathematics and Economics,* 70, pp. 117–124.

Gómez-Déniz, E., Hernandez-Bastida, A. and Vázquez-Polo, F.J., (1999). The Esscher premium principle in risk theory: a Bayesian sensitivity study. *Insurance: Mathematics and Economics,* 25, pp. 387–395.

Gómez-Déniz, E., Hernandez-Bastida, A., Pérez, J. M. and Vázquez-Polo, F. J., (2002). Measuring sensitivity in a bonus-malus system. *Insurance: Mathematics and Economics,* 31, pp. 105-113.

Gschlößl, S., Czado, C., (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal,* 2007(3), pp. 202–225.

Harrouche, L., Fellag, H. and Atil, L., (2025). Bayesian prior robustness using general ϕ-divergence measure. *Stat Papers,* 66, article 14. https://doi.org/10.1007/s00362-024-01628-z.

Heilmann, W. R., (1989). Decision theoretic foundations of credibility theory. *Insurance: Mathematics and Economics,* 8, pp. 77–95.

Hernandez-Bastida , A., Fernández-Sánchez, M.P. and Gómez-Déniz, E., (2009). The net Bayes premium with dependence between the risk profiles. *Insurance: Mathematics and Economics,* 45, pp. 247–254.

Ho, P., (2023). Global robust Bayesian analysis in large models. *Journal of Econometrics,* 235, pp. 608–642.

Hu, G., Xiao, X., (2021). Robust Bayesian estimator in a normal model with uncertain hierarchical priors. *Communications in Statistics - theory and Methods,* 52, pp. 567–582.

Lee, G. Y., Shi, P., (2019). A dependent frequency-severity approach to modeling longitudinal insurance claims. *Insurance: Mathematics and Economics,* 87, pp. 115–129.

Lee, W., Park, S.C., Ahn, J. Y., (2019). Investigating dependence between frequency and severity via simple generalized linear models. *Journal of the Korean Statistical Society,* 48, pp. 13–28.

Lemaire, J., (1995). *Bonus-Malus Systems in Automobile Insurance.* Kluwer Academic Publishers.

Nelsen, R. B., (2006). *An Introduction to Copulas.* 2nd edition, Springer, New York.

Oh, R., Shi, P. and Ahn, J.Y., (2020). Bonus-Malus premiums under the dependent frequency-severity modeling. *Scandinavian Actuarial Journal,* Vol. 2020, pp. 172–195.

Peters, G. W., Targino, R. S. and Wuthrich, M., (2017). Full Bayesian analysis of claims reserving uncertainty. *Insurance: Mathematics and Economics,* 73, pp. 41–53.

Ríos Insua, R. D., Ruggeri, F. and Vidakovic, B., (1995). Some results on posterior regret Γ-minimax estimation. *Statistics & Risk Modeling,* 13, pp. 315–332.

Ríos Insua, D., Ruggeri, F. (eds.), (2000). Robust Bayesian analysis. *Lecture Notes in Statistics,* Vol. 152. Springer-Verlag. New York.

Ruggeri, F., Sánchez-Sánchez, M., Sordo, M. A. and Suárez-Llorens, A., (2021). On a New Class of Multivariate Prior Distributions: Theory and Application in Reliability. *Bayesian Analysis,* 16, pp. 31–60.

Ruggeri, F., Sánchez-Sánchez, M. and Suárez-Llorens, A., (2025). Measuring Bayesian sensitivity in the compound Poisson process. *TEST,* 34. https://doi.org/10.1007/s11749-025-00970-0.

Sánchez-Sánchez, M., Sordo, M.A., Suárez-Llorens, A. and Gómez-Déniz, E., (2019). Deriving robust Bayesian premiums under bands of prior distributions with applications. *ASTIN Bulletin,* 49, pp. 147–168.

Shi, P., Feng, X. and Ivantsova, A., (2015). Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics,* 64, pp. 417–428.

Tomer, S. K., Rai, H., (2021). Robust Bayesian estimation of cumulative incidence function for competing risk data with missing causes. *Journal of Statistical Computation and Simulation,*92(9), pp. 1781–1804. https://doi.org/10.1080/00949655.2021.2007385 .

Young, V. R., (2004). Premium principles. *Encyclopedia of actuarial science.* John Wiley & Sons. New York, NY, USA.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 111–127, https://doi.org/10.59139/stattrans-2026-019
Received – 03.03.2025; accepted – 10.02.2026

New version of Log-Logistic distribution: properties and applications to survival time and demographic data

Arjun Kumar Gaire[1]

Abstract

This paper proposes a Deflation-Inflation Log-Logistic (DILLog) distribution as a sub-model of the Deflation-Inflation Distributed (DID) family, introduced by Alodat and Al-Rawwash (2021). The proposed model offers greater flexibility than the original model in fitting data from real-world problems, especially for survival times and demographic data. The DILLog model is characterized by unimodal right-tailed density and hazard rate functions, and its key statistical properties, including the cumulative distribution function and a closed-form quantile function,are derived. To test the performance of the distribution, a simulation study has been used as well as an application to two real data sets: the age at menarche of Nepalese girls and the survival times of patients suffering from melanoma disease. To illustrate the usefulness and application of the proposed distribution, its parameters were estimated by using the maximum likelihood estimation method. The analysis and plots of the fitted results attest to the the DILLog model being flexible enough to fit the right-skewed real data. The actual application of demographic and survival datasets demonstrates that the DILLog distribution outperforms comparative models.

Key words: deflation-inflation, Log-Logistic, melanoma, menarche, survival.

1. Introduction

In the recent literature on univariate probability distributions, various researchers, statisticians, and mathematicians have introduced numerous new distributions. These include pioneering new families and modifying the existing distribution by adding extra parameters, such as shape, scale, location, or threshold, to the original distribution. They aimed to meet the increasing demand for a flexible distribution in various aspects of everyday life. Such modified distributions have been applied to model a wide range of fields, including economics, physics, bio-statistics, engineering, and many others. Recently, such new distributions have been utilized in actuarial data analysis (Mohammad and Cooray, 2024), demography (Gaire and Aryal, 2015; Gaire et al., 2022; Gaire, 2023; Gaire et al., 2024a), and various other applications reported in the literature. In this study, a new Deflation-Inflation Log-Logistic (DILLog) distribution is introduced, based on the Deflation-Inflation (DI) distribution, the concept proposed by Alodat and Al-Rawwash (2021). Further, some statistical properties of the distribution have been formulated to illustrate its flexibility and applicability. Two real data sets of the age at menarche of Nepalese girls and the survival time of patients suffering from melanoma disease have been applied. The results of data fitting using

[1] Department of Science and Humanities, Khwopa Engineering College, Purbanchal University, Bhaktapur, Nepal. E-mail: arjun.gaire@gmail.com. ORCID: https://orcid.org/0000-0002-1958-9797.

the new model were compared to those of the LLog, Transmuted LLog (TrLLog), and Kumaraswamy LLog (KuLLog) distributions. Akaike's information criteria (AIC), Bayesian information criteria (BIC), Kolmogorov-Smirnov (KS) test, and Anderson-Darling (AD) test have been used to test the significance of fitting the data sets. The proposed model is found to be more flexible and a better fit to the data than the selected probability distributions.

The rest of the manuscript is organized as follows. Section 2 introduces a new DILLog distribution that has been formulated. In Section 3, some statistical properties and the rule for generating random numbers are derived. Section 4 includes the methods of parameter estimation. Section 5 presents the simulation study, and Section 6 contains the numerical application and model validation using two actual sets. Finally, section 7 concludes the manuscript.

2. Model formulation

In this study, a new univariate statistical distribution, named the DILLog distribution, is formulated by applying the DI-family of distributions (Alodat and Al-Rawwash, 2021). The cumulative density function (CDF) and probability density function (PDF) of the DI-family of distribution are expressed in Equations 1 and 2 as follows:

$$F(x) = \frac{ln[1+\lambda G(x)]}{ln(1+\lambda)} \tag{1}$$

$$f(x) = \frac{\lambda g(x)}{ln(1+\lambda)(1+\lambda G(x)}, \ \lambda > -1 \tag{2}$$

where the $g(x)$ and $G(x)$ are the PDF and CDF of the base distribution to be chosen. In the DI-family of distribution, the parameter λ is called the inflation-deflation parameter and controls how the new distribution deviates from the baseline distributions. When $\lambda = 0$, the DI distribution reduces to the baseline distribution, which means no inflation or deflation occurs. When $\lambda > 0$, the DI distribution exhibits inflation of the base PDF in the tail or central region. It increases the tail's heaviness, making large values more likely. Further, with the value of the parameter in the range $-1 < \lambda < 0$, the distribution exhibits deflation, reducing the PDF in the tails or central regions so that it produces lighter tails than the baseline distribution (Alodat and Al-Rawwash, 2021, P. 3). In this study, the LLog distribution is chosen as a base distribution. Because of the flexible nature of the LLog distribution, it has been used to modeling different real-world problems and also has been chosen as a base distribution in different generalized and modified versions such as Kumaraswamy LLog (De-Santana et al., 2012), Beta LLog (Lemonte, 2014) using the concept of Eugene et al. (2002) and Jones (2004), Transmuted LLog (Aryal, 2013), Marshall-Olkin extended LLog (Gui, 2013). Further, Zografos-Balakrishnan LLog (Hamedani, 2013), McDonald LLog (Tahir et al., 2014), Additive Weibull LLog (Hemeda, 2018), Transmuted generalized LLog (Adeyinka et al., 2019), Skew LLog (Gaire et al., 2019, Gaire and Gurung, 2024b),

and Rayleigh Generated LLog (Gaire and Gurung, 2024a). In the same line, in this study, the LLog distribution has been chosen as the base distribution to formulate a four-parameter DILLog distribution as a special case of the DI-family of distribution.

The PDF and the CDF of the three-parameter LLog distribution are given as:

$$g(x) = \frac{\alpha}{\beta} \frac{\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^2} \tag{3}$$

$$G(x) = \frac{\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} \tag{4}$$

where the first constant $\alpha > 0$ is the shape parameter, the second constant $\beta > 0$ is the scale parameter, and the third constant $x > \gamma$ is the location parameter. In many cases, we need a threshold parameter to guarantee that no failure occurs before a given time. Here, we chose the LLog distribution with the third parameter $\gamma > 0$, since many research studies report positive data exceeding a certain threshold. For example, the age of a girl at menarche is, of course, greater than zero.

After substituting the value of $g(x)$ and $G(x)$ of the LLog distribution in Equations 1 and 2, the CDF and PDF of the DILLog distribution are obtained and expressed in Equations 5 and 6 as follows. Both the CDF and PDF satisfy the required validity conditions, with the CDF meeting the boundary conditions and the PDF being non-negative and integrating to unity over the specified range $(0, \infty)$.

$$F(x) = \frac{1}{\ln(1+\lambda)} \ln\left(\frac{1+(1+\lambda)\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}\right) \tag{5}$$

$$f(x) = \frac{\alpha\lambda\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\beta\ln(1+\lambda)\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)\left(1+(1+\lambda)\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)}, \; x > \lambda, \; \alpha,\beta > 0, \; \lambda > -1 \tag{6}$$

The PDF and CDF of the DILLog distribution for selected parameter values are plotted in Figures 1 and 2, respectively.

The reliability function $R(x)$, which is the probability of an item not failing before some time x, is defined by $R(x) = 1 - F(x)$. Thus, the reliability function of a DILLog probability distribution is given in Equation 7, and its visual representation is shown in Figure 3.

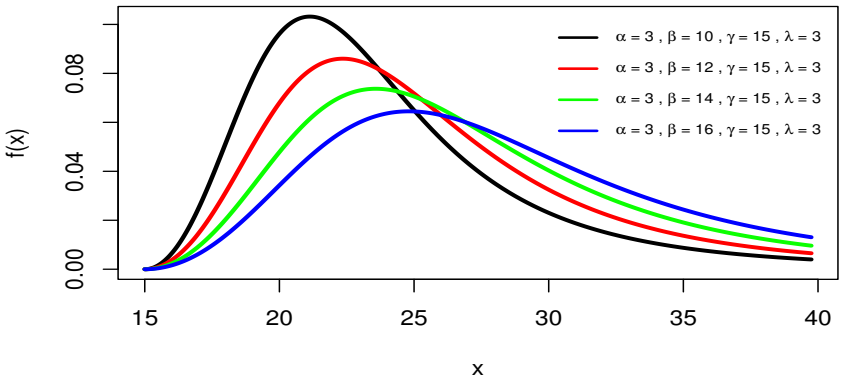


Figure 1. Plots of the PDF of DILLog distribution for different value parameters

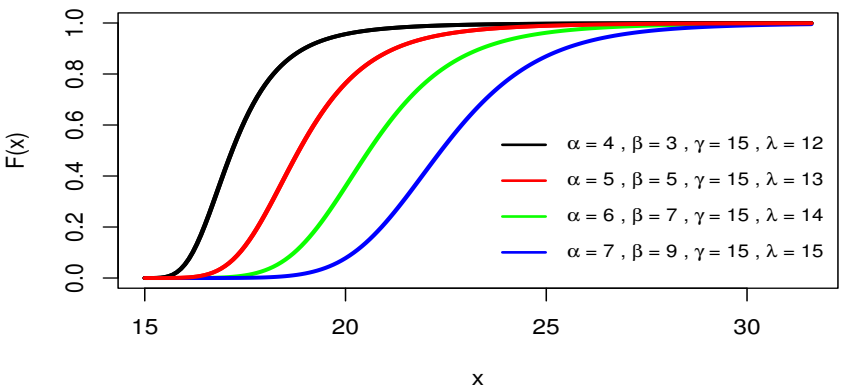


Figure 2. Graph of the CDF of DILLog for different values of parameters

$$R(x) = 1 - \frac{1}{\ln(1+\lambda)} \ln\left(\frac{1+(1+\lambda)\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}\right) \tag{7}$$

The conditional probability of failure assuming that it has survived up to the time x is given by the hazard rate function defined by $h(x) = \frac{f(x)}{1-F(x)}$. The hazard rate function of the DILLog distribution is given by Equation 8, and its graphical representation is shown in Figure 4. The hazard rate curve exhibits a rapidly increasing unimodal shape. Then it decreases steadily to attain the right-tailed curve, suggesting that this shape is suitable for modeling right-tailed data.

$$h(x) = \frac{\alpha\lambda}{\beta} \cdot \frac{\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left[\ln\left(\frac{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{\ln(1+\lambda)\left(1+(1+\lambda)\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)}\right)\right]^{-1}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)\left(1+(1+\lambda)\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)} \tag{8}$$

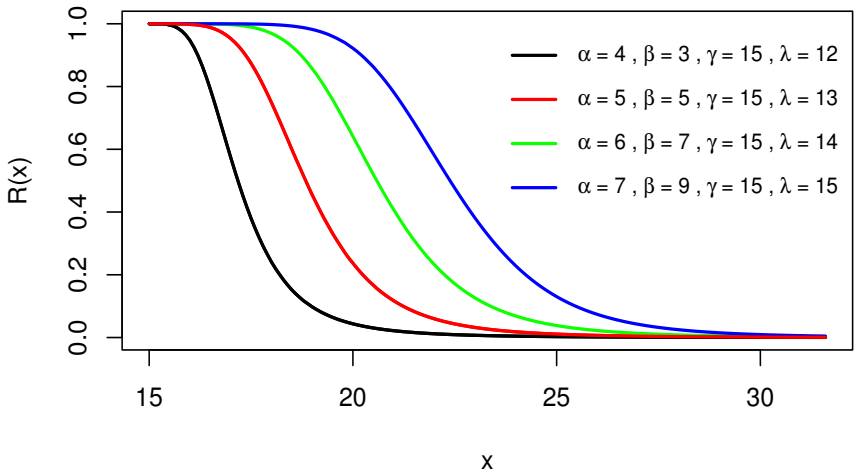


Figure 3. The plot of the reliability function of the DILLog probability distribution

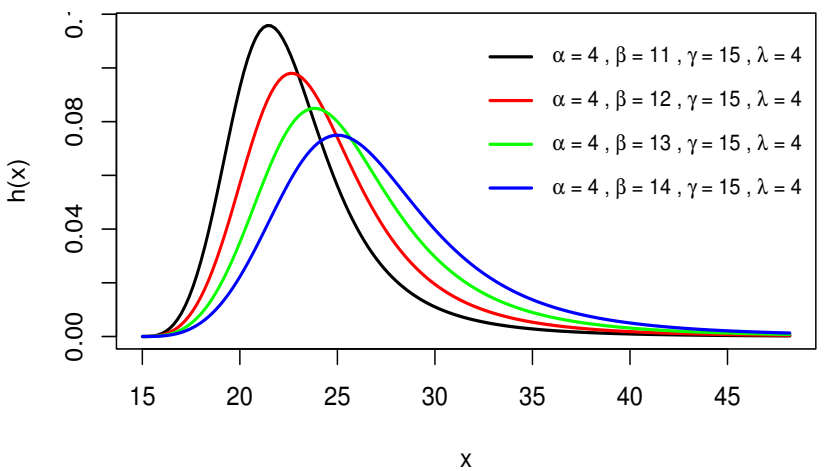


Figure 4. Plot of hazard rate function of DILLog distribution

The cumulative hazard rate function defined by $H(x) = -ln[R(x))$ of the DILLog distribution has been expressed in Equation 9 and is illustrated graphically in Figure 5. All of these functions support survival and reliability testing, actuarial data modeling, demographic data modeling, and numerous other fields.

$$H(x) = Ln\left[1 - \frac{1}{\ln(1+\lambda)} \ln\left(\frac{1+(1+\lambda)\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}\right)\right] \tag{9}$$

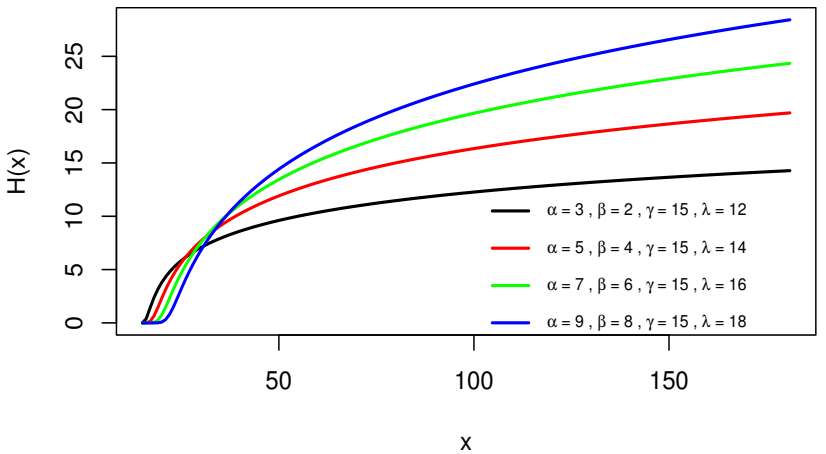


Figure 5. Graph of cumulative hazard rate function for different values of parameters

3. Statistical properties

In this section of this research, some structural properties of the DILLog distribution were presented.

3.1. Linear representation

For a more straightforward calculation of statistical features and further use of this model, the linear representations of the PDF and CDF of the DILLog distribution were presented. For this, the following series expansions were used.

The Taylor series expansion of the natural logarithm, given in Equation 10, where $m \in N$ is a positive integer.

$$ln(1+x) = \sum_{m=1}^{\infty} \frac{(-1)^x x^x}{m}, \text{ for } |x| < 1 \tag{10}$$

Also, the geometric series expansion is expressed in Equation 11, where $j \in N_0$ is a non-negative integer.

$$(1+x)^{-1} = \sum_{j=1}^{\infty} \frac{(-1)^j x^x}{m}, \text{ for } < 1 \tag{11}$$

Using these expansions Equation 12 represents the expanded form of the PDF and Equation 13 represents the expanded form CDF of DILLog distributions, where $m \in N$, and $i, j \in N_0$.

$$f(x) = \frac{\alpha\lambda}{\beta \ln(1+\lambda)} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (-1)^{i+j} (1+\lambda)^j \left(\frac{x-\gamma}{\beta}\right)^{\alpha(i+j+1)-1} \tag{12}$$

$$F(x) = \frac{1}{\ln(1+\lambda)} \sum_{m=1}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(-1)^{m+i+1}}{m} (1+\lambda)^{j} \left(\frac{x-\gamma}{\beta} \right)^{\alpha(i+j+1)-1} \tag{13}$$

3.2. Quantile function and random number generation

The random number of the DILLog distribution was generated by inverting the CDF expressed on Equation 5. For this, let us suppose $F(x) = U$, where U is the function that follows the uniform distribution in $[0,1]$.

$$F(x) = \frac{1}{\ln(1+\lambda)} \ln \left(\frac{1+(1+\lambda)\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} \right) = U$$

$$Q(X) = \gamma + \beta \left(\frac{exp\{U\ ln(1+\lambda) - 1\}}{1+\lambda - exp\{U\ ln(1+\lambda)\}} \right)^{\frac{1}{\alpha}} \tag{14}$$

The expression in Equation 14 is serves as the quantile function which used is also used to generates random numbers that follow the DILLog distribution. By specifying the parameter values for this distribution, a set of random numbers describing the future scenario can be generated using simulation.

4. Method of parameter estimation

The maximum likelihood estimates (MLEs) of the parameters involved in the DILLog distribution are given as follows: Let $X_1, X_2, \ldots, X_n$ be a random sample of observed values from a DILLog distributed random variable X. Then the likelihood function is given by

$$L = \left(\frac{\alpha\lambda}{\beta \ln(1+\lambda)} \right)^{n} \prod_{i=1}^{n} \left[\frac{\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha-1}}{\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)\left(1+(1+\lambda)\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)} \right] \tag{15}$$

After taking the natural logarithm of Equation 15 on both sides, the log-likelihood function of the DILLog distribution becomes

$$\ln L = n \ln \left(\frac{\alpha\lambda}{\beta \ln(1+\lambda)} \right) + (\alpha-1) \sum_{i=1}^{n} \ln \left(\frac{x_i-\gamma}{\beta} \right) \quad - \sum_{i=1}^{n} \ln \left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} \right)$$
$$- \sum_{i=1}^{n} \ln \left(1+(1+\lambda)\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} \right) \tag{16}$$

The components of the score vectors to estimate the parameter of the distribution are given by the partial derivative with respect to the parameters as:

$$\frac{\partial \ln L}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} ln\left(\frac{x_i-\gamma}{\beta}\right) - \sum_{i-1}^{n}\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} ln\left(\frac{x_i-\gamma}{\beta}\right) \quad (17)$$
$$-(1+\lambda)\sum_{i-1}^{n}\left(1+(1+\lambda)\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} ln\left(\frac{x_i-\gamma}{\beta}\right)$$

$$\frac{\partial \ln L}{\partial \beta} = -\frac{n}{\beta} - \left(\frac{n(\alpha-1)}{\beta}\right) + \frac{\alpha}{\beta}\sum_{i-1}^{n}\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} \quad (18)$$
$$+\frac{\alpha(1+\lambda)}{\beta}\sum_{i-1}^{n}\left(1+(1+\lambda)\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}$$

$$\frac{\partial \ln L}{\partial \gamma} = -\frac{\alpha-1}{\beta}\sum_{i=1}^{n}\left(\frac{x_i-\gamma}{\beta}\right)^{-1} + \frac{\alpha}{\beta}\sum_{i=1}^{n}\left(1+\left(\frac{x_i-\gamma}{\beta}\right)\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha-1} \quad (19)$$
$$+\frac{\alpha\,(1+\lambda)}{\beta}\sum_{i=1}^{n}\left(1+(1+\lambda)\left(\frac{x_i-\gamma}{\beta}\right)\right)^{-1}\sum_{i=1}^{n}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha-1}$$

$$\frac{\partial \ln L}{\partial \lambda} = -\frac{n}{(1+\lambda)\,(ln(1+\lambda))} + \sum_{i=1}^{n}\left(1+(1+\lambda)\left(\frac{x_i-\gamma}{\beta}\right)\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} \quad (20)$$

The parameters of the DILLog distribution are estimated by solving the nonlinear system of equations equating the score vector to zero. As these equations do not yield closed-form solutions, various numerical optimization techniques can be applied. Standard methods include the Newton-Raphson iterative method, which uses first and second derivatives for rapid convergence; the Fisher Scoring method, a variant of Newton-Raphson, which uses the expected information matrix to improve stability; the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, a quasi-Newton approach, which approximates the Hessian without requiring second derivatives; and the Nelder-Mead simplex method, a derivative-free technique suitable for non-differentiable or noisy likelihood surfaces. Depending on the problem context and computational considerations, any of these methods may be used to obtain the maximum-likelihood estimate of the distribution parameters.

5. Simulation study

To test the performance of the maximum likelihood estimate method presented in the previous section, a simulation study is performed. For this, the suggested quantile function of the new distribution was used to generate random numbers that follow the DIL-

Log distribution. Three distinct parameter sets were selected to create random samples, each representing different shapes and scales of the DILLog distribution. The first set ($\gamma = 1.0,\ \alpha = 2.0,\ \beta = 0.5,\ \lambda = 1.0$) corresponds to a distribution with a relatively higher shape parameter, indicating a more peaked and less skewed behavior. The second set ($\gamma = 2.0,\ \alpha = 1.0,\ \beta = 0.6,\ \lambda = 2.0$) models a distribution shifted to the right with moderate scale and moderate inflation effects. The third set ($\gamma = 1.0,\ \alpha = 1.0,\ \beta = 0.7,\ \lambda = 3.0$) features a lower shape parameter with a larger scale and higher inflation parameter, resulting in a heavier tail and greater skewness. For each parameter configuration, random samples of sizes $n = 50, 100$, and 500 were generated to assess the robustness of the estimation procedure across varying distributional shapes and data volumes. Each sample size was repeated $N = 1000$ times. The MLE technique was used to estimate the parameters, the algorithm used was Limited-memory BFGS with Box (L-BFGS-B) constraints. The mean estimated values of the parameters were tested using bias of estimation, variance of difference, and mean squared errors (MSEs). Tables 1 to 3 present the numerical summary of the simulation study results for Set I, Set II, and Set III separately.

Table 1. Simulation results for set I

Sample Size	Parameter	True Value	Estimate	Bias	Variance	MSE
50	γ	1.0	1.0315	0.0315	0.0011	0.0020
	α	2.0	3.7656	1.7656	1.6062	4.7234
	β	0.5	0.5798	0.0798	0.0037	0.0101
	λ	1.0	0.4276	-0.5724	0.3182	0.6458
100	γ	1.0	1.0143	0.0143	0.0002	0.0004
	α	2.0	3.9868	1.9868	2.3014	6.2489
	β	0.5	0.6297	0.1297	0.0028	0.0196
	λ	1.0	0.6037	-0.3963	0.4832	0.6403
500	γ	1.0	1.0029	0.0029	6×10^{-6}	1.4×10^{-5}
	α	2.0	3.4113	1.4113	1.0747	3.0665
	β	0.5	0.6788	0.1788	0.0008	0.0328
	λ	1.0	0.3339	-0.6661	0.3276	0.7714

Among the four parameters, the γ estimates exhibit excellent accuracy, with bias and variance decreasing as the sample size increases, confirming the reliability of its estimation. The α estimates improve significantly with larger sample sizes, as bias reduces and variance remains within a moderate range. Similarly, β estimates show steady improvement, as bias diminishes and variance decreases with increasing sample sizes. Although λ estimates are lower than the actual values, their precision improves as the sample size increases, indicating that larger samples yield more stable estimates.

Table 2. Simulation results for set II

Sample Size	Parameter	True Value	Estimate	Bias	Variance	MSE
50	γ	2.0	2.0074	0.0074	0.0001	0.0001
	α	1.0	1.1742	0.1742	0.3836	0.4140
	β	0.6	0.6491	0.0491	0.0061	0.0085
	λ	2.0	1.6870	-0.3130	1.7552	1.8531
100	γ	2.0	2.0030	0.0030	9.7×10^{-6}	1.86×10^{-5}
	α	1.0	1.0169	0.0169	0.3391	0.3393
	β	0.6	0.7069	0.1069	0.0029	0.0144
	λ	2.0	1.5438	-0.4562	2.6246	2.8327
500	γ	2.0	2.0007	0.0007	5×10^{-6}	1×10^{-6}
	α	1.0	0.8348	-0.1652	0.1523	0.1796
	β	0.6	0.7711	0.1711	0.0009	0.0302
	λ	2.0	1.1891	-0.8109	2.9235	3.5811

Table 3. Simulation results for set III

Sample Size	Parameter	True Value	Estimate	Bias	Variance	MSE
50	γ	1.0	1.0062	0.0062	4.8×10^{-5}	8.6×10^{-5}
	α	1.0	0.9636	-0.0364	0.3245	0.3259
	β	0.7	0.7045	0.0045	0.0097	0.0097
	λ	3.0	2.5322	-0.4678	2.7415	2.9603
100	γ	1.0	1.0028	0.0028	7×10^{-6}	1.47×10^{-5}
	α	1.0	0.8467	-0.1533	0.1475	0.1710
	β	0.7	0.7760	0.0760	0.0034	0.0091
	λ	3.0	2.4397	-0.5603	3.8750	4.1889
500	γ	1.0	1.0006	0.0006	2.5×10^{-7}	5.7×10^{-7}
	α	1.0	0.7389	-0.2611	0.0819	0.1500
	β	0.7	0.8330	0.1330	0.0013	0.0189
	λ	3.0	2.1278	-0.8722	3.2656	4.0263

6. Application to survival and demographic data

To test the flexibility of the DILLog distribution, two real data sets have been applied. The parameters of the proposed model and the comparative models were obtained by maximizing the negative log-likelihood as the Objective function using the OPTIM function in the Adequacy Model (Marinho et al., 2019) of the R package for probability distributions (R Core Team, 2025). The AIC, BIC, KS, and AD tests were applied as validity tools to show the flexibility and suitability of the proposed model. The KS test value measures the maximum difference between the empirical and model CDFs. The AD statistics are a weighted measure of the difference between the empirical and model CDFS.

The first data set is the age at menarche of Nepalese females, taken from the Nepal Demographic and Health Survey 2022 (MoHP, 2023). It contains 14,349 data points on the age at menarche of Nepalese females. The second data set is from Susarla and Vanryzin (1978) and comprises 46 data points on the survival times (months) of patients with melanoma disease. This second data set was recently used by Mohammad and Gaire (2025). Table 4 presents the summary statistics for both data sets.

Table 4. Summary statistics of survival time for melanoma disease and age at menarche

Statistics	Min	Q1	Median	Mode	Mean	Q3	Sk	K	max
Melanoma	3.25	6.75	12.88	4.75	15.66	22.31	1.412	2.220	58.5
Menarche	7	13	14	14	13.96	15	0.623	1.238	25

Both data sets were presented using the box-and-whisker plot in Figure 6. The first plot illustrates the age at menarche for Nepalese girls and shows skewness as well as outliers on both sides. Similarly, the second plots presents a box-and-whisker plot of melanoma survival time, which is also right-skewed and contains one outlier.

The result obtained by fitting the proposed model was compared with a three-parameter LLog distribution with PDF in Equation 3, a four-parameter Transmuted Log-Logistic (TrLLog) distribution with PDF expressed in Equation 23 taken from Aryal (2013), and a five-parameter Kumarswamy Log-Logistic (KuLLog) distribution expressed in Equation 24, taken from De-Santana et al. (2012).

$$f(x)_{\text{TrLLog}} = \frac{\alpha}{\beta}\frac{(1+\lambda)\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} + (1-\lambda)\left(\frac{x-\gamma}{\beta}\right)^{2\alpha-1}}{\left[1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]^{3}} \tag{23}$$

$$f(x)_{\text{KuLLog}} = \frac{ab\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{a\alpha-1}\left[1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]^{-(a+1)}\left\{1-\left(\frac{\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}\right)^{a}\right\}^{b-1} \tag{24}$$

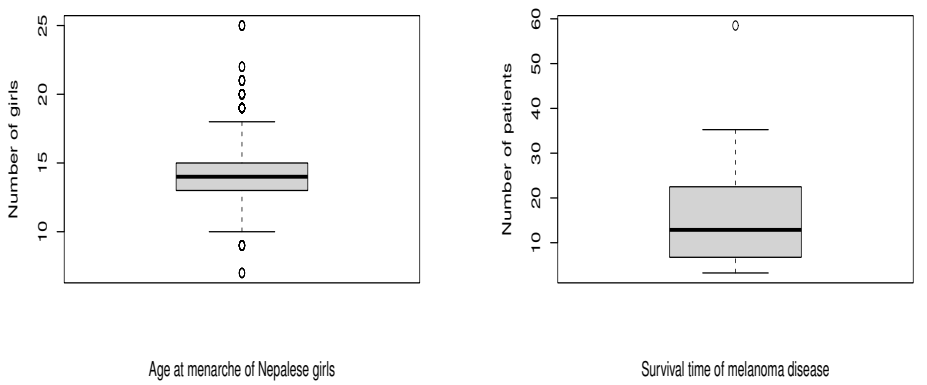


Figure 6. Box-and-whisker plots

Both TrLLog and KuLLog distributions were recently used to modeled demographic data such as the age at menopause (Gaire et al., 2023), the age of a mother at birth of a child (Gaire et al, 2024b), and the age at menarche (Gaire et al., 2024c).

Table 5 presents the goodness-of-fit test statistics and parameter estimates for the DILLog model, along with those of competing models. For the survival time data of melanoma patients, the DILLog distribution shows superior performance, with the lowest AIC (332.2680) and BIC (339.5830), as well as the smallest KS statistic (0.0586) and AD statistic (0.1798) among the models compared. Although the associated p-values for the KS and AD tests are relatively high (e.g. 0.9975 for KS and 0.9950 for AD), it is essential to emphasize that these p-values do not directly measure model fit. Instead, they indicate insufficient evidence to reject the null hypothesis that the data follow the DILLog distribution. A high p-value, in this context, reflects the absence of strong evidence against the model but does not confirm its adequacy. The model selection should primarily be based on comparative fit measures such as AIC, BIC, and the observed values of KS and AD statistics, rather than the magnitude of the p-values. Based on these criteria, the DILLog model emerges as the most appropriate choice for these data sets.

Similarly, Table 6 presents goodness-of-fit test statistics and parameter estimates for the DILLog model and other comparative models of the age at menarche of Nepalese girls. For this data set, multiple validity criteria confirm that the proposed DILLog distribution performs best in testing age-at-menarche data. The proposed model exhibits the minimum AIC (112102) and BIC (112135) values. For the age-at-menarche data of Nepalese girls, the DILLog model has the lowest KS value (0.1359) among the compared models, indicating the best fit. For this data set, the DILLog shows the lowest AD statistic (587.8871) among all compared models, showing it provides the best fit to the data. The comparative analysis reveals that the DILLog distribution provides the best overall fit among the models consid-

Table 5. Estimated parameter values and various goodness-of-fit test statistics for survival time for melanoma disease

Parameters / Test Statistics	**LLog**	**KuLLog**	**TrLLog**	**DILLog**
α	1.727	0.175	1.727	5.894
β	8.934	8005.975	8.986	36.694
γ	2.683	3.164	2.683	0.661
λ	–	–	0.01	859554.2
a	–	8.026	–	–
b	–	81477.4	–	–
AIC	335.1957	334.2525	337.1959	332.268
BIC	340.6817	343.3957	344.5104	339.583
KS	0.078 (0.941)	0.0633 (0.993)	0.0782 (0.941)	0.0586 (0.997)
AD	0.397 (0.851)	0.259 (0.965)	0.397 (0.851)	0.1798 (0.995)

Table 6. Estimated parameter values and various test statistics for the age at menarche

Parameters / Test Statistics	**LLog**	**KuLLog**	**TrLLog**	**DILLog**
α	9.829	7.742	13.576	13.535
β	8.408	12.197	12.929	10.594
γ	5.418	0.010	0.010	5.012
λ	–	–	–0.971	141.899
a	–	4.502	–	–
b	–	2.603	–	–
AIC	112511.9	112154.1	112397.4	112102.0
BIC	112536.9	112195.7	112430.7	112135.0
KS	0.143 (0.000)	0.140 (0.000)	0.152 (0.000)	0.136 (0.000)
AD	618.519 (0.000)	596.077 (0.000)	626.275 (0.000)	587.887 (0.000)

ered, as indicated by its lowest AIC, BIC, KS, and AD values. These results highlight the flexibility and robustness of the DILLog model in capturing the underlying structure of the menarche data. Although the KS test yields very small p-values of 0.0000, this should not be viewed as a sign of a poor model fit. Rather, such p-values are commonly observed in large samples, where even minor deviations from the theoretical distribution become statistically significant. Notably, model selection in this context relies primarily on AIC, BIC, KS, and AD values, rather than the associated p-values. Therefore, despite the limitations of p-values in large sample settings, the DILLog model remains the most suitable and preferred choice for the menarche data.

Figure 7 presents the histogram of observed number of melanoma survivors together with the fitted curves from different models. The DILLog model is found to better fit the data across multiple validity criteria, as confirmed by the fitted graphs and the test results.

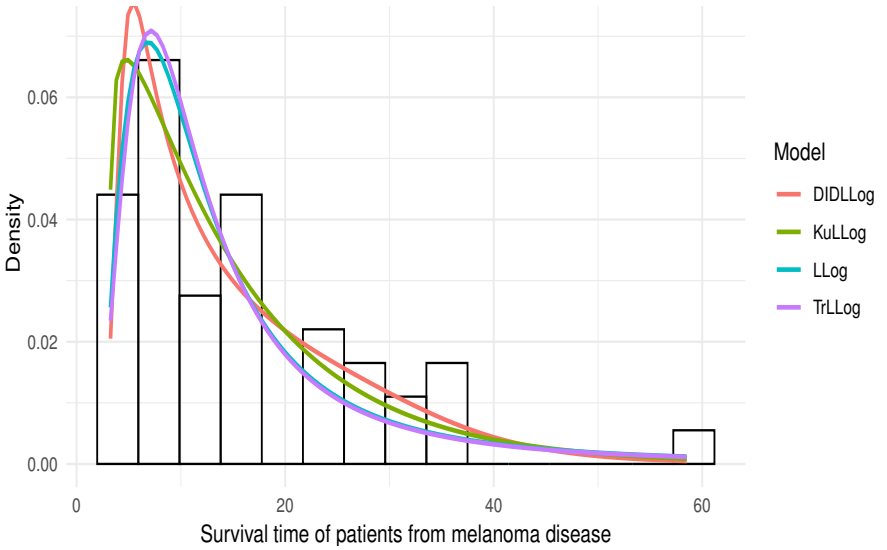


Figure 7. Histogram of observed number of patients with survival times for melanoma disease along with fitted PDFs of competing distributions

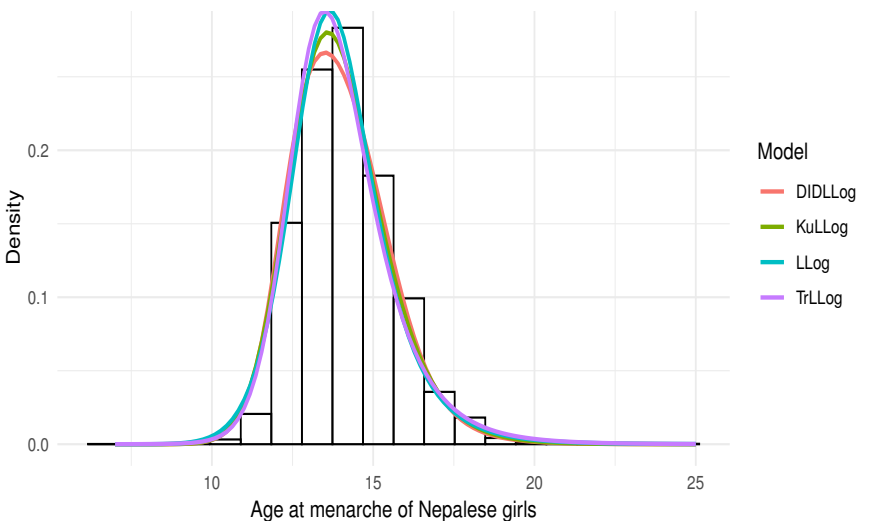


Figure 8. Histogram of observed number of girls with age-at-menarche along with fitted PDFs of competing distributions

Similarly, the histogram of the observed number of Nepalese girls at age at menarche together with the fitted curves from different models was presented on Figure 8. The DILLog model is found to better fit the data across multiple validity criteria, as confirmed by the fitted graphs and the test results.

7. Conclusions

This study introduced a new probability distribution, the DILLog distribution, formulated as a sub-model within the DI family of distributions, with the LLog distribution serving as its baseline due to its recognized flexibility and practical relevance. Several fundamental statistical properties of the proposed model were derived, providing a comprehensive theoretical foundation. Parameter estimation was carried out using the maximum likelihood method, ensuring efficient and reliable estimation across competing models. To evaluate the robustness and practical applicability of the DILLog distribution, both numerical simulations and empirical analyses based on two real data sets were conducted. The results from multiple goodness-of-fit tests consistently demonstrated that the proposed DILLog model outperformed the selected models, confirming its superior flexibility and fitting capability. The findings indicate that the DILLog distribution is a valuable addition to the DI family and holds strong potential for broader applications in statistical modeling and data analysis.

Funding

This research was supported by the Mini Research Grant provided by the Research and Development Unit of Khwopa Engineering College, Bhaktapur, Nepal (Grant Number: KhEC-SSR-08182-009).

Acknowledgment

The author expresses sincere gratitude to Khwopa Engineering College for providing valuable academic resources and a supportive research environment.

References

Adeyinka, F. S., Olapade, A. K., (2019). On transmuted four parameters generalized Log-Logistic distribution. *International Journal of Statistical Distributions and Applications*, 5(2), pp. 32–37.

Alodat, M. D. T., Al-Rawwash, M., (2021). A proposed mechanism for skewing symmetric distributions. *Communications in Statistics-Theory and Methods*, 50(11), pp. 2674–2695.

Aryal, G. R. (2013). Transmuted log-logistic distribution. *Journal of Statistics Applications and Probability*, 2(1), pp. 11–20.

De-Santana, T. V. F., Ortega, E. M., Cordeiro, G. M. and Silva, G. O., (2012). The Kumaraswamy-log-logistic distribution. *Journal of Statistical Theory and Applications*, 11(3), pp. 265–291.

Eugene, N., Lee, C. and Famoye, F., (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and Methods*, 31(4), pp. 497–512.

Gaire, A. K., (2023). Skew Lomax distribution, parameter estimation, its properties, and applications. *Journal of Science and Engineering*, 10, pp. 1–11.

Gaire, A. K., Aryal, R., (2015). Inverse Gaussian model to describe the distribution of age-specific fertility rates of Nepal. *Journal of Institute of Science and Technology*, 20(2), pp. 80–83.

Gaire, A. K., Gurung, Y. B., (2024a). Rayleigh generated Log-Logistic distribution: Properties and performance analysis. *Istatistik Journal of Turkish Statistical Association*, 15(1), pp. 13–28.

Gaire, A. K., Gurung, Y. B., (2024b). Skew Log-logistic distribution: Properties and application. *Statistics in Transition New Series*, 25(1), pp. 43–62.

Gaire, A. K., Gurung, Y. B., and Bhusal, T. P., (2024a). Age at first marriage of Nepalese women: a statistical analysis (status, differential, determinants, and distributional pattern). *Journal of Population and Social Studies*, 32, pp. 308—328.

Gaire, A. K., Gurung, Y. B., and Bhusal, T. P., (2023). Stochastic modeling of age at menopause for Nepalese women and development of menopausal life table. *Global Health Economics and Sustainability*. 1(2), pp. 1–11.

Gaire, A. K., Gurung, Y. B., and Bhusal, T. P., (2024b). Fertility model evolution: a survey on mathematical models of age-specific fertility with application to Nepalese and Malaysian data. *Global Health Economics and Sustainability*. 3(1), pp. 222–234.

Gaire, A. K., Gurung, Y. B., Bhusal, T. P., (2024c). Stochastic modeling of age at menarche for Nepalese girls and developing menarchial life table. *Population Review*, 63(2), pp. 146-161.

Gaire, A. K., Thapa G. B. and KC, S., (2019). *Preliminary results of Skew Log-logistic distribution, properties, and application. Proceeding of the 2nd International Conference on Earthquake Engineering and Post Disaster Reconstruction Planning*, 25–27 April 2019, Bhaktapur, Nepal, pp. 37–43.

Gaire, A. K., Thapa, G. B. and KC, S., (2022). Mathematical modeling of age-specific fertility rates of Nepali mothers. *Pakistan Journal of Statistics and Operation Research*, 18(2), pp. 417–426. https://doi.org/10.18187/pjsor.v18i2.3319.

Gui, W., (2013). Marshall-Olkin extended log-logistic distribution and its application in minification processes. *Applied Mathematical Science*, 7(80), pp. 3947–3961.

Hemeda, S., (2018). Additive Weibull Log Logistic distribution: Properties and application. *Journal of Advanced Research in Applied Mathematics and Statistics*, 3(4), pp. 8–15.

Hamedani, G., (2013). The Zografos-Balakrishnan log-logistic distribution: Properties and applications. *Journal of Statistical Theory and Applications*, 12(3), pp. 225–244.

Jones, M., (2004). Families of distributions arising from distributions of order statistics. *Test*, 13(1), pp. 1–43.

Lemonte, A. J., (2014). The Beta log-logistic distribution. *Brazilian Journal of Probability and Statistics*, 28(3), pp. 313–332.

Marinho, P. R. D., Silva, R. B., Bourguignon, M., Cordeiro, G. M., and Nadarajah, S., (2019). AdequacyModel: An R package for probability distributions and general-purpose optimization. *PloS One*, 14(8), e0221487.

Marshall, A. W., Olkin, I., (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, 84(3), pp. 641–652.

Ministry of Health and Population, New ERA, and ICF International Inc. (2023). Nepal demographic and health survey 2022. *Kathmandu, Nepal: Ministry of Health and Population.*

Mohammad, S., Cooray, K., (2025). Modeling actuarial data using iterated trigonometric distributions. *Communications in Statistics-Theory and Methods*, 54(16), pp. 5112–5128.

Mohammad, S., Gaire, A.K. (2025). Exponential Arctan-G Family of Distribution with Properties and Applications, *Journal of Probability and Statistics*, 25(1), pp.–13.

R Core Team, (2025). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Susarla, V., Van Ryzin, J., (1978). Empirical Bayes estimation survival distribution function from right censored data. *Anal. Statist*, 6, pp. 740–755.

Tahir, M. H., Mansoor, M., Zubair, M. and Hamedani, G., (2014). McDonald log-logistic distribution with an application to breast cancer data. *Journal of Statistical Theory and Applications*, 13(1), pp. 65–82.

Zografos, K., Balakrishnan, N., (2009). On families of Beta-and generalized Gamma-generated distributions and associated inference. *Statistical Methodology*, 6(4), pp. 344–362.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 129–144, https://doi.org/10.59139/stattrans-2026-020
Received – 09.10.2024; accepted – 10.02.2026

Progressive Type II censored exponential data analysis: the method comparison with a breakdown voltage case study

Raed R. Abu Awwad[1], **Ghassan K. Abufoudeh**[2], **Samer Alokaily**[3], **Maalee Almheidat**[4]

Abstract

This study presents a comparative analysis of different estimation methods for the parameter θ of the exponential distribution under progressive Type II censoring. We compare maximum likelihood estimation with Bayesian approaches using squared error and Kullback-Leibler loss functions under different prior specifications. The theoretical developments presented are well established in the statistical literature; our contribution lies in the systematic empirical comparison of these methods. Through simulation studies and real data application, we examine the finite-sample behavior of these estimators to provide practical guidance for researchers. A real dataset from Lawless (1982) illustrates the application of these methods.

Key words: comparative study, exponential distribution, progressive Type II censoring, Kullback-Leibler loss function, Bayesian estimation, maximum likelihood estimation.

1. Introduction

Life testing reliability studies in business, manufacturing, engineering and many other fields can be costly and time-consuming and the experimenter may not be able to get full data on the failure times for all experimental units. Multiple censoring techniques are used to reduce the time and the cost of testing. In a typical life testing experiment, n denotes the total number of items initially placed on test. The censoring techniques: types I and II are the first traditional techniques used. The experimenter tries to end the experiment at a predetermined time point, say T, in the case of Type I censoring technique, but the experiment is completed when a certain number of failures are observed, say m $(m < n)$, in the case of Type II censoring technique. One important development in the censoring techniques is the progressive Type II censoring technique, which allows the experimenter to remove surviving units during the experiment. Many authors have been discussing progressive censorship. Balakrishnan and Aggarwala (2000), Balakrishnan and Cramer (2014) and Aggarwala (1996) are good sources for additional information.

[1]Department of Mathematics, Faculty of Arts and Sciences, University of Petra, Amman, Jordan. E-mail: rabuawwad@uop.edu.jo. ORCID: https://orcid.org/0000-0002-4422-2719.
[2]Department of Mathematics, Faculty of Arts and Sciences, University of Petra, Amman, Jordan. E-mail: gabufoudeh@uop.edu.jo. ORCID: https://orcid.org/0000-0001-8520-021X.
[3]Department of Mathematics, Faculty of Arts and Sciences, University of Petra, Amman, Jordan. E-mail: samer.alokaily@uop.edu.jo. ORCID: https://orcid.org/0000-0001-8847-9519.
[4]Department of Mathematics, The University of Jordan, Amman, Jordan. E-mail: m.almheidat@ju.edu.jo. ORCID: https://orcid.org/0000-0001-7535-4135.

A progressive Type II censoring is carried out in the following way. Imagine n items are used in a life test. One or more surviving items may be arbitrarily eliminated from the life test at the time of each failure (censoring). The censoring takes place in m steps. The following information comes from a progressive Type II censored sample: $\mathbf{x} = (x_{1:m:n}, x_{2:m:n}, \ldots, x_{m:m:n})$ with $(r_1, r_2, \ldots, r_m)$ censoring scheme, where $x_{1:m:n} < x_{2:m:n} < \ldots < x_{m:m:n}$ denotes the m recorded failure times, and $r_1, r_2, \ldots, r_m$ denotes the number of items removed from the life test at each failure time.

In recent years, several authors have discussed the estimation methods of various lifetime distributions when progressive Type II censored data are used. Recent developments within the last five years include Dey et al. (2021), who analyzed progressive Type-II censored gamma distribution using computational approaches and investigated the performance of different estimators; Alshenawy et al. (2021), who developed progressive Type-II censoring schemes for extended odd Weibull exponential distribution with applications in medicine and engineering; Wu and Gui (2021), who proposed Bayesian estimation methods for Nadarajah-Haghighi distribution under progressive Type-II censoring with applications in reliability analysis; Almetwally et al. (2022), who analyzed progressive Type-II censoring for unit-Weibull distribution with optimal scheme and real data applications in reliability engineering; and Ren and Hu (2023), who developed estimation methods for inverse Weibull distribution under progressive Type-II censoring scheme using maximum likelihood, Bayesian, and inverse moment estimation approaches. Earlier foundation works include: Kundu and Raqab (2012), Pradhan and Kundu (2009) and Kim et al. (2011). Progressive Type II censored sampling is a crucial sampling technique for collecting data in lifetime research.

Let $x_{1:m:n}, x_{2:m:n}, \ldots, x_{m:m:n}$ represent the failure times of m independent and identically distributed random variables from an exponential distribution with probability density function (pdf)

$$f(x;\theta) = \begin{cases} \theta e^{-\theta x} & \text{if } x > 0, \theta > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{1}$$

and the cumulative distribution function

$$F(x;\theta) = 1 - e^{-\theta x}, \quad x > 0, \theta > 0. \tag{2}$$

The reliability and hazard functions of the exponential distribution are respectively given by

$$r(t) = e^{-\theta t} \quad \text{and} \quad h(t) = \theta$$

where $t > 0$, $\theta > 0$.

It is important to note that for the exponential distribution, the parameter θ is identical to the hazard function, i.e., $h(t) = \theta$ (constant hazard rate). This fundamental property will be acknowledged throughout our analysis.

The exponential distribution is widely recognized as one of the most fundamental probability distributions in statistical theory and practice (Lawless (1982); Balakrishnan and Aggarwala (2000)). Recent applications demonstrate the utility of exponential distribution in modeling electronic component lifetimes and mechanical system failure rates in reliabil-

ity engineering (Rasheed (2023)). Sapkota et al. (2025) utilize the exponential distribution as a base to construct a novel "New Odd-type Exponential Distribution" within a flexible family, demonstrating its superior performance in bias reduction, model selection, and goodness-of-fit tests on engineering failure times and medical survival datasets.

The theoretical methods for maximum likelihood and Bayesian estimation under progressive censoring are well established in the statistical literature. Abufoudeh et al. (2019) and Abu Awwad et al. (2019) have studied Bayesian estimation under Kullback-Leibler loss for exponential distributions, while the maximum likelihood approach has been extensively covered by Balakrishnan and Aggarwala (2000) and others.

Based on the progressive Type II sample $\mathbf{x} = (x_{1:m:n}, x_{2:m:n}, \ldots, x_{m:m:n})$ of size m, the objective of this article is to provide a comparative analysis of different estimation methods for the unknown parameter, reliability and hazard functions of the exponential distribution under the Kullback-Leibler loss and squared error loss functions. Our contribution is methodological and empirical rather than theoretical, focusing on the practical comparison of these well-established methods.

The squared error loss function (SELF) is defined as

$$L_{SE}(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

Following Kullback and Leibler (1951), who developed a divergence measure called the Kullback-Leibler divergence measure, this measure depends on the difference between the exact distribution $f(x|\theta)$ and the approximate distribution $\hat{f}(x|\hat{\theta})$ denoted by $KL(f, \hat{f})$ with property $KL(f, \hat{f}) \geq 0$. For further information, one can see Abufoudeh et al. (2019) and Abu Awwad et al. (2019). The Kullback-Leibler loss function is given by

$$KL(f, \hat{f}) = \int_{-\infty}^{\infty} f(x|\theta) \log \frac{f(x|\theta)}{\hat{f}(x|\hat{\theta})} dx = \frac{\hat{\theta}}{\theta} - \log \frac{\hat{\theta}}{\theta} - 1.$$

This Kullback-Leibler loss function (KELF) is denoted by $L_{KL}(\theta, \hat{\theta})$, where

$$L_{KL}(\theta, \hat{\theta}) = \frac{\hat{\theta}}{\theta} - \log \frac{\hat{\theta}}{\theta} - 1$$

The remainder of the article is structured as follows. Section 2 reviews the well-known method of deriving the maximum likelihood estimators of the unknown parameter θ of the exponential distribution. Since $h(t) = \theta$ for the exponential distribution, the estimator for the hazard function is identical to that for θ. Section 3 presents the standard Bayesian estimation methods under both SELF and KELF based on progressive Type II censored data. Section 4 presents our main contribution: a comparative simulation study of the estimation results based on multiple progressive Type II samples of different schemes, emphasizing that we are comparing different statistical paradigms rather than determining which is superior. Also, analysis of a real dataset from reliability studies is presented in Section 4. Section 5 presents the conclusions.

2. Frequentist method

This section reviews the standard maximum likelihood estimation approach for progressive Type II censored exponential data, following the well-established methodology in the literature.

Suppose we observe a progressive Type II censored sample $\mathbf{x} = (x_{1:m:n}, x_{2:m:n}, \ldots, x_{m:m:n})$ of size m from a total sample of size n, where the underlying distribution is exponential with parameter θ. As established by Balakrishnan and Aggarwala (2000), the likelihood function of progressive Type II sample from exponential distribution is defined as

$$L(\theta|\mathbf{x}) = C\prod_{i=1}^{m} f(x_{i:m:n})[1-F(x_{i:m:n})]^{r_i}, \tag{3}$$

where $C = n(n-1-r_1)(n-2-r_1-r_2)\cdots(n-m+1-r_1-\cdots-r_{m-1})$.

By using (1), (2) and (3), we directly obtain

$$L(\theta|\mathbf{x}) = C\theta^m e^{-\theta\sum_{i=1}^{m}(1+r_i)x_{i:m:n}} \tag{4}$$

Using the logarithmic function for both sides in Eq. (4), we obtain the log-likelihood function

$$\ln L(\theta|\mathbf{x}) = \ln C + m\ln\theta - \theta\sum_{i=1}^{m}(1+r_i)x_{i:m:n}.$$

Setting the derivative of $\ln L(\theta|\mathbf{x})$ to zero produces the maximum likelihood estimator (MLE) of θ, which is the standard result:

$$\hat{\theta}_{MLE} = \frac{m}{\sum_{i=1}^{m}(1+r_i)x_{i:m:n}}. \tag{5}$$

By the invariance property of MLEs:

$$\hat{r}(t)_{MLE} = e^{-t\hat{\theta}_{MLE}} = e^{-\frac{mt}{\sum_{i=1}^{m}(1+r_i)x_{i:m:n}}} \tag{6}$$

Since $h(t) = \theta$ for the exponential distribution, we have $\hat{h}(t)_{MLE} = \hat{\theta}_{MLE}$. This eliminates the need for separate hazard function calculations.

3. Bayesian estimation and credible intervals

This section presents the standard Bayesian estimation procedures for exponential distributions under progressive censoring. The theoretical developments reviewed here are well established in the Bayesian literature.

Based on the observed m progressive Type II sample $\mathbf{x} = (x_{1:m:n}, x_{2:m:n}, \ldots, x_{m:m:n})$, we apply standard Bayesian methodology to estimate the unknown parameter of the exponential distribution. Since the hazard function $h(t) = \theta$, estimating θ provides estimate for the hazard functions. KELF and SELF are used to obtain the point estimation of the unknown parameter θ and the reliability function.

3.1. Bayesian estimation for θ

Following standard Bayesian theory, we derive the Bayesian estimator of the unknown parameter θ under both squared error and Kullback-Leibler loss functions. The posterior distribution combines the information in the sample and the prior information, making it straightforward to use the Bayes technique. The posterior distribution of θ given $\mathbf{x} = (x_{1:m:n}, x_{2:m:n}, \ldots, x_{m:m:n})$ is obtained as

$$\pi(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})\pi_1(\theta|a,b)}{\int_0^\infty L(\theta|\mathbf{x})\pi_1(\theta|a,b)d\theta}. \tag{7}$$

If θ has a conjugate gamma prior $\pi_1(\theta|a,b)$ with hyper-parameters $a > 0$ and $b > 0$, then the pdf of θ given a and b is defined as

$$\pi_1(\theta|a,b) = \begin{cases} \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta} & \text{if } \theta > 0 \\ 0 & \text{if } \theta \leq 0 \end{cases} \tag{8}$$

By substituting Eqs. (3) and (8) in Eq. (7), using standard conjugate prior theory, we obtain

$$\pi(\theta|\mathbf{x}) = \frac{(b+\sum_{i=1}^m (1+r_i)x_{i:m:n})^{m+a}}{\Gamma(m+a)}\theta^{m+a-1}e^{-\theta(b+\sum_{i=1}^m (1+r_i)x_{i:m:n})}. \tag{9}$$

In other words, the posterior distribution of θ given $\mathbf{x}$ is Gamma$(m+a, b+\sum_{i=1}^m (1+r_i)x_{i:m:n})$.

The Bayes estimate of θ under the squared error loss function is the posterior mean:

$$\hat{\theta}_{BSE} = E_{posterior}(\theta|\mathbf{x}) = \frac{m+a}{b+\sum_{i=1}^m (1+r_i)x_{i:m:n}}. \tag{10}$$

The Bayes estimate of θ under the Kullback-Leibler loss function is obtained by minimizing the risk function as follows:

$$E_{posterior}(KL(\theta,\hat{\theta})) = \int_0^\infty \left(\frac{\hat{\theta}}{\theta} - \log\frac{\hat{\theta}}{\theta} - 1\right)\pi(\theta|\mathbf{x})d\theta.$$

Setting the derivative of $E_{posterior}(KL(\theta,\hat{\theta}))$ to zero produces

$$\int_0^\infty \left(\frac{1}{\theta} - \frac{1}{\hat{\theta}}\right)\pi(\theta|\mathbf{x})d\theta = 0.$$

Solving for $\hat{\theta}$ yields

$$\hat{\theta}_{BKL} = (E_{posterior}(\theta^{-1}|\mathbf{x}))^{-1}$$

For the gamma posterior distribution, this gives the Bayes estimate of θ under KELF:

$$\hat{\theta}_{BKL} = \frac{m+a-1}{b+\sum_{i=1}^m (1+r_i)x_{i:m:n}}. \tag{11}$$

3.2. Bayesian estimation for $r(t)$

We provide the Bayes estimate of the reliability function under squared error and Kullback-Leibler loss functions. The Bayes estimate of $r(t)$ under squared error loss is computed as follows:

$$\begin{aligned}\hat{r}(t)_{BSE} &= E_{posterior}(r(t)|\mathbf{x}) \\ &= E_{posterior}(e^{-\theta t}|\mathbf{x}) \\ &= \left(\frac{b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n}}{t+b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n}}\right)^{m+a}. \end{aligned} \tag{12}$$

The Bayes estimate of a general function of parameter θ, say $g(\theta)$ with respect to KELF is obtained as follows:

$$\hat{\theta}_{B_{KE}} = (E_{posterior}(g(\theta)^{-1}|\underset{\sim}{x}))^{-1}.$$

The Bayes estimate of $r(t)$ under Kullback-Leibler loss is computed as follows:

$$\begin{aligned}\hat{r}(t)_{BKE} &= (E_{posterior}((e^{-\theta t})^{-1}|\underset{\sim}{x}))^{-1} \\ &= \left(\frac{(-t+b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n})}{(b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n})}\right)^{m+a}. \end{aligned} \tag{13}$$

3.3. Credible intervals for θ and $r(t)$

Using standard results for gamma distributions, the credible intervals of θ are constructed as follows. Since θ has gamma posterior distribution, the $(1-\tau)100\%$ credible interval of θ, say (I_L, I_U), can be calculated by solving the equations:

$$P(I_L < \theta < \infty) = 1 - \frac{\tau}{2}, \tag{14}$$

$$P(I_U < \theta < \infty) = \frac{\tau}{2}. \tag{15}$$

The $(1-\tau)100\%$ lower credible bound for θ is derived by solving Eq. (14) with respect to I_L, as follows:

$$\int_{I_L}^{\infty} \frac{\left(b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n}\right)^{m+a}}{\Gamma(m+a)} \theta^{m+a-1} e^{-\theta\left(b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n}\right)} d\theta = 1 - \frac{\tau}{2}.$$

By letting the substitution $u = \theta(b + \sum_{i=1}^{m}(1+r_i)x_{i:m:n})$ directly we get

$$\int_{\left(b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n}\right)I_L}^{\infty} u^{m+a-1}\, e^{-u}\, du = \left(1-\frac{\tau}{2}\right)\Gamma(m+a).$$

The incomplete gamma function $\Gamma(c,d)$, defined as

$$\Gamma(c,d) = \int_{d}^{\infty} x^{c-1}e^{-x}\, dx,\ c>0,\ d>0,$$

is employed to derive the relation:

$$\Gamma\left(m+a,\ (b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n})\, I_L\right) = \left(1-\frac{\tau}{2}\right)\Gamma(m+a). \tag{16}$$

Following the same approach from Eq. (15), the $(1-\tau)100\%$ upper credible bound for θ can be obtained by solving the following equation for I_U:

$$\Gamma\left(a+m,\ (b+\sum_{i=1}^{m}(1+r_i)x_{i:m:n})\, I_U\right) = \frac{\tau}{2}\Gamma(m+a). \tag{17}$$

The nonlinear Eqs. (16) and (17) prevents analytical solution, requiring appropriate numerical techniques to compute the values of I_L and I_U, respectively.

To find the $(1-\tau)100\%$ credible interval of the reliability function $r(t)$ we use the following algorithm:

Algorithm 1

- Step 1: Generate θ from $\pi(\theta|\mathbf{x})$ to get a sample of size M, $\{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Step 2: Evaluate $r(t) = e^{-\theta t}$ for each θ_i, $i = 1,2,\ldots,M$ to get $r_1(t), r_2(t), \ldots, r_M(t)$.
- Step 3: Order $r_1(t), r_2(t), \ldots, r_M(t)$ to get the order statistics $r_{(1)}(t), r_{(2)}(t), \ldots, r_{(M)}(t)$.
- Step 4: The $(1-\tau)100\%$ credible interval of the reliability function $r(t)$ is given by $[r_{[\frac{M\tau}{2}]}(t), r_{[M(1-\frac{\tau}{2})]}(t)]$, where $[x]$ represents the greatest integer less than or equal to x.

4. Simulation study

This section presents our main empirical contribution: a systematic comparison of the estimation methods described above. We emphasize that this is a comparative study examining the finite-sample behavior of different statistical approaches, each with its own theoretical foundation and practical considerations. We do not claim that one paradigm

is inherently superior to another, as maximum likelihood and Bayesian methods represent different philosophical approaches to statistical inference.

Our objective is to understand how these methods perform under various scenarios, which can provide practical guidance for researchers. The comparison focuses on finite-sample properties rather than establishing theoretical dominance.

Using simulated progressive Type II censored data under different censoring schemes, this research explores and assesses the performance and characteristics of maximum likelihood estimators and Bayes estimators. The following censoring schemes are considered:

- Case 1: $(n = 30, m = 10, r_1 = \ldots = r_4 = 5, r_5 = \ldots = r_{10} = 0)$
- Case 2: $(n = 30, m = 15, r_1 = \ldots = r_6 = 0, r_7 = r_8 = r_9 = 5, r_{10} = \ldots = r_{15} = 0)$
- Case 3: $(n = 30, m = 20, r_1 = \ldots = r_{18} = 0, r_{19} = r_{20} = 5)$
- Case 4: $(n = 40, m = 10, r_1 = \ldots = r_{10} = 3)$
- Case 5: $(n = 40, m = 20, r_1 = \ldots = r_{10} = 2, r_{11} = \ldots = r_{20} = 0)$

For the simulation study, we generated progressive Type II censored samples from an exponential distribution with parameter $\theta = 2$ using the technique given by Balakrishnan and Aggarwala (2000). In the Bayesian approach, we have computed the Bayesian estimates for the exponential parameter θ and reliability function with respect to both loss functions: squared error (L_{SE}) and Kullback-Leibler (L_{KL}). For conducting Bayesian analysis, the gamma density function with shape and scale parameters a and b, respectively, is assumed as a prior density of θ. To compute the different Bayes estimates under the two considered loss functions SELF and KELF, we assume two priors:
Prior 0: $a = 0.01$, $b = 0.01$ (weakly informative prior)
Prior 1: $a = 1$, $b = 0.5$ (informative prior with prior mean $E[\theta] = 2$)
For comparing the performance of the estimators of θ and $r(t)$, we have computed the MSEs based on $M = 1000$ iterations, where

$$MSE = \frac{\sum_{i=1}^{M} (\theta_i - \hat{\theta}_E)^2}{M}, \tag{18}$$

where $\hat{\theta}_E$ stands for the point estimate computed by one of the considered method.

The results of the maximum likelihood estimates (MLEs) and the Bayes estimates (BEs) of θ and $r(t)$ under SELF and KELF are presented in Tables 1 and 2 for Prior 0 and Prior 1, respectively. The mean square errors (MSEs) are shown in parentheses. Table 3 compares the average lengths (A.L) of credible intervals for θ and $r(t)$ between Priors 0 and 1, and presents the coverage percentage (C.P) for all schemes considered.

Table 1. Results of MLEs and BEs when Prior 0 is applied under SELF and KELF ($\theta = 2$, $r(1) = \exp(-2) = 0.1353$)

n	m	Censoring Scheme	MLE (MSE)	BE under SELF (MSE)	BE under KELF (MSE)
30	10	$(4\times5, 6\times0)$	θ: 2.1218 (0.5019)	2.1189 (0.4975)	1.9072 (0.4002)
			$r(1)$: 0.1465 (0.0070)	0.1707 (0.0081)	0.1217 (0.0070)
30	15	$(6\times0, 3\times5, 6\times0)$	θ: 2.0849 (0.3338)	2.0832 (0.3318)	1.9444 (0.2861)
			$r(1)$: 0.1417 (0.0040)	0.1585 (0.0045)	0.1247 (0.0040)
30	20	$(18\times0, 2\times5)$	θ: 2.0787 (0.2791)	2.0775 (0.2778)	1.9736 (0.2460)
			$r(1)$: 0.1389 (0.0030)	0.1518 (0.0033)	0.1260 (0.0030)
40	10	(10×3)	θ: 2.1847 (0.5092)	2.1817 (0.5045)	1.9637 (0.3833)
			$r(1)$: 0.1379 (0.0069)	0.1622 (0.0075)	0.1132 (0.0072)
40	20	$(10\times2, 10\times0)$	θ: 2.1078 (0.2945)	2.1065 (0.2931)	2.0012 (0.2543)
			$r(1)$: 0.1365 (0.0032)	0.1493 (0.0035)	0.1237 (0.0033)

Table 2. Results of BEs when Prior 1 is applied under SELF and KELF ($\theta = 2$, $r(1) = exp(-2) = 0.1353$)

n	m	Censoring Scheme	BE under SELF (MSE)	BE under KELF (MSE)
30	10	$(4\times5, 6\times0)$	θ: 2.1101 (0.3662)	1.9183 (0.2993)
			$r(1)$: 0.1630 (0.0059)	0.1173 (0.0054)
30	15	$(6\times0, 3\times5, 6\times0)$	θ: 2.0591 (0.2665)	1.9304 (0.2360)
			$r(1)$: 0.1579 (0.0039)	0.1258 (0.0035)
30	20	$(18\times0, 2\times5)$	θ: 2.0770 (0.2050)	1.9781 (0.1810)
			$r(1)$: 0.1489 (0.0029)	0.1242 (0.0027)
40	10	(10×3)	θ: 2.1024 (0.4489)	1.9112 (0.3702)
			$r(1)$: 0.1682 (0.0072)	0.1232 (0.0062)
40	20	$(10\times2, 10\times0)$	θ: 2.0351 (0.2147)	1.9382 (0.1974)
			$r(1)$: 0.1551 (0.0033)	0.1304 (0.0028)

Table 3. Results of average lengths (A.L.) and coverage probabilities (C.P.)

			Prior 0		Prior 1	
n	m	Censoring Scheme	A.L.	C.P.	A.L.	C.P.
30	10	$(4\times 5, 6\times 0)$	θ: 2.91	0.94	2.47	0.96
			$r(1)$: 0.31	0.94	0.30	0.94
30	15	$(6\times 0, 3\times 5, 6\times 0)$	θ: 2.09	0.96	2.01	0.95
			$r(1)$: 0.27	0.96	0.26	0.95
30	20	$(18\times 0, 2\times 5)$	θ: 1.81	0.96	1.76	0.96
			$r(1)$: 0.23	0.96	0.22	0.95
40	10	(10×3)	θ: 2.67	0.95	2.46	0.94
			$r(1)$: 0.32	0.95	0.31	0.93
40	20	$(10\times 2, 10\times 0)$	θ: 1.80	0.93	1.73	0.94
			$r(1)$: 0.24	0.93	0.23	0.94

The following interpretations and comments can be obtained from these tables:

(1) Table 1 shows that increasing m improves the performance of MLEs of θ and $r(t)$ in terms of biases and mean square errors (MSEs), which is consistent with asymptotic theory.

(2) Tables 1 and 2 show that the Bayesian estimates of the parameter θ, derived from both Prior 0 and Prior 1 while employing the two loss functions SELF and KELF, show different characteristics. Under the weakly informative Prior 0, the Bayesian estimates under SELF are identical to the MLEs, as expected. The Kullback-Leibler loss function produces estimates that are generally closer to the true value in terms of bias. The performance differences between methods become less pronounced as the sample size increases, which is consistent with asymptotic theory.

(3) Table 2 indicates that when using the informative Prior 1, both Bayesian methods show improved performance compared to the weakly informative prior, demonstrating the value of accurate prior information. The Kullback-Leibler loss function continues to show competitive performance relative to squared error loss.

(4) Table 3 shows that the credible intervals behave as expected: average lengths decrease as sample size increases, and coverage probabilities are close to the nominal 95% level. The informative prior generally produces shorter intervals while maintaining appropriate coverage.

These results illustrate the practical implications of different methodological choices rather than establishing the superiority of one approach over another. Each method has its place depending on the availability of prior information and the specific loss structure of the problem.

Real data example

The dataset considered by Lawless (1982), as presented in Table 1.1 on page 3, represents the duration (measured in minutes) until the breakdown of an insulating fluid situated

between electrodes at a voltage of 34 kV. The complete dataset, comprising 19 recorded breakdown times, properly ordered, is: 0.19, 0.78, 0.96, 1.31, 2.78, 3.16, 4.15, 4.67, 4.85, 6.50, 7.35, 8.01, 8.27, 12.06, 31.75, 32.52, 33.91, 36.71, 72.89.

In order to ascertain the suitability of this dataset for the exponential distribution, we employed the Kolmogorov-Smirnov (KS) test. The rate parameter of the exponential distribution was estimated using the maximum likelihood estimation (MLE) method from the entire dataset ($n = 19$), yielding $\hat{\theta} = 1/\bar{x} = 0.06966$, where $\bar{x}$ is the sample mean. The KS test statistic was calculated using the standard definition $D_n = \max|F(x) - \hat{F}(x)|$, where $F(x)$ is the empirical cumulative distribution function and $\hat{F}(x)$ is the theoretical cumulative distribution function of the exponential distribution with the estimated parameter $\hat{\theta} = 0.06966$. The test statistic ($D_n = 0.2463$) and the corresponding p-value (0.1993) were computed using Minitab Statistical Software (Version 15) based on the asymptotic Kolmogorov distribution formula:

$$P(D_n > \frac{z}{\sqrt{n}}) = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2z^2)$$

where D_n is the Kolmogorov-Smirnov statistic, $\frac{z}{\sqrt{n}} = 0.2463$, and $n = 19$. Graphical diagnostics including empirical and theoretical cumulative distribution functions (CDFs) and the quantile-quantile (Q-Q) plot were created using R software with the `ggplot2` package and are presented in Figures 1 and 2. With p-value $= 0.1993 > 0.05$, we fail to reject the null hypothesis at the $\alpha = 0.05$ significance level, indicating that the exponential distribution is appropriate for modeling this dataset.

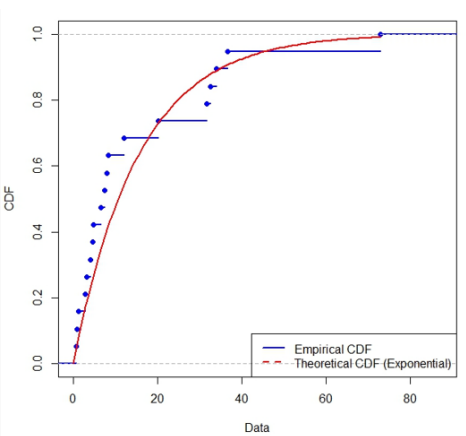


Figure 1. Empirical and theoretical cumulative distribution functions for the breakdown voltage data. The close agreement between the empirical CDF (step function) and theoretical exponential CDF (smooth curve) supports the exponential distribution assumption

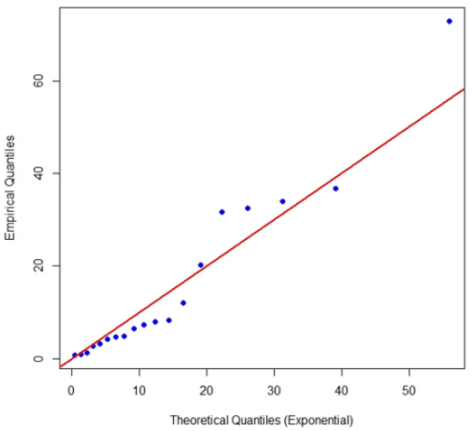


Figure 2. Q-Q Plot: Empirical vs. Theoretical Quantiles (Exponential). The points approximately follow the diagonal line, indicating good fit to the exponential distribution, with some deviation in the upper tail as commonly observed in reliability data

To demonstrate the methods developed in this study, we analyze this dataset and extract $m = 12$ progressive Type II sample from this data using the scheme:

i	1	2	3	4	5	6	7	8	9	10	11	12
r_i	1	0	0	2	0	0	1	0	1	0	0	2
$x_{i:m:n}$	0.19	0.96	1.31	2.75	4.67	4.85	6.50	8.01	8.27	31.75	32.52	33.91

This progressive censoring scheme works as follows: at the first failure (0.19), we remove 1 additional item (0.78); at the fourth failure (2.78), we remove 2 items (3.16, 4.15); at the seventh failure (6.50), we remove 1 item (7.35); at the ninth failure (8.27), we remove 1 item (12.06); and at the final failure (33.91), we remove the remaining 2 items (36.71, 72.89). This accounts for all 19 original observations: 12 observed failures + 7 censored items = 19 total.

For this dataset, we applied our methods and obtained the maximum likelihood estimates of θ and $r(t)$. All computations were performed using Mathematica 12.0 software. The MLE was computed using Equation (5) from Section 2. The Bayesian estimates were obtained using Equations (10) and (11) from Section 3 under both squared error and Kullback-Leibler loss functions. The credible intervals were computed numerically by solving Equations (16) and (17) using Mathematica's `FindRoot` function with precision settings `WorkingPrecision`→16 and `MaxIterations`→1000. The estimates were determined to be approximately 0.0536 and 0.9479, respectively, and 95% confidence intervals for θ and $r(t)$ were determined to be respectively (0.0233, 0.0839) and (0.9191, 0.9766). Since $h(t) = \theta$ for the exponential distribution, $\hat{h}(t)_{MLE} = 0.0536$. Furthermore, as illustrated in Table 4, the Bayes estimates along with the corresponding 95% credible intervals for both Priors 0 and 1 are presented. The estimates from different methods show reasonable variation, reflecting the different loss functions and prior specifications and fall within

the established 95% credible intervals. When Prior 1 is employed, the estimates reflect the influence of the informative prior, showing how prior knowledge affects the results.

Table 4. Estimation results for the real data with 95% credible intervals
$((\theta = 0.0536, r(1) = exp(-0.0536) = 0.9478))$

Scheme: n=19 ,	m=12	$(1,2\times 0,2,2\times 0,1,0,1,2\times 0,2)$		
		SELF	*KELF*	95% credible interval
Prior 0	θ	0.0536	0.0492	$(0.0277, 0.0879)$
	$r(t)$	0.9479	0.9476	$(0.9158, 0.9727)$
Prior 1	θ	0.0579	0.0535	$(0.0308, 0.0934)$
	$r(t)$	0.9438	0.9436	$(0.9109, 0.9696)$

The results demonstrate the practical application of the different estimation methods. The MLE and Bayesian estimates under weakly informative priors are quite similar, while the informative prior shows its influence on the estimates. The Kullback-Leibler loss function consistently produces somewhat different estimates, reflecting its different optimization criterion compared to squared error loss.

Following Lawless (1982)'s framework for Breakdown Voltage Data, Figures 3 and 4 visually compare the performance of all estimation methods.

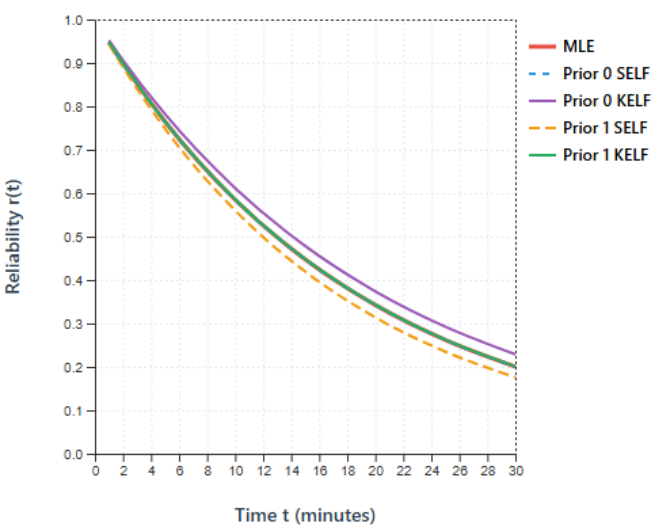


Figure 3. Reliability Function $r(t)$ Estimates under Different Loss Functions and Priors

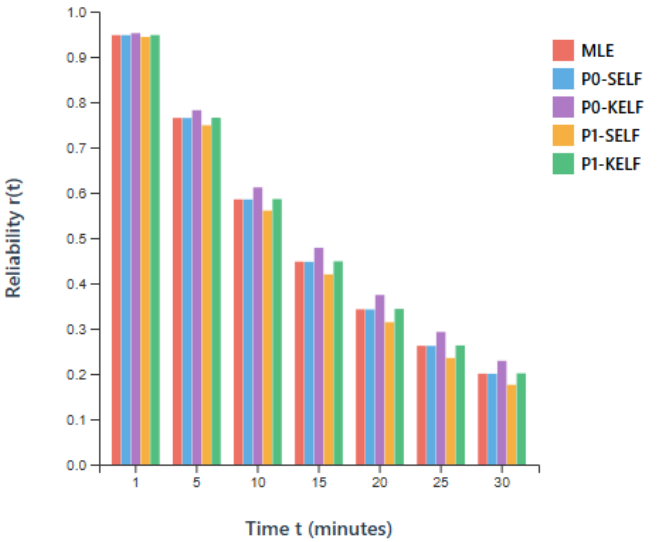


Figure 4. Reliability Estimates at Selected Time Points by Estimation Method

5. Conclusion

This paper provides a comparative analysis of frequentist and Bayesian estimation methods for the parameter, reliability and hazard functions of the exponential distribution using progressive Type II censored data. The theoretical methods presented are well established in the statistical literature; our contribution lies in the systematic empirical comparison of these approaches.

The maximum likelihood estimate and Bayes estimates of the parameter, reliability and hazard functions have been computed using standard procedures implemented via "Mathematica 12" software. Two loss functions have been considered: squared error and Kullback-Leibler for Bayesian computation under two priors: weakly informative and informative. The performance of all estimators has been assessed through mean square errors.

The key contributions include:

- under weakly informative priors, Bayesian estimates with squared error loss and maximum likelihood estimators show similar performance, confirming theoretical expectations about their asymptotic equivalence,
- the Kullback-Leibler loss function provides a meaningful alternative with different optimization characteristics,
- informative priors can improve estimation performance when they accurately reflect prior knowledge,
- the choice between methods should be guided by the specific context, available prior information, and loss structure rather than claims of universal superiority,

- proper acknowledgment that $h(t) = \theta$ for exponential distributions, avoiding redundant calculations.

The analysis of real reliability data illustrates the practical application of these methods and shows reasonable agreement among different approaches.

Acknowledgements

The authors acknowledge that the theoretical foundations presented are well established in the statistical literature. We thank the reviewers for their constructive comments that helped clarify the scope and contribution of this comparative study.

References

Abu Awwad, R. R., Bdair, O. M. and Abufoudeh, G. K., (2019). Statistical Inference of Exponential Record Data under Kullback-Leibler Divergence Measure. *Statistics in Transition*, 20(2), pp. 1–14.

Abufoudeh, G., Bdair, O. and Abu Awwad, R., (2019). Bayesian estimation under Kullback-Leibler divergence measure based on exponential data. *Investigacion Operacional*, 40(1), pp. 61–72.

Aggarwala, R., (1996). *Advances in life testing: Progressive censoring and generalized distribution*. PhD thesis. McMaster University.

Almetwally, E. M., Jawa, T. M., Sayed-Ahmed, N., Park, C., Zakarya, M., and Dey, S., (2023). Analysis of unit-Weibull based on progressive Type-II censored with optimal scheme. *Alexandria Engineering Journal*, 63, pp. 321–338.

Alshenawy, R., Al-Alwan, A., Almetwally, E. M., Afify, A. Z., and Almongy, H. M., (2021). Progressive Type-II censoring schemes of extended odd Weibull exponential distribution with applications in medicine and engineering. *Mathematics*, 8(10), 1679.

Balakrishnan, N., Aggarwala, R., (2000). *Progressive censoring: Theory, methods and applications*. Boston, MA: Birkhäuser.

Balakrishnan, N., Cramer, E., (2014). *The art of progressive censoring: Applications to reliability and quality*. New York: Springer.

Dey, S., Elshahhat, A., and Nassar, M., (2022). Analysis of progressive Type-II censored gamma distribution. *Computational Statistics*, 38, pp. 481–508.

Kim, C., Jung, J., and Chung, Y., (2011). Bayesian estimation for the exponentiated Weibull model under Type II progressive censoring. *Statistical Papers*, 52(1), pp. 53–70.

Kullback, S., Leibler, R.A., (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), pp. 79–86.

Kundu, D., Raqab, M. Z., (2012). Bayesian inference and prediction for a Type-II censored Weibull distribution. *Journal of Statistical Planning and Inference*, 142(1), pp. 41–47.

Lawless, J. F., (1982). *Statistical models and methods for lifetime data*. New York: John Wiley & Sons.

Pradhan, B., Kundu, D., (2009). On progressively censored generalized exponential distribution. *Test*, 18(3), pp. 497–515.

Rasheed, M., (2023). Analyzing applications and properties of the exponential continuous distribution in reliability and survival analysis. *Journal of Positive Sciences*, 3(1), pp. 15–23.

Ren, H., Hu, X., (2023). Estimation for inverse Weibull distribution under progressive Type-II censoring scheme. *AIMS Mathematics*, 8(10), pp. 22808–22829.

Sapkota, L. P., Bam, N., and Kumar, V., (2025). A new exponential family of distributions with applications to engineering and medical data. *Scientific Reports*, 15(1), 33649.

Wu, M., Gui, W., (2021). Estimation and prediction for Nadarajah-Haghighi distribution under progressive Type-II censoring. *Symmetry*, 13(6), 999.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 145–160, https://doi.org/10.59139/stattrans-2026-021
Received – 10.04.2025; accepted – 20.01.2026

Effectiveness of bankruptcy prediction models constructed for differently selected diagnostic variables

Bernard Kokczyński[1], **Dorota Witkowska**[2]

Abstract

The purpose of the article is to compare the effectiveness of discriminant models for predicting corporate bankruptcy, constructed using different methods of diagnostic variables selection. We compared several methods, such as arbitrary selection, the forward stepwise method applied after the initial selection of variables by means of a significance test of differences between group averages (further called the two-step method), the Hellwig method, the *tstatistics* and the backward stepwise method. We also assessed the models' accuracy in terms of the synthetic measure. It was constructed by applying eight measurements of the classification effectiveness, such as the values of Wilks' lambda statistic and AUC together with the percentage of correctly identified companies, i.e. total, bankrupts and non-bankrupts in training and testing sets. The results show that the backward stepwise method and the two-step method generate models with the highest accuracy of classification. In addition, the study found that Wilks' lambda statistic is not a good approximation of the classification abilities of bankruptcy models. The contribution of our paper is a comparative methodological study, focusing on the impact of alternative diagnostic variable selection techniques on the linear discriminant function accuracy used to bankruptcy prediction.

Key words: diagnostic variable selection, linear discriminant function, bankruptcy prediction, Euclidean distance.

1. Introduction

Shi and Li (2019) list the currently popular bankruptcy prediction methods, i.e.: logistic regression, classification trees and artificial neural networks. In contrast, Kopczyński (2022) emphasizes that linear discriminant analysis still retains its relevance and application in predicting corporate bankruptcy, despite the development of

[1] Department of Corporate Financial Management, Faculty of Management, University of Lodz, Lodz, Poland. E-mail: bernard.kokczynski@gmail.com. ORCID: https://orcid.org/0000-0002-9379-0376.

[2] Department of Corporate Financial Management, Faculty of Management, University of Lodz, Lodz, Poland. E-mail: dorota.witkowska@uni.lodz.pl. ORCID: https://orcid.org/0000-0001-9538-9589.

newer methods. However, regardless of the bankruptcy prediction method used, the key issue remains the choice of diagnostic variables[3].

According to Grzybowska and Karwański (2023), explanatory variables are selected to reflect expected influences based on theory, previous research, and local context in temporal and spatial dimensions. Gruszczyński (2012) points out that the process of selecting diagnostic variables in practice is often limited to looking for such variables that intuitively appear to be causal for the explained variable, which is then subject to empirical verification. Witkowska (2023) emphasizes that independent variables should comprehensively describe the most important aspects of the phenomenon under analysis, providing general information about individual units rather than unique data. Many economic variables are characterized by high mutual correlation, which leads to information redundancy. In such a case, there is a need to reduce the number of diagnostic variables, which can also be caused by a limited sample size or a large number of estimated parameters.

According to Woo Ahn (2016), the problem of selecting the optimal set of variables is one of the most active areas of research in statistics. This phenomenon is reflected in numerous studies on the development of new methods of variable selection. For example, Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996), Smoothly Clipped Absolute Deviation (SCAD) proposed by Fan and Li (2001), Iterative BIC-Based Variable Selection for Model-Based Clustering (IBVS-MBC) proposed by Raftery and Dean (2006), Hierarchical Distance-Based Variable Grouping and Selection (HDB-VGS) proposed by Korzeniewski (2016), Supervised Factor Models (SFM) using constraint optimization proposed by Wang et al. (2023), One Covariate at a Time Multiple Testing (OCTMT) proposed by Chudik et al. (2024), to name a few.

Despite the development of modern machine learning and penalized regression methods, linear discriminant analysis is still widely used in applied bankruptcy studies, particularly in Central and Eastern European research and practice.

The literature provides a number of proposals for the selection of diagnostic variables. Altman (1968) sought to determine the relative contribution of each independent variable to the model, assessed the correlations between explanatory variables, checked the predictive accuracy of the models, and subjected them to judgment as an analyst. Gajdka and Stos (1996) selected discriminating variables by analyzing the correlations between the value of the dependent variable Y and the values of 20 arbitrarily chosen financial indicators, selecting those most correlated with the value of Y. Mączyńska and Zawadzki (2006) created an initial set of 45 diagnostic variables based on the literature and assumptions about the mechanism of financial deterioration of companies. The

[3] In the article, phrases such as discriminant, diagnostic, independent or explanatory variables will be used interchangeably.

selection process included an evaluation of the discriminatory ability of the variables on the basis of: the test of differences of the financial indicators characterizing the groups, the Mahalanobis distance, Wilks' lambda statistic, the accuracy of the classification obtained from the application of single-indicator discriminant models, and the value of the correlation coefficient between the variables. The final selection of indicators also considered expert judgment. Herman (2018) used three variable selection techniques: the method called by him *tstatistics*, the selection of independent variables based on the Spearman correlation coefficient between them and the dependent variable, and the stepwise forward selection of variables, taking into account the variables that reduce Wilks' lambda statistics[4] the most. In the collective work of Valaskov et al. (2023), the results of the significance test of differences between group averages were used to select discriminant variables.

The aim of the paper is to investigate how different variable selection strategies affect model quality which is understood as effective classification of companies. The empirical study compares linear discriminant models constructed on the basis of diagnostic variables distinguished using various techniques for their selection, and indicating which variable selection techniques lead to the most effective linear discriminant models from the point of view of multivariate evaluation of the model validity.

In assessing their effectiveness, eight commonly used accuracy measures were considered, i.e. Wilks' lambda statistic, the value of AUC, the percentage of properly recognized bankrupts, non-bankrupts and all analyzed enterprises in the training and testing samples. The model with the lowest value of Euclidean distance from the defined pattern was considered the best. In other words, the aim of analysis is to point out which method of variable selection specified the most effective bankruptcy prediction models. An additional purpose of the study is to find out whether Wilks' lambda statistic is a good approximation of the discriminatory power of the models.

2. Research methods

In the study a linear discriminant function was used to identify bankrupt and going concern enterprises. Verification of the ability to differentiate between groups of a discriminant model is usually done by using Wilks' lambda statistic, which is calculated as the ratio of the determinant of the within-group variance and covariance matrix to the determinant of the total variance and covariance matrix (Pociecha et al., 2014; Herman, 2018). Many researchers point out that Wilks' lambda statistic can be used as a measure to evaluate the discriminatory ability of classification models. Kopczyński (2022) emphasizes that the value of this statistic makes it possible to assess the effectiveness of the

[4] Before applying the selection techniques, the author eliminated explanatory variables that were highly correlated among themselves (i.e. Spearman correlation coefficient > 0.90).

entire model in separating groups. In a similar vein is Herman (2018), who uses a step-wise selection of forward variables, including in the model those variables that most significantly cause Wilks' lambda value to fall. Mączyńska and Zawadzki (2006) point out that a lower Wilks' lambda value means better discriminatory ability of variables. Similar conclusions are found in the work of Pociecha et al. (2014), where a range of Wilks' lambda values from 0 (best classification ability) to 1 (no discriminatory power) is defined.

In order to verify the discriminatory power of the model, it is testified whether Wilks' lambda is significantly less than unity, i.e. a pair of hypotheses is posed:

$$H_0: \lambda = 1$$
$$H_0: \lambda < 1$$

the test statistic has χ^2 distribution and it is of the form (Jagiełło, 2013):

$$\chi^2 = -\left(N - \frac{k+p}{2} - 1\right) \ln(\hat{\lambda}) \tag{1}$$

where: $\hat{\lambda}$ is the value of Wilks' lambda statistic estimated from the sample.

Evaluation of the discriminant model efficiency is based on an analysis of prediction accuracy, which can be carried out on a training sample, used to estimate model parameters, or on a test sample, not used in the estimation process, to assess the accuracy of classification. Ptak-Chmielewska (2012) emphasizes that classification using the model distinguishes between an overall classification error and two types of partial errors if the consequences of misclassification into one group are "more costly" than for the other group. When considering the problem of bankruptcy prediction, the classification error of the first type E_1 occurs if the method used misclassifies a bankrupt as a non-bankrupt (2), and the error of the second type E_2 occurs if the model recognizes a non-bankrupt as a bankrupt (3). The overall classification error E is the sum of errors of the first and second type with respect to the total number of observations, reflecting the total percentage of misclassified cases in the analyzed model (4). Errors[5] are most often expressed as a percentage:

$$E_1 = \frac{Fb}{Tb+Fb} \cdot 100 \tag{2}$$

$$E_2 = \frac{Fn}{Tn+Fn} \cdot 100 \tag{3}$$

$$E = \frac{Fb+Fn}{Tb+Tn+Fb+Fn} \cdot 100 \tag{4}$$

where: Fb, Fn - the number of misidentified bankrupts and non-bankrupts, respectively, Tb, Tn – the number of correctly identified bankrupts and non-bankrupts,

[5] Most analyses of the quality of decision rules are based on a comparison of the classification errors of different classifiers. Classification error is not an ideal indicator of a classifier's predictive value. This measure does not consider different a priori probabilities of objects belonging to classes and the problem of unbalanced classes. Misztal (2014) points out that these problems can be partially solved by taking into account different weights or costs of misclassifications, if they can be estimated.

respectively. Thus, the classification efficiency is indicated by the percentage of correctly recognized objects, which completes the classification error to 100%.

Alternatively, the analysis of the classifier quality can be broadened by using additional measures of quality. In this case, the most commonly used are sensitivity (the ability of a classifier to correctly identify bankrupts) and specificity (the ability of a classifier to correctly recognize non-bankrupts). Misztal (2014) emphasizes that these measures are crucial in constructing Receiver Operating Characteristic (ROC) curves. The area under the ROC curve, referred to as AUC, is an integrated measure of a classifier's ability to distinguish between bankrupts and non-bankrupts. An AUC value of 1 indicates a perfect model, while an AUC value of 0.5 indicates a model operating at random. Kopczyński (2022) limits the comparison of two ROC curves solely to comparing the AUC values, omitting their graphical representation. In this case, determining the optimal AUC value becomes crucial. Harańczyk (2010) points out that an AUC value at a certain level is not unambiguously good or bad. It depends on the field and the specifics of the problem under consideration. According to Kopczyński (2022), on the other hand, the AUC values above 0.7, in the context of bankruptcy prediction models, are considered a sign of satisfactory discriminatory ability of the model, while values above 0.6 indicate sufficient discriminatory ability.

2.1. Techniques for selecting discriminating variables for models

The model construction started by applying an arbitrary selection of variables. In the first model, one financial indicator was selected from each of the following groups: liquidity indicators, operating efficiency indicators, debt indicators and profitability indicators. The second model was constructed using a tool based on the generative artificial intelligence Chat GPT (version 3.5), which selected variables based on an implemented list of primary diagnostic variables and an analysis of available literature sources. In both cases, variable selection was based on their "popularity" in the literature, rather than on the statistical properties of the variables derived from the data analysis.

The next technique used for the selection of independent variables, was a two-step method consisting of a test of significance of differences between group means, performed with the SPSS program[6], by one-way ANOVA analysis of variance and consisting of stepwise forward selection of variables. In other words, the following null hypothesis was verified:

$$H_0: E(X_1) = E(X_2) \tag{5}$$

where: $E(X_1), E(X_2)$ denotes the expected values of the random variables derived from the first and second groups, respectively. The test statistic expresses the ratio of the

[6] The study used the SPSS program Statistics version 29.0.0.

between-group variance to the within-group variance, which in the case of two groups can be written as (Malarska, 2005):

$$F = \frac{\sum_{j=1}^{2}(\bar{x}_j - \bar{x})^2 n_j}{\frac{1}{n-2}\sum_{i=1}^{n_j}\sum_{j=1}^{2}(x_{ij} - \bar{x}_j)^2} \tag{6}$$

where: $\sum_{j=1}^{2}(\bar{x}_j - \bar{x})^2 n_j$ is the intergroup variability expressing the variation of the two group averages around the overall average, $\sum_{i=1}^{n_j}\sum_{j=1}^{2}(x_{ij} - \bar{x}_j)^2$ is the within-group variability expressing the variation of individual observations around group averages, $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$ are group averages calculated for each of the distinguished groups (j=1, 2), $\bar{x} = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{2} x_{ij}$ is the overall average calculated from all observations, n_j is the number of observations in each group, $n = \sum_{j=1}^{2} n_j$ is the number of total observations, x_{ij} is the observation of a distinguished feature in the i-th object located in the j-th group.

In the first stage of the model building, the variables most differentiating the two groups of companies (bankrupt and non-bankrupt) were identified based on Wilks' lambda and Fisher-Snedecor statistics implemented in the SPSS package. The study assumed a significance level of 0.05. In the second stage, models were constructed using the stepwise progressive method. In this variable selection procedure, at each step, the variable that most contributes to improving the model's classification performance was added to the model, starting with the most important determinant of the phenomenon under study (as in Pociecha et al. (2014) and Herman (2018)).

The further method used in the study, was so-called Hellwig's method of selecting diagnostic variables extensively described in Witkowska's work (2023). One of the formal procedures for selecting diagnostic features, in the analysis of multivariate objects, is based on the **R** correlation matrix, which analyses potential diagnostic features:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1K} \\ r_{21} & 1 & r_{23} & \dots & r_{2K} \\ \dots & \dots & \dots & \dots & \dots \\ r_{K1} & r_{K2} & r_{K3} & \dots & 1 \end{bmatrix} \tag{7}$$

where: for each pair of variables x_i and x_j (i, j = 1, 2, …, K), r_{ij} is Pearson's linear correlation coefficient.

Witkowska (2023) emphasizes that the criterion for the selection of diagnostic variables is based on the formally determined or arbitrarily adopted critical value of the correlation coefficient r^*, which in this study was adopted at the level of 0.8. On this basis, clusters, i.e. subsets of features for which the absolute values of correlation coefficients do not exceed the critical value, are distinguished from the matrix. Within the clusters, one so-called central feature and a number of so-called satellite features are

determined. The features in the clusters are called systemic, while the remaining features are referred to as isolated. Central and isolated features may form a set of diagnostic features.

In order to distinguish individual features, the sum of the elements of each column of the correlation matrix **R** is calculated. Then the column p for which the sum of absolute values is the largest is searched for, and the feature corresponding to this column is considered the central feature. In the highlighted column, elements that satisfy the inequality are identified (Witkowska, 2023):

$$\left|r_{pj}\right| \geq r^* \tag{8}$$

The features corresponding to the highlighted rows are considered satellite features. The highlighted columns and rows are then removed from the correlation matrix **R**, resulting in a reduced correlation matrix. This procedure is repeated until the set of features is exhausted, resulting in the identification of more cluster points and the creation of new reduced correlation matrices.

The next method of the variable selection used in the study is the method called by Herman (2018) *t-statistics*. According to this procedure, 5 variables were selected that had the highest absolute values of the Student's t-statistic, in a test comparing the mean value of indicators in the study groups, for independent samples (5). The test statistic is calculated for samples with heterogeneous variances as (Wiktorowicz et al., 2020):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{9}$$

where: $\bar{x}_1; \bar{x}_2$ are averages from samples, S_1^2 ; S_2^2 are the variances from samples, $n_1; n_2$ are sample sizes. The statistic (9) has t-Student distribution with $\left(\frac{1}{2} + \frac{S_1^2 S_2^2}{S_1^2 + S_2^2}\right)(n_1 + n_2 - 2)$ degrees of freedom. The rejection of the null hypothesis occurs if *p-value* < 0.05. This approach was supplemented by verification of information redundancy, carried out using Pearson's correlation matrix, where 0.8 was taken as the critical value.

The last method used to select diagnostic variables for discriminant models, which was implemented in SPSS software, was the backward stepwise method. Once all variables are entered, those that meet the elimination criteria are removed until all variables meeting the criteria are exhausted. The process begins with the full set of explanatory variables assumed initially in the theoretical model. Then, in each step, the variable most weakly related to the phenomenon under study is eliminated, and the process continues until only significant variables remain. Once removed, the variable is not included in the model again in further stages of the analysis (Wiktorowicz et al., 2020). In SPSS, independent variables were included or removed from the model based on the value of the Fisher-Snedecor statistic (6).

2.2. Models Evaluation

In the study several methods of discriminant model accuracy assessment are applied, such as: Wilks' lambda value, AUC value, classification efficiency of bankrupts, non-bankrupts and overall in the training and the test samples, assessed as the percentage of properly recognized firms. Values of these measurements were used to construct the synthetic measures in the k-dimensional space as Euclidean distance from the pattern, which is defined as follows:

$$\left[\sum_{j=1}^{k}\left(z_{ij}-z_{0j}\right)^2\right]^{\frac{1}{2}} \tag{10}$$

where k is the number of variables used for the models' assessment (k = 2, 4, 6 or 8), z_{ij} and z_{0j} are the standardized[7] variables in the i-th (i =1, 2,…, 8) model and the pattern, respectively. Equal weighting of all accuracy measures was adopted in order to avoid introducing subjective preferences and to ensure a neutral evaluation of model performance.

The Euclidean-distance-based synthetic measure is intended as a decision-support tool rather than a purely statistical construct. It allows for a simultaneous, multivariate assessment of the model performance when several accuracy measures are considered jointly. Such an approach reflects the practical situation faced by analysts, who must evaluate classification models under multiple, often conflicting, criteria instead of relying on a single performance indicator.

In the study two types of patterns were defined – the ideal model and the best hypothetical model. Pattern values for the ideal model were assumed as extremal values showing the perfect discrimination strength or excellent classification ability of the model, whereas the best hypothetical model was characterized by the lowest Wilks' lambda value, highest classification efficiency and highest AUC value among all considered models. Each model was assigned ranks in each category, where the highest (i.e. rank 1) was given to the model for which Euclidean distance from the pattern was the lowest.

To find out which accuracy measurements give similar information, all models were ranked from the best to the worst according to each measurement. Then the Spearman correlation coefficients between the ranks of models ordered due to Wilks' lambda statistics and ordered according to other characteristics of the models were calculated.

3. Data and Results

The analyses were based on data concerning 416 Polish non-financial non-public enterprises. The companies formed a choice-based, matched, and balanced sample.

[7] Standardization formula: $z_{ij} = (x_{ij} - \bar{x}_j)/S_j)$ where: x_{ij} denotes the value of the model assessment measurement, $\bar{x}_j$ and S_j are the average value and standard deviation of the j-the variable (j = 1, 2,…, k), respectively.

Bankrupts were defined as companies that had filed a bankruptcy petition with the court in the years 2019-2022. Non-bankrupts were defined as companies that had not filed a bankruptcy petition with the court, they operated in the same industry and had a similar size of annual revenues as companies considered bankrupt.

Firms included in the research belong to three economic sectors in equal proportions: trade, manufacturing and services. The source of the data used in the study was the Emerging Markets Information Service (EMIS). The extracted financial data of the companies from their financial statements was used to calculate 56 financial indicators, which are the most commonly used in the literature in the context of bankruptcy prediction. These were indicators of liquidity (denoted by the letter P), debt (- Z), operating efficiency (- S), profitability (- R), dynamics (-D), and company size (- W). Based on these, a preliminary set of discriminant features was created from which variable selection was carried out for the construction of discriminant models. Models were estimated based on a training sample in which there were 332 objects (80%), and verified on a test sample with 84 objects (20%). Objects for the test sample were selected by random drawing, using a random number generator in Excel[8]. Companies considered bankrupt in each industry were drawn, and their counterparts in the non-bankrupt group were paired. In order to maintain the representativeness of economic sector participation in the study, the drawing was carried out separately in trade, manufacturing and services. The results of the sampling from each industry were then summed to form a balanced teaching and testing sample, which reflects the industry structure of the available database.

In the course of the study, all potential 56 diagnostic variables were analyzed for their use in building bankruptcy prediction models. For this purpose, the methods for selecting independent variables discussed in subsection 2.1 were used. Table 1 provides a description of each set of diagnostic variables, along with information on the method used to select them.

Among the variables presented in Table 1, the most frequently occurring variable was W03 (logarithm of assets), which appeared in six different sets. The frequently occurring variables were P02 (quick ratio), Z02 (debt-to-equity ratio), R10 (return on average current assets), W02 (logarithm of asset structure) in three sets, and Z04 (share of equity in total assets), R02 (operating profitability to total assets), R04 (net margin), R09 (average gross profit to assets), S01 (asset turnover ratio), S07 (conversion of receivables), which were present in two sets. Variables such as: Z03 (long-term debt to equity), Z05 (proportion of current liabilities to total assets), R01 (EBITDA), R05 (return on equity), R07 (operating profit to assets), R12 (operating margin after depreciation), R13 (return on current assets), S03 (ratio of inventory to operating

[8] The use of Excel was limited to generating random numbers and does not affect the randomness or validity of the sampling procedure.

expenses), S04 (ratio of inventory to sales revenue), S09 (inventory turnover in days), S18 (averaged net cash conversion cycle), S19 (averaged coverage of current liabilities by operating expenses), S20 (averaged asset turnover ratio), W01 (asset structure), D01 (revenue dynamics), D02 (equity dynamics) appeared in only one of the sets.

Table 1. Sets of financial variables selected using various selection methods

Selection method	Symbol of the variable set	List of variables in each set
arbitrary own choice	A	P02, R04, S01, Z02
arbitrary GPT Chat selection	B	P02, R04, S07, W03, Z02
two-step method (variant I)	C	R02, R09, W03
two-step method (variant II)	D	R10, W03
Hellwig's selection method diagnostic variables (central variables)	E	P02, R09, R10, R12, R13, S01, S04, S18, S19
Hellwig's selection method diagnostic variables (isolated variables)	F	D01, D02, R01, R02, R05, S03, S07, S09, W01, W02, W03, Z02, Z03, Z04
tstatistics	G	R07, W02, W03, Z04, Z05
backward stepwise method	H	R10, S20, W02, W03

Source: own work.
Note: D - dynamics, P - liquidity, R - profitability, S - operating efficiency, W - company size, Z - debt.

Table 2. Model estimation results

Discriminant function	Wilks' lambda	AUC
A = 0.001·P02 + 0.007·Z02 + 0·R04 + 0.093·S01 -0.435	0.981	0.500
B = 0·P02 - 0.002·Z02 + 0·R04 + 0·S07 + 1.068·W03 -3.413	0.867	0.723
C = 0.217·R02 + 1.026·W03 + 0.032 ·R09 -3.322	0.870	0.722
D = 1.043·W03 + 0.003·R10 -3.357	0.863	0.730
E = 0·S18 + 0.271·R09 – 0.001·P02 + 0.028·R13 + 0·R14 – 0.018·R16 – 0.181 ·S04 + 0·S19 + 0.002·R10 + 0.243	0.952	0.715
F = -0.002·Z02 + 0.078·Z03 + 0·R01 + 0.129·R02 + 0·R05 – 0.061·S03 + 0·S07 – 0.051·W01 + 0.358·W02 + 0·D01 + 0.034·Z04 + 0.779·W03 – 0.001·D02 + 0 ·S09 -2.133	0.798	0.763
G = 0.873·W03 + 0.074·Z04 + 0.286·W02 + 0.245·R07 + 0.045·Z05 - 2.606	0.856	0.718
H = 0.309·W02 + 0.926·W03 + 0.003·R10 + 0.081·S20 - 2.971	0.831	0.730

Source: own work.
Note: The zero values of some model parameters are due to the rounding to three decimal places used.

On the basis of selected financial variables, discriminant models were estimated. Parameter estimates together with the values of Wilks' lambda statistic and AUC are presented in Table 2. The lowest Wilks' lambda value was obtained by the model F

(0.798), and the highest by the model A (0.981). All Wilks' lambda values for the presented models are relatively high, which may indicate their limited discriminatory power for classifying objects. The obtained AUC values (except for the model A) were greater than 0.7, which indicates the satisfactory discriminatory power of the models.

Table 3. Classification performance

Model	Classification effectiveness in training set			Classification effectiveness in test set		
	bankrupts	non-bankrupts	total	bankrupts	non-bankrupts	total
A	32.70%	76.20%	54.45%	37.50%	70.00%	53.75%
B	65.50%	72.60%	69.05%	70.00%	72.50%	71.25%
C	66.70%	77.40%	72.05%	70.00%	75.00%	72.50%
D	68.50%	77.40%	72.95%	70.00%	80.00%	75.00%
E	41.10%	91.10%	66.10%	42.50%	92.50%	67.50%
F	65.50%	76.80%	71.15%	72.50%	67.50%	70.00%
G	64.90%	73.80%	69.35%	67.50%	65.00%	66.25%
H	63.70%	79.20%	71.45%	67.50%	75.00%	71.25%

Source: own work.

Table 4. Assigned tied ranks associated with model features

Model	Classification effectiveness in training set			Classification effectiveness in test set			Wilks' lambda	AUC
	bankrupts	non-bankrupts	total	bankrupts	non-bankrupts	total		
A	8	6	8	8	6	8	8	8
B	3.5	8	6	3	5	3.5	5	4
C	2	3.5	2	3	3.5	2	6	5
D	1	3.5	1	3	2	1	4	2.5
E	7	1	7	7	1	6	7	7
F	3.5	5	4	1	7	5	1	1
G	5	7	5	5.5	8	7	3	6
H	6	2	3	5.5	3.5	3.5	2	2.5

Source: own work.

The classification efficiency of the models (Table 3) was verified using data from the test sample, which includes data not used in the estimation of the parameters of the discriminant function, and compared with the classification results obtained for the training sample. It was noted that higher classification performance of bankrupts was obtained in the test sample when comparing it to that obtained for the training sample. An inverse relationship is found in the context of classifying non-bankrupts (except for models D and E). In the training and test samples, the highest overall classification

efficiency was distinguished by model D. The most effective in classifying bankrupts was model F, and non-bankrupts was model E (Table 4).

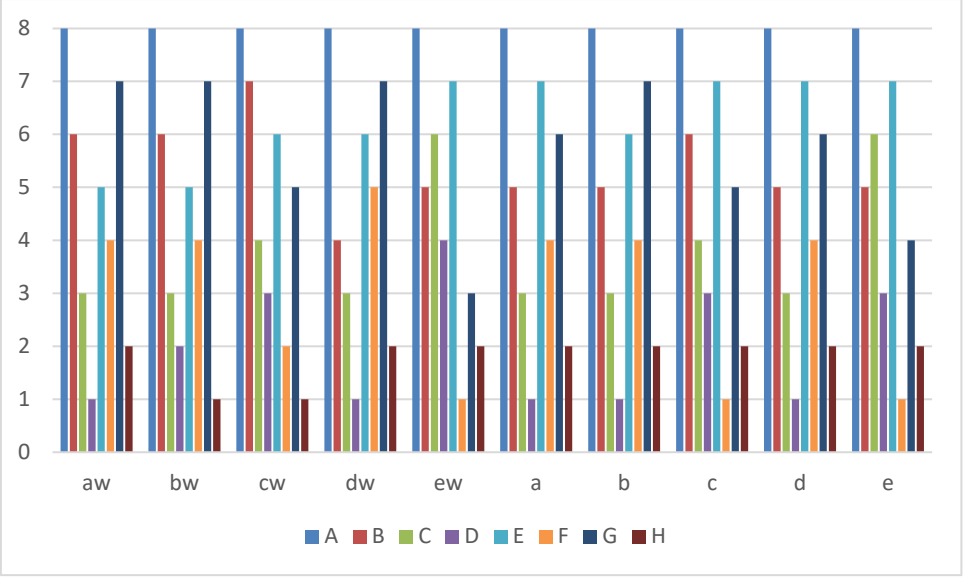
The assigned tied ranks were used to calculate the Spearman correlation coefficients between positions of models obtained due to values of Wilks' lambda statistic and rankings made according to other characteristics of the models (Table 5). It is noticeable that strong correlation is visible only for rankings made due to AUC.

Table 5. Spearman's rank correlation coefficients for classification model features

Classification effectiveness in training set			Classification effectiveness in test set			AUC
bankrupts	non-bankrupts	total	bankrupts	non-bankrupts	total	
0.3631	-0.0417	0.5238	0.6131	-0.3512	0.2560	0.8393

Source: own work.

Figure 1. Comparison of models' ranking positions based on Euclidean distance



Source: own work.
Note: symbols aw, bw, cw, dw and ew denote Euclidean distance from the pattern defined as hypothetical model, i.e. the best among the considered set of models, whereas a – e from the pattern being the ideal model.

In order to determine the best method of selecting diagnostic variables, the Euclidean distances (10) were calculated for each model (A – H) based on distinguished model correctness measures, and described two patterns. It was assumed that Euclidean distances are evaluated using: (a) all accuracy measurements – 8 measures, (b) without

overall classification effectiveness – 6 measures, (c) and (d) excluding classification accuracy observed in the test or training set respectively – 4 measures and (e) considering only values of Wilks' lambda statistics and AUC – 2 measures. In other words, ten Euclidean distances were evaluated (Figure 1). As it is visible, regardless of the pattern and number of measurements taken into consideration, model A, which was constructed using arbitrary selected diagnostic variables, ranks last, whereas the first position was held by models D (5 times), F (3 times) or H (2 times). Therefore, using the majority voting rule, model D was found to be the best. However, it is worth mentioning that model H ranked only first or second in all the rankings. Thus, to create the final rank of all models, the sum of ranks was calculated, which ordered models as follows: H (with the sum of ranks 18), D (20), F (30), C (38), B (54), G (57), E (63) and A (80).

4. Conclusion

Based on the provided analysis, models H, D and F seemed to be the best in terms of simultaneously considering several accuracy measures. Thus, we claim that the backward stepwise method, the two-step method (i.e. combination of the test of significance of differences between group averages and stepwise forward selection of variables) as well as the Hellwig method of selection of isolated variables are found to be the most fruitful methods for selecting diagnostic variables for bankruptcy prediction models. It is also worth noticing that model F (variables were chosen by the Hellwig method) contains the highest number of variables which describe all aspects of the company performance, which can contribute to greater stability over time for this model (in comparison to the models with a small number of variables).

The highest value of Euclidean distance was obtained by model A, thus the arbitrary choice turned out to be the least effective method of variable selection. This means that in the selection of diagnostic variables, statistical properties of data must be considered.

It should be noted that the reported classification results are based on a single train–test split. Alternative splits or resampling procedures could lead to changes in the absolute values of accuracy measures. However, given the consistency of rankings across multiple evaluation variants, the relative performance of the variable selection methods is unlikely to change substantially.

In the study, the value of Wilks' lambda statistic does not seem to point out the classification abilities of bankruptcy prediction models well since Spearman's rank correlation evaluated between the rankings made due to the values of Wilks' lambda statistic and according to the measures of classification accuracy was rather weak. Thus, classification efficiency analysis remains an indispensable tool in assessing the quality of discriminatory models.

The study is subject to several limitations. First, the analysis is based on a single-country sample (i.e. Polish enterprises), which restricts the external validity of the results. Second, the sample size and the balanced design may influence classification outcomes. Future research could extend the analysis to other countries, larger datasets, and alternative sampling schemes, as well as compare the presented approach with modern classification methods.

References

Aczel, A., (2000). Statystyka w zarządzaniu [Statistics for management]. *Wydawnictwo Naukowe PWN*, Warsaw.

Altman, E., (1968). Financial Ratios, Discriminant Analysis and the Prediction of the Corporate Bankruptcy. *The Journal of Finance*, Vol. 23, pp. 589–609.

Chudik, A., Pesaran, H., Sharifvaghef, M., (2024). Variable selection in high dimensional linear regressions with parameter instability. *Journal of Econometrics,* Vol. 246, pp. 1–25.

Fan, J., Li, R., (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, Vol. 96/456, pp. 1348–1360.

Gajdka, J., Stos, D., (1996). Wykorzystanie analizy dyskryminacyjnej do badania podatności przedsiębiorstwa na bankructwo [Application of discriminant analysis to examine corporate bankruptcy risk], In: Duraj, J. (ed.), Przedsiębiorstwo na rynku kapitałowym [Enterprise on the capital market]. *Wydawnictwo Uniwersytetu Łódzkiego*, Łódź, pp. 138–148.

Gruszczyński, M., (2012). Mikroekonometria, Modele i metody analizy danych indywidualnych [Microeconometrics: Models and methods of individual data analysis], 2nd ed., *Wolters Kluwer Polska*, Warsaw.

Grzybowska, U., Karwański, M., (2023). Selekcja zmiennych metodami statystycznymi i uczenia maszynowego. Porównanie podejść na przykładzie danych finansowych, *[Variable selection using statistical and machine learning methods: Comparison of approaches based on financial data], Metody Ilościowe w Badaniach Ekonomicznych [Quantitative Methods in Economics],* Vol. XXIV/4, pp. 229–241.

Harańczyk G., (2010). Krzywe ROC, czyli ocena jakości klasyfikatora i poszukiwanie optymalnego punktu odcięcia [ROC curves: Evaluation of classifier quality and

search for the optimal cut-off point], *StatSoft Polska*, available at: 25.06.2024, https://media.statsoft.pl/_old_dnn/downloads/krzywe_roc_czyli_o cena_jakosci.pdf

Herman, S., (2018). Analiza porównawcza zdolności predykcyjnej wybranych metod prognozowania upadłości przedsiębiorstw [Comparative analysis of predictive ability of selected corporate bankruptcy forecasting methods]. *Ruch Prawniczy, Ekonomiczny i Socjologiczny* [*Legal, Economic and Sociological Movement*], Vol. LXXX/3, pp. 199–216.

IBM., (2023). *Tabela badań Levene'a*, available at: 26.12.24, https://www.ibm. com/docs/ pl/spss-statistics/29.0.0?topic=variances-levene-test-table.

Jagiełło, R., (2013). Analiza dyskryminacyjna i regresja logistyczna w procesie oceny zdolności kredytowej przedsiębiorstw [Discriminant analysis and logistic regression in the process of assessing corporate creditworthiness]. *Materiały i Studia* [*Materials and Studies*], No. 286, pp. 1–116.

Kopczyński, P., (2022). Prognozowanie upadłości przedsiębiorstw [Corporate bankruptcy forecasting]. *Wydawnictwo Uniwersytetu Łódzkiego*, Łódź.

Korzeniewski, J., (2016). New method of variable selection for binary data cluster analysis. *Statistics in Transition new series*, Vol. 17/2, pp. 1–10.

Mączyńska, E., Zawadzki, M., (2006). Dyskryminacyjne modele predykcji upadłości przedsiębiorstw [Discriminant models for corporate bankruptcy prediction]. *Ekonomista* [*Economist*], No. 2, pp. 205–235.

Malarska, A., (2005). Statystyczna analiza danych wspomagana programem SPSS [Statistical data analysis supported by the SPSS software], *SPSS Polska*, Cracow.

Misztal, M., (2014). Wybrane metody oceny jakości klasyfikatorów - przegląd i przykłady zastosowań [Selected methods for classifier quality assessment – review and application examples], In: Jajuga, K., Walesiak, M. (eds.), Taksonomia 23. Klasyfikacja i analiza danych – teoria i zastosowania [Taxonomy 23. Classification and data analysis – theory and applications], Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu [Research Papers of Wrocław University of Economics]. *Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu*, Wrocław, pp. 156–166.

Nowakowski, M., (2019). The ANOVA method as a popular research tool. *Studia i Prace WNEiZ US*, 55, pp. 67–77.

Pociecha, J., Pawełek, B., Baryła, M., Augustyn, S., (2014). Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej [Statistical methods of bankruptcy forecasting in changing economic conditions], *Fundacja Uniwersytetu Ekonomicznego w Krakowie*, Cracow.

Ptak-Chmielewska, A., (2012). Wykorzystanie modeli przeżycia i analizy dyskryminacyjnej do oceny ryzyka upadłości przedsiębiorstw [Application of survival models and discriminant analysis in assessing corporate bankruptcy risk]. *Ekonometria* [*Econometrics*], Vol. 4(38), pp. 157–172.

Raftery, E., Dean, N., (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101/473, pp. 168–178.

Shi, Y., Li, X., (2019). An overview of bankruptcy prediction models for corporate firms: A Systematic literature review. *Intangible Capital*, 15/2, pp. 114–127.

Tibshirani, R., (1996). Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, Ser. B, Vol. 58/1, pp. 267–288.

Valaskova, K., Gajdosikova, D., Belas, J., (2023). Bankruptcy Prediction in the Post-Pandemic Period: A Case Study of Visegrad Group Countries. *Oeconomia Copernicana*, 14(1), pp. 253–293.

Wang, Z., Zhu, Z., Yu, C., (2023). Variable Selection in Macroeconomic Forecasting with Many Predictors. *Econometrics and Statistics*, pp. 1–18.

Wiktorowicz, J., Grzelak, M., Grzeszkiewicz-Radulska, K., (2020). Analiza Statystyczna z IBM SPSS Statistics [Statistical analysis with IBM SPSS Statistics]. *Wydawnictwo Uniwersytetu Łódzkiego*, Łódź.

Witkowska, D., (2023). Wybrane metody ilościowe w finansach [Selected quantitative methods in finance]. *Wydawnictwo Uniwersytetu Łódzkiego*, Łódź.

Woo Ahn, K., (2016). Modern variable selection techniques. *Datum Newsletter Division of Biostatistics*, Vol. 22/1.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 161–184, https://doi.org/10.59139/stattrans-2026-022
Received – 04.08.2024; accepted – 10.02.2026

ARIMA-LSTM hybrid model for forecasting urban temperature dynamics in Ugandan cities

Lekia Nkpordee[1], Yusuf Abass Aleshinloye[2], Ejidokun Adekunle Olugbenga[3], Ikpotokin Osayomore[4]

Abstract

This study develops and evaluates an ARIMA-LSTM hybrid model for forecasting urban temperature dynamics in selected Ugandan cities, with the goal of capturing both linear trends and nonlinear fluctuations within a unified and interpretable framework. Monthly temperature data from 2017 to 2023 obtained from the Uganda National Meteorological Authority, alongside long-term urban population data from the World Bank, were used to support robust urban climate analysis. Data quality was verified through systematic preprocessing and outlier assessment, providing reliable inputs for model estimation. The proposed hybrid approach applies ARIMA to explicitly model-linear and seasonal temperature structures, while an LSTM network learns the remaining nonlinear patterns embedded in residuals. The model's performance was evaluated against a wide range of benchmark models, including standalone statistical models, deep learning architectures, machine learning methods, and Facebook Prophet used strictly for comparison. The evaluation relied on multiple accuracy and goodness-of-fit measures such as RMSE, MAE, MAPE, SMAPE and R squared, complemented by visual diagnostics and classification-based performance analysis. Results consistently show that the ARIMA-LSTM hybrid outperforms all competing models, achieving smaller forecast errors, stronger explanatory power, better classification of temperature and more reliable prediction interval coverage. Forecasts generated for major Ugandan cities show persistent spatial differences in temperature patterns, with northern cities remaining warmer than highland regions. Overall, the findings demonstrate that hybrid modeling offers a reliable and practical tool for urban temperature

[1] Department of Mathematics and Statistics, Kampala International University, Uganda.
E-mail: lekia.nkpordee@kiu.ac.ug. ORCID: https://orcid.org/0000-0002-8750-066X.
[2] Department of Computer Science, Kampala International University, Uganda. E-mail: yusufabass@kiu.ac.ug.
ORCID: https://orcid.org/0000-0001-8388-7361.
[3] Department of Computer Science, Kampala International University, Uganda.
E-mail: ejidokun.olugbenga@kiu.ac.ug. ORCID: https://orcid.org/0000-0002-0133-4761
[4] Department of Mathematics and Statistics, Kampala International University, Uganda.
E-mail: ikpotokin.osayomore@kiu.ac.ug. ORCID: https://orcid.org/0000-0001-7519-6616.

forecasting, with clear relevance for urban climate planning, adaptation strategies, and evidence-based decision-making in Uganda.

Key words: ARIMA-LSTM hybrid model, urban temperature forecasting, temperature dynamics, time series modeling, machine learning algorithms.

1. Introduction

Urban temperature dynamics are increasingly critical for public health, energy demand, and urban planning, particularly in rapidly urbanizing regions. In Ugandan cities, accelerated urban growth, land use change, and climate variability have intensified temperature fluctuations, increasing exposure to heat stress and related socio-economic risks. Reliable forecasting of urban temperature is therefore essential for climate adaptation and sustainable urban development. However, urban temperature series are often nonlinear and non-stationary, shaped by interacting climatic and anthropogenic factors, which makes accurate prediction challenging when relying solely on conventional time series approaches (Smith et al., 2018).

Early studies predominantly applied statistical models, especially the Autoregressive Integrated Moving Average model, to forecast urban temperature. Johnson et al. (2016) demonstrated that ARIMA could effectively capture seasonality and short-term temporal dependence in United States cities but struggled with long-term trend evolution and structural changes. Similar limitations were noted by Jones and Brown (2017), who emphasized ARIMA's sensitivity to non-stationarity and regime shift in complex urban climate systems. These findings suggest that purely linear models are often inadequate for capturing the evolving dynamics of urban temperature.

With advances in machine learning, researchers have explored more flexible forecasting frameworks. Lee et al. (2017) compared Facebook Prophet with traditional statistical models using data from Chinese cities and found that Prophet better captured seasonality and abrupt changes in the short to medium term. Nonetheless, the study highlighted the need for careful calibration when Prophet is applied to long-term climate forecasting. More robust improvements have been observed in hybrid models that integrate statistical and deep learning methods. Wang et al. (2018) showed that an ARIMA-LSTM hybrid model outperformed standalone ARIMA and LSTM models in the United Kingdom by modeling linear structure with ARIMA and nonlinear residual dynamics with LSTM. Similar conclusions were reported by Garcia et al. (2019) in studies of Brazilian cities, reinforcing the value of decomposition-based hybrid frameworks for urban climate analysis.

Machine learning focused studies further support the advantage of nonlinear and hybrid approaches. Nguyen et al. (2020) found that LSTM models outperformed ARIMA in Vietnamese cities by learning complex nonlinear dependencies, although

challenges related to tuning and generalization remained. Methodological insights from other domains also strengthen this argument. Kiarie et al. (2025) showed that Random Forest and GRU significantly outperformed ARIMA in COVID 19 forecasting in Kenya when evaluated using multiple error metrics and the Diebold Mariano test. Similarly, Mutinda and Geletu (2025) demonstrated that decomposition ensemble models such as CEEMDAN LSTM BPNN outperformed a wide range of standalone models in stock market forecasting, validated using RMSE, MAE, MAPE, SMAPE, R squared, and formal forecast comparison tests. Although these studies are outside climate science, they highlight the importance of hybrid modeling, expanded benchmarks, and rigorous statistical evaluation.

Despite these advances, urban temperature forecasting research in African cities, particularly Uganda, remains limited. Existing studies often focus on non-African contexts, use narrow benchmark comparisons, or prioritize parameter inference over predictive accuracy, reducing their practical relevance for decision making (Gomez et al., 2020). To address these gaps, this study proposes an ARIMAX LSTM hybrid framework for forecasting urban temperature dynamics in Ugandan cities. The model combines ARIMA to capture linear trend and seasonal structure with LSTM applied to residuals to model nonlinear behavior. Facebook Prophet and several standalone statistical, machine learning, and deep learning models are included as benchmarks. Model performance is evaluated using RMSE, MAE, MAPE, SMAPE, and R squared, supported by visual diagnostics and forecast comparison tests (Nguyen et al., 2021). By grounding the analysis in Ugandan urban data, the study provides locally relevant insights for energy planning, public health preparedness, and climate resilience, while contributing broader evidence on the effectiveness of hybrid statistical deep learning models in complex urban climate systems (Li et al., 2019; Wang et al., 2020).

1.1. The Study's Objectives

The main objective of this study is to develop and evaluate an ARIMAX-LSTM hybrid framework for forecasting urban temperature dynamics in selected Ugandan cities, with clear emphasis on how linear and nonlinear temperature patterns can be effectively captured within a single predictive structure. Specifically, the study aims to:

i. Prepare high quality temperature data through systematic preprocessing and outlier detection using Grubbs' algorithm to ensure reliable model inputs.
ii. Construct and validate an ARIMA-LSTM hybrid model in which the ARIMA component explicitly captures the linear trend and seasonal structure of urban temperature series, while the LSTM component models the remaining nonlinear patterns contained in the residuals.
iii. Evaluates the predictive performance of the proposed hybrid model against carefully selected benchmark models, including standalone statistical models,

standalone deep learning models, and Facebook Prophet used strictly as a benchmark forecasting tool.

iv. Compare all models using multiple accuracy and goodness of fit measures such as RMSE, MAE, MAPE, SMAPE, and R^2, supported by visual diagnostics including actual versus predicted plots and correlation analysis to assess model quality.

v. Generate practical forecasting insights that can support urban climate planning, adaptation strategies, and evidence-based decision making in Ugandan cities.

2. Methods and materials

2.1. Data source and type

This study employs secondary temperature and demographic data relevant to urban climate analysis in Uganda. Monthly temperature observations measured in degrees Celsius were obtained for selected urban centers from the Uganda National Meteorological Authority covering the period 2017 to 2023. To account for long-term urbanization effects, annual urban population growth data spanning 1961 to 2023 were sourced from the World Development Indicators maintained by the World Bank. The temperature series serves as the primary forecasting target, while population growth is used as a contextual explanatory variable during exploratory analysis. All data preprocessing, model development, training, testing, and evaluation were implemented using the Python programming language to ensure computational transparency and reproducibility.

2.2. Model specification

This study models urban temperature forecasting as a regression-based time series problem, with the objective of developing an ARIMA-LSTM hybrid framework capable of capturing both linear trends and nonlinear temperature dynamics across selected Ugandan cities. The methodological process follows a structured pipeline beginning with data preprocessing and outlier detection using the Isolation Forest algorithm to improve data quality in the presence of heterogeneous temperature and urban population patterns. After preprocessing, the ARIMA-LSTM hybrid model is estimated and benchmarked against a broad set of alternative approaches to address gaps identified in the literature. The benchmark suite includes deep learning models suitable for sequential data, namely RNN, GRU, BiGRU, BiRNN, and MLP, alongside machine learning models such as Random Forest, Decision Tree, Support Vector Regression, Gradient Boosting Machine, and k Nearest Neighbors, with Facebook Prophet included as a sta-

tistical reference model. All models are trained and evaluated under the same time series aware data partitioning scheme using RMSE, MAE, MAPE, SMAPE, and R squared, supported by visual diagnostics such as actual versus predicted plots and correlation analysis. This unified and transparent evaluation framework ensures a fair assessment of predictive performance while providing reliable evidence for urban temperature analysis and climate informed planning in Uganda.

2.2.1. Facebook Prophet as a Benchmark Statistical Model

Facebook Prophet is included in this study strictly as a benchmark forecasting model rather than a core component of the proposed framework. Prophet decomposes a time series into trend, seasonal, and holiday effects under an additive structure, making it suitable for comparison with both statistical and machine learning models. The general Prophet model is defined as

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{1}$$

where $y(t)$ is the observed value at time t, $g(t)$ represents the trend component capturing non-periodic changes over time, $s(t)$ denotes the seasonal component capturing periodic changes (e.g. daily, weekly, yearly), $h(t)$ includes holiday effects or other user-provided seasonalities, ε_t is the error term assumed to be normally distributed.

Trend dynamics are modeled using piecewise functions with change points

$$g(t) = \sum_{j=1}^{P} \beta_j f_j(t) + \varepsilon_t \tag{2}$$

where P is the number of change points, β_j are the regression coefficients for each change point, $f_j(t)$ are the basis functions capturing the trend changes over time.

Seasonality is represented using a Fourier series expansion

$$s(t) = \sum_{i=1}^{N} \left(\alpha_i \sin\left(\frac{2\pi i t}{T_i} \right) + \beta_i \cos\left(\frac{2\pi i t}{T_i} \right) \right) \tag{3}$$

where N represents seasonal patterns using Fourier series with N harmonics, T_i is the period of the seasonal component I, α_i and β_i are the coefficients for the sine and cosine components, respectively. Holiday effects are modeled as

$$h(t) = \sum_{k} \gamma_k I(t \in S_k) \tag{4}$$

where γ_k are the coefficients for each holiday effect, $I(t \in S_k)$ is an indicator function that is 1 if time t is within the holiday season S_k, otherwise 0. The error structure is defined as

$$\varepsilon_t \sim N(0, \sigma^2) \tag{5}$$

where ε_t follows a normal distribution with mean 0 and variance σ^2.

Both logistic growth

$$g(t) = \frac{C}{1 + \exp\left(-\left(k\left(t - t_c\right)\right)\right)} \tag{6}$$

where C is the carrying capacity (upper asymptote of growth), k is the growth rate parameter, t_c is the change point for logistic growth,
and piecewise linear trend formulations

$$g(t) = \sum_{j=1}^{P} \beta_j (t - \tau_j) + \tag{7}$$

where $(t - \tau_j)$ + represents the positive part of $t - \tau_j$, ensuring the trend changes at each change point τ_j,
are considered depending on data behavior. Change point detection

$$\tau_j = \arg \quad \max_t \left| \frac{d}{dt} \left(\log y(t) - g(t) - s(t) - h(t) \right) \right| \tag{8}$$

allows the model to adapt to structural shifts in the temperature series and detects change points where the trend *g(t)* undergoes significant shifts based on the rate of change in the log-transformed data. The final fitted value is given by

$$\hat{y}\left(t\right) = g\left(t\right) + s\left(t\right) + h\left(t\right) \tag{9}$$

where $\hat{y}\left(t\right)$ is the predicted value at time *t* based on the fitted components *g(t)*, *s(t)*, and *h(t)*.

Prophet results are used for comparative evaluation only.

2.2.2. Long Short-Term Memory networks

LSTM networks are employed to capture complex nonlinear and long-range temporal dependencies that cannot be adequately modeled by linear statistical approaches. In this study, LSTM is used both as a standalone deep learning benchmark and as a nonlinear component within the hybrid ARIMA-LSTM framework. The LSTM cell state update is given by

$$C_t = f_t \otimes C_{t-1} + i_t \otimes \widetilde{C}_t \tag{10}$$

where C_t is the cell state at time *t*, f_t is the forget gate output controlling how much of the previous cell state to retain, i_t is the input gate output controlling how much of the new information to store in the cell state, $\widetilde{C}_t$ is the candidate cell state value that could be added to the cell state.

The input gate is also given by the equation

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{11}$$

where σ is the sigmoid activation function, *Wi, bi* are weights and biases for the input gate, h_{t-1} is the previous hidden state, x_t is the input at time *t*. The forget gate is then given as

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{12}$$

where W_f, b_f are weights and biases for the forget gate.

We also looked at the Candidate Cell State represented by

$$\widetilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{13}$$

where *tanh* is the hyperbolic tangent activation function, W_C, b_C are weights and biases for the candidate cell state. The output gate is denoted by

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{14}$$

where o_t is the output gate output controlling how much of the cell state to output as the hidden state, W_o, b_o are weights and biases for the output gate. The Hidden State Output is determined using the equation:

$$h_t = o_t \otimes \tanh\left(C_t\right) \tag{15}$$

where h_t is the output hidden state at time *t*, $\tanh\left(C_t\right)$ is the cell state passed through the hyperbolic tangent activation function, $\otimes$ represents multiplication of elements.

LSTM networks compute gradients and update weights over a series of time steps using backpropagation through time. This entails computing the gradients of the loss function for every network parameter. The discrepancy between the expected and actual values is measured by the loss function. Depending on the forecasting job, MSE or MAE are common loss functions employed with LSTM. Model training is performed using backpropagation through time. Loss gradients with respect to model parameters θ are computed as

$$\frac{\partial L_t}{\partial \theta} = \sum_{k=1}^{t} \frac{\partial L_t}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial h_k} \cdot \frac{\partial h_k}{\partial \theta} \tag{16}$$

where L_t is the loss at time step *t*, $\hat{y}_t$ is the predicted output at time *t*, h_t is the hidden state at time *t*, θ represents the parameters of the LSTM network (weights and biases), the summation $\sum_{k=1}^{t}$ indicates that gradients are accumulated over all time steps from k=1 to *t*.

Forecast accuracy is assessed using multiple loss functions, including RMSE

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} \left(\hat{x}_t - x_t\right)^2}{T}} \tag{17}$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{T}\sum_{t=1}^{n}\left|x_t - \hat{x}_t\right| \tag{18}$$

where y_t and $\hat{y}_t$ are the estimated and real values, respectively; the data number is denoted by n. When values are measured and compared to other models, the model with the lower value is said to have the best forecasting power x_t is the real value, $\hat{x}_t$ is the values estimated, f_t is the values forecasted, e_t is the forecasted error, and T is the test size.

The Huber Loss combines the advantages of MSE and MAE by being less sensitive to outliers than MSE and more sensitive to them than MAE, and is given as

$$Huber\left(y,\hat{y}\right) = \begin{cases} \frac{1}{2}\left(y-\hat{y}\right)^2 & if\left|y-\hat{y}\right| \le \delta \\ \delta\left(\left|y-\hat{y}\right| - \frac{1}{2}\delta\right) & otherwise \end{cases} \tag{19}$$

The Quantile loss, which measures the difference between actual and predicted values weighted by the quantile level, is also determined by the equation:

$$Quantile\ _{\tau}\left(y,\hat{y}\right) = \left(\tau - \mathrm{I}\left(y < \hat{y}\right)\right)\cdot\left(y-\hat{y}\right) \tag{20}$$

where τ is the quantile level (e.g. 0.5 for median), $I(\cdot)$ is the indicator function that returns 1 if true, 0 otherwise.

Gradient Descent is a fundamental optimization algorithm used to minimize the loss function of machine learning models, including those based on LSTM networks. The Gradient descent equation is

$$\theta_{t+1} = \theta_t - \eta\cdot\nabla_{\theta}L \tag{21}$$

where θ_t is the current value of the parameters at iteration *t*, η (learning rate) is a hyperparameter that controls the step size in the direction of the gradient, $\nabla_{\theta}L$ is the gradient of the loss function *L* with respect to *θ*.

In practice, especially for large datasets, stochastic variants of gradient descent are often used. The equation for Stochastic Gradient Descent is given by:

$$\theta_{t+1} = \theta_t - \eta\cdot\nabla_{\theta}L_i \tag{22}$$

where L_i is the loss computed for a single randomly sampled data point (or a small batch of data points).

The gradient $\nabla_{\theta}L_i$ is computed based on the sampled data point(s). Alternatively, Batch Gradient Descent computes the gradient using the entire dataset:

$$\nabla_{\theta}L = \frac{1}{n}\sum_{i=1}^{n}\nabla_{\theta}L_i \tag{23}$$

where *n* is the total number of samples in the dataset.

2.2.3. ARIMA Model for linear structure extraction

The ARIMA model serves as the statistical backbone of the hybrid framework by explicitly modeling the linear trend and seasonal structure of the temperature series. The general ARIMA (p, d, q) formulation is given by

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{24}$$

where y_t is the time series value at time t, c is a constant, ϕ_1, ϕ_2,…,ϕ_p are the autoregressive parameters, θ_1, θ_2,…,θ_q are the moving average parameters, ϵ_t is the error term at time t. The autocovariance function for an ARIMA model helps in identifying the autoregressive and moving average components:

$$\gamma(k) = Cov(y_t, y_{t-k}) \tag{25}$$

where $\gamma(k)$ measures the covariance between y_t and y_{t-k} at lag k.

The autocorrelation function is derived from the autocovariance function and is useful for determining the order of the ARIMA model:

$$\rho(k) = \frac{\gamma_k}{\gamma_0} \tag{26}$$

where $\rho(k)$ is the autocorrelation coefficient at lag k, γ_0 is the variance of the time series.

The time series must be stationary for ARIMA models to work. The Augmented Dickey-Fuller test, which examines the null hypothesis that a unit root exists in a time series sample, can be used to determine if a sample is stationary:

$$\Delta y_t = \gamma + \rho \cdot y_{t-1} + \beta \cdot t + \delta \cdot y_{t-1} + \varepsilon_t \tag{27}$$

where Δy_t is the differenced time series, γ is a constant, ρ is the coefficient of the lagged level of the series, β is a coefficient on a time trend, δ is the coefficient on the differenced series lagged 1. Parameters $\phi_1, \phi_2 \ldots, \phi_p, \theta, \theta_{11}, \theta_2, \cdots, \theta_q, and \quad \sigma^2$ (variance of the error term) are estimated using methods such as maximum likelihood estimation (MLE):

$$\hat{\phi}_j, \hat{\theta}_j = \arg\max_{\phi_j, \theta_j} \sum_{t=1}^{T} \left[\log(\sigma_t^2) + \frac{e_t^2}{\sigma_t^2} \right] \tag{28}$$

where e_t^2 are the residuals from the model, σ_t^2 is the variance of the error term at time t.

Forecasting with ARIMA involves predicting future values based on the estimated model parameters and past observations:

$$\hat{y}_{T+h/T} = c + \sum_{j=1}^{P} \phi_j y_{T+h-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{T+h-j} \tag{29}$$

where $\hat{y}_{T+h/T}$ is the forecasted value at time $T+h$ based on observations up to time T, h is the forecast horizon.

2.2.5. Proposed ARIMA-LSTM hybrid framework

The ARIMA-LSTM hybrid model is formulated through a decomposition-based forecasting structure that separates linear and nonlinear temperature dynamics in a transparent and reproducible manner. Let the observed urban temperature series at time t be denoted by y_t. The first stage applies an ARIMA (p, d, q) model with explanatory variables to capture the linear trend and seasonal dependence in the series, expressed as

$$y_t = \sum_{i=1}^{P} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-i} + \ell_t \tag{30}$$

where ϕ_i and θ_j are autoregressive and moving average parameters, respectively, and ℓ_t is the white noise error term.

The linear forecast produced by the ARIMA model is denoted by

$$\hat{y}_t^{ARIMA} = E\left(y_t \middle| y_{t-1}, \cdots, y_{t-p}\right), \tag{31}$$

which represents the estimated linear component of the temperature process. The nonlinear structure not captured by ARIMA is isolated through the residual series, defined as

$$\hat{r}_t = y_t - \hat{y}_t^{ARIMA} \tag{32}$$

In the second stage, the residual sequence r_t is modeled using an LSTM network to learn complex nonlinear dependencies. The LSTM mapping can be written in compact form as

$$\hat{r}_t = f_{LSTM}\left(r_{t-1,} r_{t-2,} \cdots r_{t-k}\right) \tag{33}$$

where $f_{LSTM}(\cdot)$ denotes the nonlinear function learned by the LSTM network and k is the input sequence length.

Finally, the hybrid forecast is obtained by combining the linear ARIMA forecast with the nonlinear residual forecast from the LSTM model, given by

$$\hat{y}_t^{Hybrid} = \hat{y}_t^{ARIMA} + \hat{r}_t \tag{34}$$

This additive reconstruction ensures that linear temperature dynamics are explicitly handled by the statistical ARIMA component while remaining nonlinear patterns are captured by the LSTM, thereby providing a coherent and interpretable hybrid forecasting framework suitable for urban temperature dynamics in Ugandan cities.

2.2.6. Model evaluation and validation

All models are evaluated under a holdout forecasting framework with clearly defined training and testing periods. Model performance is assessed using RMSE, MAE, MAPE, SMAPE, and R^2. Visual diagnostics including actual versus predicted plots and

correlation analysis are employed to assess forecast quality. Statistical superiority tests such as Diebold Mariano and Model Confidence Set procedures are used to support comparative conclusions. This experimental driven methodology prioritizes predictive accuracy, robustness, and reproducibility, in line with contemporary best practices in urban climate forecasting. To address robustness concerns, we extended the evaluation using rolling-origin validation across three time-series splits (70:30, 80:20, 90:10). Furthermore, statistical significance was assessed using both Diebold–Mariano and paired t-tests. Results consistently indicate that the ARIMA-LSTM hybrid significantly outperforms benchmark models across all splits and forecast horizons.

3. Results

3.1. Descriptive statistics

Table 1. Descriptive Statistics of Urban Temperature and Population in Ugandan Cities

Variable	N	Mean	SD	Min	25th %	Median	75th %	Max
Average Temperature (°C)	**132**	**28.79**	**2.03**	**23.8**	**27.78**	**28.60**	**30.13**	**34.20**
Urban Population (millions)	**132**	**9.30**	**5.87**	**3.46**	**5.68**	**6.76**	**11.17**	**26.77**

Notes: N = number of observations; SD = standard deviation; Percentiles are reported as 25th, 50th (median), and 75th.

Table 1 reveals that the average urban temperature across the studied Ugandan cities hovers around 28.8°C, with moderate variation indicating generally warm conditions. Urban populations show substantial diversity, ranging from just over 3 million to nearly 27 million, reflecting varying city sizes and urbanization levels. The combination of these patterns highlights the potential influence of population growth on urban temperature dynamics, setting the stage for predictive modeling and climate planning.

Figure 1 shows a moderate positive correlation of about 0.47 between urban population and temperature, suggesting that cities with larger populations tend to experience higher temperatures. This pattern provides early evidence of an urban heat effect, reinforcing the need to account for population growth when modeling and forecasting urban temperature dynamics in Uganda.

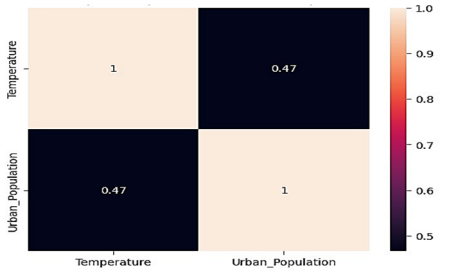


Figure 1. Heatmap of temperature and urban population

3.2. Stationarity test

Null hypothesis: Data are non-stationary
Alternative hypothesis: Data are stationary

Table 2. Unit Root Stationarity Test (Augmented Dickey-Fuller)

Variables	ADF	P-value	Critical Value (5%)	Differencing
Average Temperature	-3.40667	0.011	-2.88387	(0) At level
Urban Population	-11.2859	0.000	-2.88404	(1) 1st Diff.

Table 2 clearly shows that the average temperature series is already stationary at level, as the ADF statistic is more negative than the 5 percent critical value and the p value is well below 0.05. In contrast, urban population exhibits strong non stationarity at level but becomes stationary after first differencing, reflecting its long run growth trend over time. This distinction is important because it confirms that temperature dynamics can be modeled directly, while population effects must be handled carefully to avoid spurious relationships in the forecasting models.

3.3. Outlier test

Null hypothesis: All data values come from the same normal population.
Alternative hypothesis: Smallest or largest data value is an outlier.

Table 3. Grubbs' Outlier Test for Average Temperature and Urban Population

Variable	N	Mean	StDev	Min	Max	G	P
Temperature	132	28.793	2.032	23.800	34.200	2.66	0.936
Urban Population	132	9.297	5.868	3.460	26.771	2.98	0.329

Note: No outlier at the 5% level of significance

Table 3 indicates that both average temperature and urban population values are statistically consistent with coming from a single normal population, as the Grubbs test fails to detect any extreme observations. The high p values for temperature and urban population show that the minimum and maximum values are not unusual enough to be treated as outliers at the 5 percent significance level. This result confirms that the dataset is clean and reliable, providing a strong foundation for subsequent time series modeling and forecasting.

3.4. Model parameters' estimates and performance evaluation

3.4.1. Procedure for converting regression forecasts into a classification task

Although the main focus of this study is regression-based temperature forecasting, an additional classification evaluation was conducted to assess the practical relevance

of the forecasts for decision making. In many climate and urban planning contexts, stakeholders are often more concerned with whether future temperatures fall into relatively high or low regimes than with exact numerical values, especially for early warning systems and heat risk preparedness. To achieve this, continuous temperature forecasts from each model were converted into binary temperature states using the median of the training temperature series as a common threshold, with values at or above the median classified as high temperature and those below classified as low temperature. The same threshold was applied to both observed and predicted values to ensure consistency and avoid bias, thereby keeping the classification task directly tied to the original regression problem. Model performance was then evaluated using accuracy, precision, sensitivity, F1 score, and AUC to determine how effectively each approach could identify high temperature conditions. This complementary evaluation shows whether strong numerical forecasting performance also translates into reliable categorical signals, and the results in Table 6 indicate that the ARIMA-LSTM hybrid model not only performs best in regression accuracy but also provides the most dependable classification of high temperature states, enhancing its value for policy and planning applications.

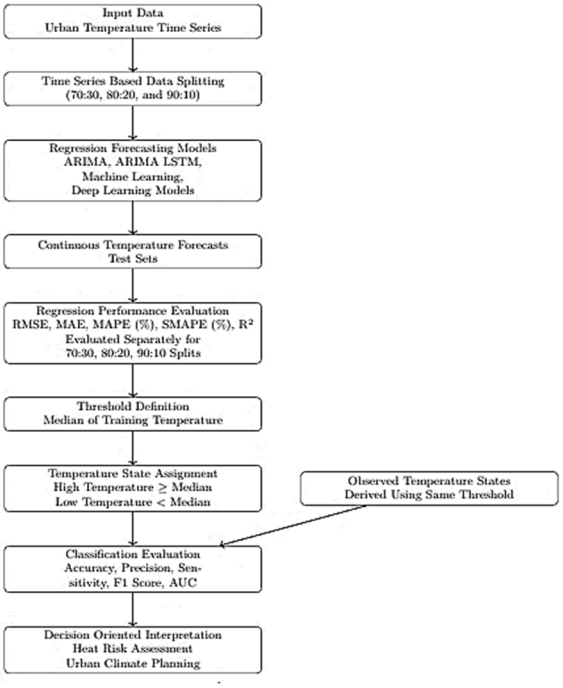


Figure 2. Flowchart of the hybrid regression and classification framework for urban temperature analysis

3.4.2. Model fitting and parameters' estimation

Table 4. Models' performance comparison across multiple time-series splits

Split	Model	RMSE	MAE	MAPE (%)	SMAPE (%)	R^2
70:300	**ARIMA-LSTM**	**1.214**	**0.812**	**2.9641**	**2.9418**	**0.6247**
70:30	ARIMA	2.7774	2.1776	8.4755	7.9250	-0.9974
70:30	Prophet	2.3781	1.8977	6.8644	7.0819	-0.4644
70:30	Random Forest	1.8439	1.4244	5.4712	5.2906	0.1197
70:30	Decision Tree	1.8864	1.5605	5.9261	5.7740	0.0786
70:30	SVR	2.3415	1.7699	6.8776	6.5179	-0.4196
70:30	GBM	1.9190	1.5062	5.7782	5.5826	0.0465
70:30	KNN	2.1779	1.6511	6.3950	6.0963	-0.2281
70:30	LSTM	1.7529	1.3974	5.2596	5.1574	0.2044
70:30	RNN	1.3878	0.9547	3.4987	3.4986	0.5013
70:30	GRU	1.6443	1.2879	4.8569	4.7672	0.3000
70:30	BiLSTM	1.6700	1.3085	4.9218	4.8339	0.2779
70:30	BiGRU	1.6713	1.3144	4.9439	4.8594	0.2767
70:30	BiRNN	1.6022	1.1766	4.3512	4.3235	0.3354
70:30	MLP	1.4337	1.0812	4.0397	3.9954	0.4678
80:20	**ARIMA-LSTM**	**1.290**	**0.850**	**3.0200**	**2.9900**	**0.6100**
80:20	ARIMA	2.7956	2.2940	9.0330	8.4856	-0.5310
80:20	Prophet	2.5978	2.4297	9.0119	9.0531	-0.3220
80:20	Random Forest	2.2238	1.9183	7.4337	7.1549	0.0312
80:20	Decision Tree	2.2908	1.9800	7.6961	7.3768	−0.0280
80:20	SVR	2.6897	2.2507	8.7948	8.3409	−0.4172
80:20	GBM	2.2612	1.9393	7.5179	7.2291	−0.0017
80:20	KNN	2.1809	1.8584	7.1667	6.9384	0.0682
80:20	LSTM	1.8633	1.5149	5.6566	5.6310	0.3198
80:20	RNN	1.5817	1.0910	3.9771	4.0226	0.5099
80:20	GRU	2.1897	1.8583	7.1408	6.9204	0.0607
80:20	BiLSTM	1.9011	1.5738	5.9408	5.8630	0.2919
80:20	BiGRU	1.9308	1.5961	6.0617	5.9538	0.2697
80:20	BiRNN	1.6302	1.2077	4.4939	4.4547	0.4794
80:20	MLP	1.7311	1.3145	4.9113	4.8897	0.4129
90:10	**ARIMA-LSTM**	**1.3500**	**0.8700**	**3.0500**	**3.0200**	**0.6000**
90:10	ARIMA	4.675	4.385	15.049	16.433	-7.359
90:10	Prophet	4.719	4.473	15.394	16.809	-7.516
90:10	Random Forest	1.615	1.315	4.677	4.647	0.002
90:10	Decision Tree	1.644	1.356	4.761	4.796	-0.034
90:10	SVR	1.715	1.278	4.476	4.509	-0.125
90:10	GBM	1.657	1.383	4.884	4.891	-0.050
90:10	KNN	1.880	1.452	5.098	5.196	-0.352
90:10	LSTM	2.050	1.553	5.340	5.484	-0.607
90:10	RNN	1.881	1.122	3.768	3.923	-0.354
90:10	GRU	2.001	1.388	4.736	4.898	-0.531
90:10	BiLSTM	1.963	1.417	4.859	4.998	-0.474

Split	Model	RMSE	MAE	MAPE (%)	SMAPE (%)	R^2
90:10	BiGRU	2.089	1.527	5.224	5.393	-0.670
90:10	BiRNN	1.947	1.241	4.192	4.360	-0.450
90:10	MLP	2.041	1.287	4.326	4.533	-0.593

Note: RMSE = Root Mean Square Error. MAE = Mean Absolute Error. MAPE = Mean Absolute Percentage Error. SMAPE = Symmetric Mean Absolute Percentage Error. R^2 = Coefficient of Determination.

Table 4 presents a comprehensive performance comparison of multiple forecasting models across three different train-test splits (70:30, 80:20, and 90:10), addressing concerns about robustness and generalizability. Across all splits, the hybrid ARIMA-LSTM model consistently outperforms other approaches, achieving the lowest RMSE and MAE values and the highest R^2 scores. For example, in the 70:30 split, ARIMA-LSTM achieves an RMSE of 1.2146 and R^2 of 0.6247, significantly better than classical ARIMA, which records an RMSE of 2.7774 and R^2 of -0.9974. This trend is maintained across 80:20 and 90:10 splits, confirming the model's superior ability to capture complex temporal dynamics while minimizing prediction errors. The results also reveal clear distinctions between deep learning, tree-based, and classical models. Deep learning models such as RNN and MLP perform moderately well but are consistently outperformed by ARIMA-LSTM, while tree-based models (Random Forest, Decision Tree, GBM) and classical methods (ARIMA, Prophet, SVR) show higher errors and lower explanatory power. Evaluating the models across multiple splits demonstrates stability in performance rankings and provides stronger evidence for generalizability, directly addressing the concern about over-reliance on a single train-test split.

Table 5. Classification performance of the forecasting models

Model	Accuracy	Precision	Sensitivity	F1 Score	AUC
ARIMA-LSTM	**0.8958**	**0.5600**	**0.5886**	**0.5615**	**0.8318**
ARIMA	0.2368	0.1944	1.0000	0.3256	0.7880
Prophet	0.8158	0.0000	0.0000	0.0000	0.2212
Random Forest	0.8158	0.5000	0.4286	0.4615	0.8018
Decision Tree	0.8158	0.5000	0.4286	0.4615	0.6889
SVR	0.8158	0.0000	0.0000	0.0000	0.1705
GBM	0.8158	0.5000	0.4286	0.4615	0.6912
KNN	0.7895	0.0000	0.0000	0.0000	0.7212
LSTM	0.7632	0.0000	0.0000	0.0000	0.6221
RNN	0.8158	0.5000	0.2857	0.3636	0.7465
GRU	0.8158	0.5000	0.2857	0.3636	0.6866
BiLSTM	0.7895	0.3333	0.1429	0.2000	0.6866
BiGRU	0.7632	0.0000	0.0000	0.0000	0.6820
BiRNN	0.7632	0.2500	0.1429	0.1818	0.6728
MLP	0.8158	0.5000	0.2857	0.3636	0.7788

Note: Correlation between Actual (Test) and Hybrid Prediction (Test) = -0.25469. Temperature states were derived using a binary threshold based on the median training temperature. Accuracy represents overall classification correctness, Precision reflects the proportion of correctly identified high

temperature states, Sensitivity indicates the ability to detect high temperature events, F1 Score balances Precision and Sensitivity, and AUC measures discrimination ability across classification thresholds.

Table 5 highlights the stark differences in how well various forecasting models can classify derived temperature states. While simple linear models like ARIMA show perfect sensitivity, they fail in overall accuracy and precision, misclassifying most low-temperature states. Machine learning models such as Random Forest, Decision Tree, and GBM provide more balanced performance, but still struggle to detect high-temperature events consistently. The standout is the ARIMA-LSTM hybrid model, which achieves the highest accuracy, F1 score, and AUC, indicating it effectively captures both linear trends and nonlinear residual fluctuations in temperature. Interestingly, the negative correlation between the actual test values and hybrid predictions underscores the complex, short-term variability in urban temperature dynamics, suggesting that even the best model cannot perfectly anticipate all rapid changes, yet it remains the most reliable for classifying temperature state transitions.

Table 6. Diebold–Mariano test comparing forecast accuracy of ARIMA-LSTM against competing models

Model 1	Model 2	DM Statistic
ARIMA-LSTM	ARIMA	5.6540
ARIMA-LSTM	Prophet	4.8104
ARIMA-LSTM	Random Forest	4.3452
ARIMA-LSTM	Decision Tree	3.5689
ARIMA-LSTM	SVR	4.6577
ARIMA-LSTM	GBM	4.0952
ARIMA-LSTM	KNN	4.3350
ARIMA-LSTM	LSTM	3.1826
ARIMA-LSTM	RNN	3.2595
ARIMA-LSTM	GRU	3.3527
ARIMA-LSTM	BiLSTM	3.2811
ARIMA-LSTM	BiGRU	3.3408
ARIMA-LSTM	BiRNN	3.2726
ARIMA-LSTM	MLP	2.8090

The Diebold–Mariano test results in Table 6 above clearly indicate that the ARIMA-LSTM hybrid model consistently outperforms all other evaluated forecasting models, with DM statistics well above the critical threshold for statistical significance at conventional levels ($p < 0.05$). The highest DM statistic of 5.654 against ARIMA confirms a substantial improvement in forecast accuracy, while comparisons with machine learning models such as Random Forest, GBM, and SVR also show significant superiority. These findings provide strong statistical evidence that ARIMA-LSTM is not only more accurate but also robust across alternative models, supporting its use as the preferred forecasting approach in this study.

Table 7. Paired sample t-test comparing ARIMA-LSTM with other forecasting models

Model 1	Model 2	t Statistic	p Value
ARIMA-LSTM	ARIMA	6.31	0.0000
ARIMA-LSTM	Prophet	5.57	0.0000
ARIMA-LSTM	Random Forest	4.65	0.0000
ARIMA-LSTM	Decision Tree	2.83	0.0074
ARIMA-LSTM	SVR	4.66	0.0000
ARIMA-LSTM	GBM	3.83	0.0005
ARIMA-LSTM	KNN	4.49	0.0001
ARIMA-LSTM	LSTM	3.36	0.0018
ARIMA-LSTM	RNN	3.96	0.0003
ARIMA-LSTM	GRU	3.75	0.0006
ARIMA-LSTM	BiLSTM	3.65	0.0008
ARIMA-LSTM	BiGRU	3.66	0.0008
ARIMA-LSTM	BiRNN	4.01	0.0003
ARIMA-LSTM	MLP	3.15	0.0032

The paired t-test results in Table 7 reveal that ARIMA-LSTM consistently outperforms all alternative forecasting models across the evaluated splits, with statistically significant differences in predictive accuracy ($p < .01$ for all comparisons). The highest t-statistics were observed when comparing ARIMA-LSTM to classical models like ARIMA ($t = 6.31$) and Prophet ($t = 5.57$), indicating a strong margin of improvement. These findings confirm the robustness of ARIMA-LSTM and provide rigorous statistical evidence supporting its superior forecasting performance over both traditional and machine learning-based approaches.

Table 8. Prediction interval coverage at 95 percent confidence level across competing models

Model	Prediction Interval Coverage
ARIMA	0.0789
Prophet	0.6579
ARIMA-LSTM	**0.9868**
Random Forest	0.5000
Decision Tree	0.6316
SVR	0.0526
GBM	0.5789
KNN	0.3421
LSTM	0.7105
RNN	0.9737
GRU	0.7895
BiLSTM	0.8421
BiGRU	0.7632
BiRNN	0.9211
MLP	0.9474

Note: Prediction interval coverage measures the proportion of observed temperature values that fall within the estimated 95 percent forecast intervals.

Table 8 reveals a striking contrast in how well different models capture the uncertainty in urban temperature forecasts. The ARIMA-LSTM hybrid demonstrates outstanding reliability, with nearly all observed temperatures (98.7%) falling within its 95 percent prediction intervals, highlighting its superior ability to quantify forecast uncertainty. Similarly, deep learning variants like BiRNN (92.1%) and MLP (94.7%) perform remarkably well, outperforming traditional statistical and machine learning models such as ARIMA (7.9%) and SVR (5.3%), which severely underestimate variability. Models like Prophet, RNN, and BiLSTM also show strong interval coverage, while simpler approaches including KNN and Random Forest provide only moderate reliability.

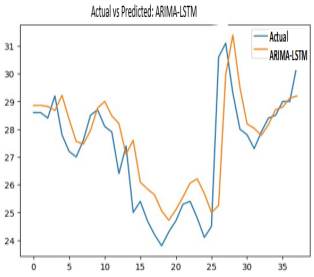


Figure 3. ARIMA-LSTM Plot

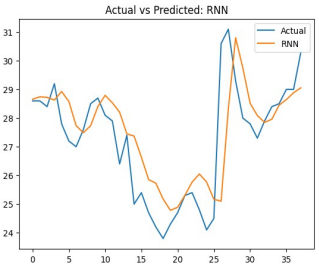


Figure 4. RNN Plot

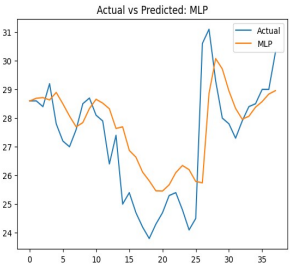


Figure 5. MLP Plot

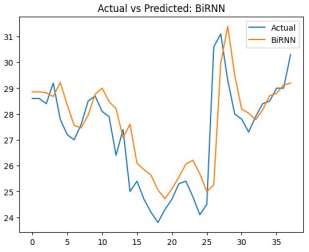


Figure 6. BiRNN Plot

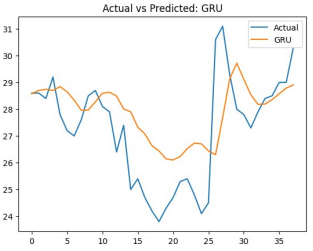


Figure 7. GRU Plot

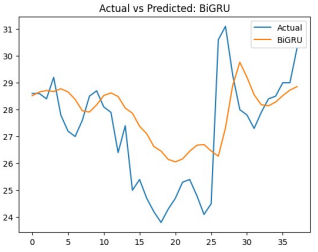


Figure 8. BiGRU Plot

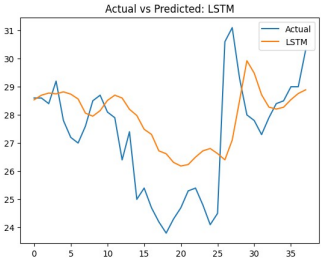


Figure 9. LSTM Plot

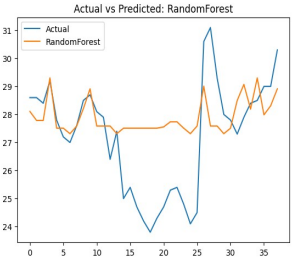


Figure 10. RF Plot

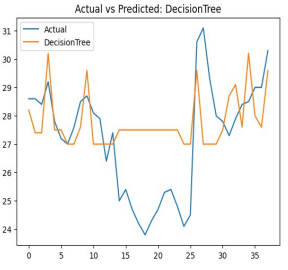


Figure 11. DT Plot

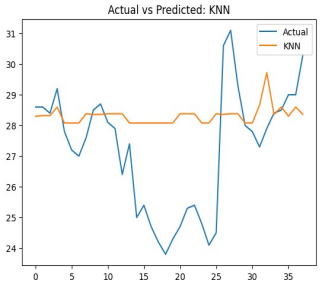


Figure 12. KNN Plot

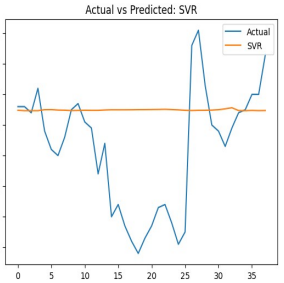


Figure 13. SVR Plot

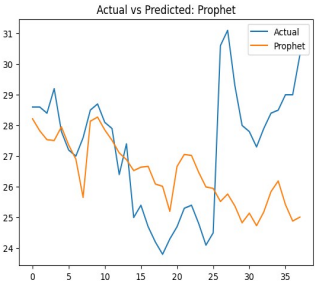


Figure 14. Prophet Plot

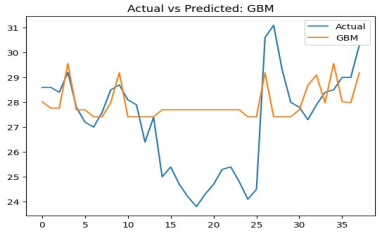


Figure 12. Actual vs Predicted GBM Plot

3.5. Forecast of the temperature dynamics in selected cities in Uganda

Table 9. Forecasted annual maximum temperature for selected Ugandan cities (2024–2030) using the ARIMA-LSTM hybrid model

City	2024	2025	2026	2027	2028	2029	2030
Arua	30.244	30.065	29.556	28.754	28.414	28.342	28.437
Entebbe	27.079	28.243	27.079	28.295	27.176	28.328	27.233
Gulu	31.211	31.487	31.169	30.728	30.335	30.027	29.883
Kampala	27.208	27.543	27.444	27.414	27.383	27.360	27.350
Kasese	30.579	30.833	30.642	30.781	30.732	30.756	30.754
Lira	30.974	31.105	31.420	31.165	31.234	31.271	31.229
Masindi	29.715	29.840	29.767	29.677	29.678	29.654	29.638
Jinja	28.302	28.188	28.117	28.069	28.051	28.031	28.017
Mbarara	27.281	27.085	27.232	27.086	27.084	27.089	27.067
Kabale	24.519	24.410	24.471	24.399	24.401	24.402	24.389
Soroti	30.161	29.689	29.095	28.279	27.910	27.895	28.179

Note: Values represent forecasted annual maximum temperature in degrees Celsius generated using the ARIMA-LSTM hybrid model. Forecasts span seven years beginning from 2024.

The forecasted temperature trends in Table 9 reveal notable variations across Ugandan cities over the seven-year period from 2024 to 2030. Northern and central cities such as Gulu, Lira, and Arua consistently show higher maximum temperatures, often exceeding 30°C, while cooler highland areas like Kabale remain around 24–25°C. The ARIMA-LSTM hybrid model captures both gradual long-term trends and short-term fluctuations, highlighting that some cities may experience slight year-to-year increases or dips in temperature rather than steady rises. This suggests that urban climate patterns in Uganda are influenced by complex interactions, possibly including population growth, local urban heat effects, and regional climatic variability, emphasizing the importance of city-specific planning for heat mitigation and adaptation strategies.

4. Discussion of findings

The findings of this study show that urban temperature dynamics in Ugandan cities are driven by a combination of persistent linear trends and short-term nonlinear fluctuations, and that this structure is best captured using a hybrid modeling approach. Descriptive results indicate consistently high urban temperatures, averaging close to 29°C, alongside substantial variation in urban population levels. The observed moderate positive correlation of about 0.47 between population size and temperature provides empirical support for the urban heat effect, where denser and rapidly growing cities experience elevated temperatures. This relationship aligns with established urban climate theory and corroborates recent African based studies by Kiarie et al. (2025) and Mutinda and Geletu (2025), who similarly documented population induced amplification of urban heat.

The methodological rigor of the study is reinforced by the stationarity and data quality diagnostics. Temperature series were found to be stationary at level, indicating dominance of short-term variability, while urban population series required first differencing due to their long run growth behavior. This justified modeling temperature directly while treating population effects cautiously to avoid spurious relationships. Grubbs outlier tests and visual diagnostics confirmed the absence of extreme observations in both datasets, supporting the reliability of the inputs. These findings are consistent with assumptions reported in prior temperature forecasting studies, including Nguyen et al. (2020), and provide a solid foundation for the subsequent modeling stages.

Performance comparisons across multiple models and validation schemes clearly establish the superiority of the ARIMA-LSTM hybrid framework. Across all-time series splits of 70:30, 80:20, and 90:10, the hybrid model consistently achieved the lowest RMSE and MAE values and the highest R^2 scores, demonstrating strong generalization

and addressing concerns related to overfitting associated with single train test evaluations. While standalone deep learning models such as RNN and MLP outperformed classical approaches in some cases, they remained consistently inferior to the hybrid model, confirming that nonlinear learning alone is insufficient when linear structure is ignored. Traditional models such as ARIMA, Prophet, SVR, and KNN exhibited higher errors and weaker explanatory power, echoing conclusions reported by Wang et al. (2018), Garcia et al. (2019), and Nguyen et al. (2020). Formal robustness checks further strengthened these findings, with Diebold Mariano tests and paired sample t tests confirming statistically significant performance gains for the ARIMA-LSTM model over all alternatives at the 1 percent significance level.

Beyond numerical accuracy, the results demonstrate strong practical relevance for decision making. The classification analysis showed that the ARIMA-LSTM model most reliably identified high temperature states, achieving the best balance of accuracy, F1 score, and AUC, which is critical for heat risk assessment and urban planning. Uncertainty analysis revealed near perfect 95 percent prediction interval coverage, indicating reliable representation of forecast uncertainty and outperforming several classical and standalone deep learning models. City specific forecasts further revealed persistent spatial contrasts, with northern cities such as Gulu, Lira, and Arua experiencing higher maximum temperatures, while highland cities like Kabale remaining relatively cooler. The presence of modest interannual variability alongside stable long-term trends highlight the value of combining ARIMA and LSTM components to capture both persistence and nonlinear behavior. Overall, the study contributes to African urban climate research by demonstrating, through multi split validation and formal statistical testing, that ARIMA-LSTM hybrid models provide a robust, interpretable, and generalizable framework for forecasting urban temperature dynamics in Uganda.

5. Validation of results, limitations, and policy implications

The study demonstrates that the ARIMA-LSTM hybrid model provides consistently superior performance for forecasting urban temperature dynamics across Ugandan cities. Validation was carried out using a broad set of regression and classification metrics, including RMSE, MAE, MAPE, SMAPE, R^2, accuracy, precision, sensitivity, F1 score, and AUC. To ensure robustness and address concerns about dependence on a single data split, time series aware rolling origin validation was applied across three split ratios of 70:30, 80:20, and 90:10. Across all splits, the ARIMA-LSTM hybrid achieved the lowest forecast errors and the highest explanatory power, with stable performance that indicates strong generalizability and reduced risk of overfitting. Statistical significance testing using the Diebold Mariano test and paired sample t tests

further confirmed that the hybrid model significantly outperformed all benchmark models, including classical statistical, machine learning, and deep learning approaches. These improvements were systematic and not driven by random variation or specific data partitions. In addition, a decision-oriented classification framework showed that the hybrid model translated numerical forecast accuracy into reliable identification of high temperature conditions, which is critical for early warning and risk management applications.

6. Conclusion

This study confirms that integrating ARIMA with LSTM in a hybrid framework offers a reliable and effective approach for forecasting urban temperature dynamics in Ugandan cities. By jointly capturing long-term linear trends and short-term nonlinear fluctuations, the ARIMA-LSTM model consistently outperformed traditional statistical methods and standalone machine learning models, highlighting the complex nature of urban climate behavior. The results underscore the value of combining statistical and deep learning techniques for climate prediction, while also pointing to the need for future research that incorporates longer time horizons and additional climatic variables such as rainfall and humidity. Importantly, the projected temperature increases, particularly in northern and central Uganda, signal an urgent need for city specific climate adaptation strategies, including urban greening, energy efficient infrastructure, and sustainable urban planning to reduce heat-related risks associated with climate change and rapid urbanization.

Recommendations

The findings of this study call for targeted and practical actions to address rising urban temperatures in Uganda. City specific climate adaptation plans are needed, especially in fast growing urban centers such as Gulu, Arua, and Lira, where heat risks are likely to intensify with continued population growth. Investment in green infrastructure including urban parks, tree planting, and green roofs should be prioritized as cost effective measures for reducing urban heat and improving thermal comfort. At the same time, sustainable urban planning practices that balance expansion with environmental protection are essential to limit heat exposure and excessive energy demand. From a research perspective, future studies should integrate additional climate variables such as humidity, rainfall, and wind speed to improve the realism and predictive strength of temperature models. Expanding the temporal coverage of datasets and including a wider range of urban environments will also enhance the models' capacity to capture long-term climate variability and rare extreme events, thereby strengthening their usefulness for policy and planning.

References

Chen, X., Liu, Y., (2018). Integrating ARIMA and LSTM for improved time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, No. 5, pp. 1923–1935.

Cheng, S., Shi, J., Cheng, Q., Zhou, X. and Zeng, S., (2025). Hybrid model for medium-term load forecasting in urban power grids. *Energies*, Vol. 18, No. 16, p. 4378. https://doi.org/10.3390/en18164378.

Garcia, M., Silva, R., Pereira, L., Santos, A. and Rocha, D., (2019). Urban climate resilience: Statistical modeling of temperature variability in Brazilian cities. *Climate Resilience Reviews*, Vol. 15, No. 1, pp. 150–165.

Gomez, F., Sanchez, M., Patel, R., Li, Q. and Zhao, Y., (2020). Hybrid statistical-machine learning models for environmental forecasting. *Journal of Environmental Science and Technology*, Vol. 18, No. 4, pp. 456–470.

Johnson, A., Smith, B. and Williams, C., (2016). Urban temperature dynamics: ARIMA modeling for climate forecasting. *Journal of Climate Change*, vol. 8, no. 3, pp. 210–225.

Jones, E., Brown, D., (2017). Limitations of ARIMA models in urban temperature forecasting. *Urban Climate Dynamics Review*, Vol. 12, No. 3, pp. 210–225.

Kiarie, J., Mwalili, S., Mbogo, R., Mutinda, J. and Langat, A. (2025). Statistical, machine learning, and deep learning models for COVID-19 forecasting in Kenya. *Computational and Mathematical Biophysics*, 13(1), 20250026. https://doi.org/10.1515/cmb-2025-0026.

Lee, X., Chen, Y., Huang, Z., Wang, M. and Zhang, L., (2017). Forecasting urban temperature dynamics: A comparison of Facebook Prophet and traditional models. *Environmental Science and Technology*, Vol. 12, No. 2, pp. 300–315.

Li, Z., Johnson, A., Perez, M., Brown, C. and Taylor, S., (2019). Urban climate resilience strategies: A review of recent literature. *Climate Resilience Reviews*, Vol. 15, No. 4, pp. 321–335.

Lynda, D., Logeswari, G., Tamilarasi, K. and Rakesh, S., (2025). Hybrid Bayesian deep learning model for predicting urban heat island intensity in African cities. *Scientific Reports*, Vol. 15, Article No. 31280. https://doi.org/10.1038/s41598-025-13492-4.

Mutinda, J. K., Geletu, A. (2025). Stock market index prediction using CEEMDAN-LSTM-BPNN-decomposition ensemble model, *Journal of Applied Mathematics*, *7706431*, 32 pages, https://doi.org/10.1155/jama/7706431.

Nguyen, H., Smith, J., Wang, T., Chen, L. and Lee, X., (2021). Advancements in environmental modeling using hybrid models. *Environmental Modeling and Software*, Vol. 30, No. 2, pp. 89–102.

Nguyen, T., Huynh, Q., Tran, P., Le, M. and Vo, D., (2020). Machine learning for urban temperature forecasting: A comparative study with traditional models. *Journal of Computational Statistics*, Vol. 30, No. 3, pp. 500–515.

Smith, A., Jones, B. and Brown, C., (2018). Urban temperature dynamics: Challenges and opportunities. *Journal of Climate Change*, Vol. 5, No. 2, pp. 123–135.

Tang, S., Jiang, Y., Su, H., Lim, K. M., Zheng, Z. and Zhu, Y., (2025). Wear defect detection of hydraulic pump using a hybrid method of VGG and LSTM. *International Journal of Hydromechatronics*, Forthcoming Articles, https://doi.org/10.1504/IJHM.2025.10073233.

Taylor, S. J., Letham, B., (2017). Facebook Prophet: A forecasting tool for the R programming language. *Journal of Computational Statistics and Data Visualization*, Vol. 25, No. 3, pp. 301–315.

Wang, H., Liu, J., Zhao, F., Chen, Q. and Yang, S., (2018). Hybrid ARIMA-LSTM modeling for urban temperature forecasting. *Journal of Environmental Science*, Vol. 25, No. 4, pp. 400–415.

Wang, L., Thompson, D., Green, S., Kim, J. and Lee, Y., (2020). Machine learning applications in environmental science: A comprehensive review. *Environmental Science: Processes & Impacts*, Vol. 22, No. 7, pp. 1567–1583.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 185–204, https://doi.org/10.59139/stattrans-2026-023
Received – 18.12.2024; accepted – 09.02.2026

Comparison of two types of topological networks for the foreign exchange market: one based on correlation coefficients and the other on the concept of causality

Joanna Landmesser-Rusek[1]

Abstract

Topological networks make it possible to recognize structural properties of the currency market. Such networks can be constructed on the basis of the values of correlation coefficients between currency pairs, and the popular minimum spanning tree (MST) algorithm allows an understanding of significant relationships on the market. An alternative measure of distance, based on the concept of causality for time series, makes it possible to measure not only the strength of relationships between currency pairs but also the directionality of these relationships. There is even an equivalent of MST on a directed graph – minimum-cost arborescence (MCA). The purpose of this study is to compare correlation and causality networks built for the foreign exchange market. The networks were constructed in a stepwise manner for the most important world currencies in the period from 3rd Jan. 2020 to 18th Oct. 2024. The comparison was carried out using certain topological characteristics of the networks, such as density, average distance, diameter, centralization index and degrees of vertices. The study details the properties of both approaches.

Key words: topological networks, foreign exchange market, correlation, causality.

1. Introduction

Network analysis focuses on modeling various phenomena from the real world as a system. Network tools are used to study, for example, social networks, financial networks or computer networks. In each of these cases, a complex system is represented as a network with nodes playing the role of agents and edges representing the interconnections between nodes. By analyzing the connections between nodes in the topological structure of a network, hidden information about the network can be

[1] Institute of Economics and Finance, Warsaw University of Life Sciences—SGGW, Warsaw, Poland. E-mail: joanna_landmesser@sggw.edu.pl. ORCID: https://orcid.org/0000-0001-7286-8536.

obtained. In the context of financial networks, such studies are conducted based on statistical relationships between series of returns.

Complex networks have been the subject of much research since the late 1990s. At first, small-world networks and scale-free networks were studied. The former refer to social networks in which the nodes and edges are people and their interactions (Watts and Strogatz, 1998). Here, individuals can be connected to each other with just a few random connections. The term scale-free network refers to a network whose degree distribution follows a power law (Barabási and Bonabeau, 2003). Both of these structures point to the unique features of the network. In particular, they emphasize that the real-world network does not have a random topology, but rather a centralized one with several hubs. The impact of the outbreak of certain events on network topology has also been studied. A network's topology is important for its resilience to external perturbations, such as failures or attacks (Albert et al., 2000; Cohen et al., 2000). For example, research has been conducted on the spread of diseases (Liu et al., 2004). Social problems such as traffic (e.g. Wu et al., 2008) and mobile communication (e.g. Hidalgo and Rodriguez-Sickert, 2008) have been solved using the network. The distribution of edges has led to consideration of the network community structure (Fortunato, 2010; Porter et al., 2009). Research has been conducted on the flow of information through nodes (Liu et al., 2016). The time-varying characteristics of networks were also analyzed (e.g. Palla et al., 2007).

Mantegna (1999) and Mantegna and Stanley (1999) introduced networks to the financial literature as a way to deal with the scale and number of complex relationships between economic agents. In this article, we attempt to model the structure of the global foreign exchange market. In the global foreign exchange market, the economic situation of each country and the interest rate policies implemented affect the exchange rates of neighbouring countries.

The purpose of the study is to compare two types of topological networks for the foreign exchange market: those based on correlation coefficients and those based on the concept of Granger causality for time series. We construct both correlation networks and causality networks, utilizing both undirected and directed edges. While correlation networks measure the strength of relationships between currency pairs, causality networks feature links that reflect significant directional effects from one currency to another. Beyond mere comparison, the study aims to demonstrate that causality networks act as a complementary framework to traditional correlation models. While the latter are effective for identifying overall market integration, the causality-based approach provides a deeper layer of 'informational richness' by uncovering the hidden directional architecture and lead-lag relationships of the FX market.

To achieve the general aim of this study, we formulate the following specific research questions:

Q1: do causality-based networks exhibit systematically different topological dynamics during crisis periods than correlation-based networks?

Q2: does the inclusion of directionality in the analysis change the identification of key 'driver' currencies in the global market compared to traditional undirected structures?

Q3: does the causality-based approach provide economically meaningful insights regarding market contagion and leadership that cannot be captured by analyzing correlations alone?

The comparative analysis of the two approaches is based on three criteria:

- Informational richness: The ability to identify lead-lag relationships and the direction of shock transmission.
- Sensitivity to market shifts: How quickly the network topology (e.g. density or diameter) responds to geopolitical shocks.
- Interpretability: The extent to which the resulting tree structure aligns with known economic dependencies (e.g. regional trade blocks).

Currency topological networks have already been studied, for example by McDonald et al. (2005), Ortega and Matesanz (2006), Naylor et al. (2007), Górski et al. (2008). Many researchers have focused on the structural evolution of the foreign exchange market during periods of crisis. Jang et al. (2011) and Feng and Wang (2010) noted that the correlation coefficient between currencies decreased during crises, while the length of the tree increased. Wang et al. (2012) studied the position of dominant world currencies. The correlation structure of the currency network was also studied by Kazemilari et al. (2018), Cao et al. (2020), Miśkiewicz (2021). An example of the research on the evolution of the currency network in the context of COVID-19 is the one conducted by Gupta and Chatterjee (2020). Another type of network is Granger causality networks, which are a common tool in mapping the human brain (Bullmore and Sporns, 2009), but they are also used in the financial literature, for example in the work of Billio et al. (2011), Výrost et al. (2015), Park et al. (2020), Jiang et al. (2022).

Although correlation-based MSTs and Granger causality networks have been individually applied to financial markets, there is a notable lack of comparative studies that evaluate their consistency and divergence during concurrent global shocks. Most existing literature treats these methods as alternatives rather than complements. Our study fills this gap by examining whether the 'causal architecture' of the FX market remains robust when the 'correlation architecture' shifts, thereby providing a more granular view of currency leadership that standard MST models cannot provide.

In the study, we analyzed changes in network topology based on time series of exchange rates of the most important world currencies from the period 03/01/2020 – 18/10/2024. Networks were constructed for the whole period as well as for 100-day rolling subsamples. Then the topological characteristics of the obtained graphs were analyzed.

The paper is organized as follows: Section 2 presents the two methods of analysis: correlation and causality networks, Section 3 describes the data used, Section 4 presents empirical results for currency market modeling obtained using both approaches, and Section 5 concludes.

2. Method of analysis

2.1. Correlation networks

An undirected graph is an ordered pair $G = (V, E)$, where V is the set of vertices (nodes) and E is the set of edges, which are two-element subsets of V : $E \subseteq \{\{u, v\}: u, v \in V\}$. A weighted graph is a graph in which each edge is assigned a weight that is some number (usually non-negative): $G = (V, E, w)$, where $w: E \to \mathbb{R}$.

One way to determine weights for edges in graphs (distances between nodes $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$) is to use Pearson's linear correlation coefficient:

$$\rho_p(X, Y) = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2}} \in [-1,1] \quad (1)$$

$$d_p(X, Y) = \sqrt{2[1 - \rho_p(X, Y)]} \in [0,2] \quad (2)$$

Links between objects in correlation networks can be defined in various ways. Pearson's linear correlation coefficient is preferred (Mantegna and Stanley (1999), Mizuno et al. (2006), Naylor et al. (2007), Jang et al. (2011)). Alternatively used measures of correlation are Spearman's rank correlation coefficient, Kendall's coefficient, partial correlation coefficient (Kenett et al., 2010; Basnarkov et al., 2019) or the coefficient of tail dependence estimated from the copula function (cf. Marti et al., 2021).

Starting from the correlation matrix between returns, the most important correlations can be extracted, resulting in a substantially sparser representation. Two approaches to filtering out the most important relationships are: (i) hierarchical methods and (ii) threshold methods. Among hierarchical methods, minimum spanning trees (MSTs) are the best known. MSTs, introduced by Kruskal (1956), compress information about network structure and simplify analysis by lowering the number of elements to be compared.

A spanning tree is a graph that is consistent and acyclic, i.e. there is a path between any two vertices and it is the only possible path between them. A minimum spanning

tree T for a weighted undirected graph $G = (V, E, w)$ is a spanning tree (containing all vertices in the set V) for which the sum of the weights of all edges

$$w(T) = \sum_{(u,v)\in T} w(u, v) \tag{3}$$

is minimal (a minimum cost spanning tree). MST has $|V|$-1 edges, where $|V|$ is the number of vertices. To find MST, the Prim and Kruskal algorithms are used. Based on the topology of the network, a hierarchical structure can be built using the Girvan-Newman method, which uses the so-called edge betweenness.

2.2. Granger causality networks

A directed graph is an ordered pair $G = (V, A)$, where V is the set of vertices and A is the set of directed edges (arcs), which are two-element subsets of V, with edge $\{a, b\}$ understood to be directed from vertex a to b.

In weighted directed networks, the weights may be determined by testing causality in the Granger sense (Granger 1969, 1980). By definition, a variable X is a cause of Y in the Granger sense if current values of Y can be predicted with greater accuracy using past values of X than without using them, with the remaining information unchanged. In the linear Granger causality test for pairs of variables, we estimate the equations of a VAR model with an equal number of lags for both variables, k, and apply a test of the joint significance of the lags of a given variable in the equation explaining the other variable:

$$y_t = \alpha_{10} + \sum_{j=1}^{k} \alpha_{1j} y_{t-j} + \sum_{j=1}^{k} \beta_{1j} x_{t-j} + \varepsilon_{1t} \tag{4}$$

$$x_t = \alpha_{20} + \sum_{j=1}^{k} \alpha_{2j} x_{t-j} + \sum_{j=1}^{k} \beta_{2j} y_{t-j} + \varepsilon_{2t} \tag{5}$$

$H_0: \beta_{11} = \beta_{12} = \cdots = \beta_{1k} = 0$ means there is no causal relationship in the Granger sense from X to Y (X is not the cause of Y). $H_0: \beta_{21} = \beta_{22} = \cdots = \beta_{2k} = 0$ means no causal dependence in the Granger sense from Y to X (Y is not the cause of X). In this paper, causality testing was performed using a wrapper in R: HDGC_VAR_all_I0 (Granger Causality Network in High Dimensional Stationary VARs) for stationary time series (I(0)), with lag k=1. The choice of the lag length was motivated by both economic and statistical considerations. In highly liquid foreign exchange markets, information is processed rapidly, and the primary interactions typically materialize within a single trading day. Statistically, we verified the optimal lag length for a representative subsample of pairs using the AIC criterion, which consistently pointed to k=1 or k=2 as the most appropriate structure. To maintain consistency and parsimony across all rolling windows and currency pairs in the network construction, a uniform lag of k=1 was adopted.

While foreign exchange markets often exhibit nonlinearities and tail dependencies, we employ the linear Granger causality framework as a robust baseline for identifying

directional information flows in the conditional mean. This approach allows for a direct comparison with correlation-based networks and ensures the tractability of the resulting topological metrics. However, we acknowledge that this captures primarily short-term predictive relationships, and future research could extend this by employing non-linear or frequency-domain causality tests to capture higher-moment dependencies.

The intention of someone, given a directed graph $G = (V, A)$, would be to find a minimal spanning tree on it. However, this task is problematic. In MST, all vertices should be connected, and in a directed graph, not every node is reachable from every other node. Thus, directed graphs do not satisfy the requirement that all vertices are connected.

The equivalent of a minimum spanning tree on a directed graph is a spanning arborescence of minimum weight (MSA) or otherwise optimum branching. An *r*-arborescence of a graph *G* is a directed tree *T* that contains a directed path from a specified node *r* to each node of a subset V' of the set $V \setminus \{r\}$. The node *r* is called the root of the arborescence. The algorithm for finding MSA is the Chu-Liu/Edmonds algorithm.

The construction of undirected and directed graphs in R is possible thanks to functions available in the *igraph* (the *mst* function allows the creation of MSTs) and *optrees* (*msArborEdmonds* allows building of MSAs) packages.

2.3. Topological characteristics of networks

Topological network indexes were used to study the dynamically changing structure of constructed networks. The following measures were applied at the level of the entire graph:

- density – the ratio of the number of edges ($|E|$) to the largest possible number of edges. Low density indicates greater independence between agents.

$$density(G) = \frac{|E|}{|V|(|V|-1)/2} \tag{6}$$

- mean distance (*apl*) – an average distance between all pairs of nodes in the graph. A low *apl* value indicates the efficiency of information flow in the network and indicates a structure susceptible to infection by negative events.

$$apl(G) = \frac{1}{|V|(|V|-1)} \sum_{i \neq j} d(v_i, v_j) \tag{7}$$

  $|V|$ – the number of vertices, $d(v_i, v_j)$ – the length of the shortest path between nodes *i* and *j*.

- diameter – the length of the longest shortest path between two nodes. A smaller diameter favors the transmission of information.

- degree centrality - the graph's centralization index regarding the number of links that nodes have. The higher it is, the higher the risk that nodes will intercept everything that flows through the network.

At the node level, the following indicators were used:

- degree – the degree of a vertex, i.e. the number of edges entering and leaving the vertex,
- in-degree – the input degree of a vertex, i.e. the number of edges entering the vertex,
- out-degree - the output degree of a vertex, i.e. the number of edges leaving the vertex.

To calculate the values of the above metrics, functions available in R within the *igraph* package were used.

3. Data used in the study

Daily data for the exchange rates of 15 currencies against the New Zealand Dollar (X/NZD) for the period 3/01/2020 – 18/10/2024 were obtained from https://stooq.com. A list of the currencies studied, along with their abbreviations, is shown in Table 1.

Table 1. List of the currencies studied

Abbreviation	Currency name	Abbreviation	Currency name
CAD	Canadian Dollar	KRW	South Korean Won
CHF	Swiss Franc	NOK	Norwegian Krone
CNY	Chinese Yuan	PLN	Polish Zloty
EUR	Euro	RUB	Russian Ruble
GBP	Pound Sterling	SEK	Swedish Krona
HKD	Hong Kong Dollar	SGD	Singapore Dollar
ILS	New Israeli Shekel	USD	US Dollar
JPY	Japanese Yen		

Source: authors' work.

The choice of base currency (numéraire) is a problem for which there is no standard solution. Currencies are valued against each other, so there is no independent numéraire. Different choices will yield different results. In this study, we chose the New Zealand Dollar (NZD) as the reference currency. This decision is grounded in the methodological framework proposed by Kwapień et al. (2009), who demonstrated that using a dominant global currency (like the USD or EUR) as a numéraire 'leads to a star-like MST structure (...) and does not represent the true relationships among the remaining currencies.' By choosing a liquid yet peripherally located currency that is not a 'driver' for a major regional trade bloc, we minimize the risk of artificial centralization and common-factor-induced distortions.

Furthermore, the stability of the network's topological hierarchy under this approach is supported by our previous research (Andrzejak et al., 2024), which specifically investigated the impact of different distance measures and reference frames on the consistency of currency networks. Our findings in that study confirmed that while the numéraire affects absolute correlation levels, the relative hierarchical positions of major currencies (such as the centrality of the SGD or EUR) remain robust. Consequently, the observed changes in network topology reported in this paper reflect genuine structural shifts in the market rather than artifacts of the data transformation. Additionally, the use of causality-based networks provides a further layer of robustness, as these measures are inherently less sensitive to the common-factor bias introduced by the numéraire compared to traditional correlation-based MSTs.

On the other hand, it is known that the choice of base currency strongly affects Pearson's linear correlation, while partial correlations would be invariant in this aspect (Basnarkov et al., 2019).

Prior to the analysis, all exchange rate series were transformed into logarithmic returns to ensure stationarity according to the formula:

$$R(t) = log\left(\frac{P(t+1)}{P(t)}\right) = log\, P(t+1) - log\, P(t) \tag{8}$$

Augmented Dickey-Fuller (ADF) tests were performed for all series across all sub-periods, confirming that the variables are I(0) at the 5% significance level. Furthermore, by employing a rolling window approach, we explicitly account for local non-stationarity and potential regime changes triggered by the pandemic and the conflict in Ukraine, allowing the network topology to evolve as market conditions shift.

4. Empirical results

In the constructed graphs, vertices represent exchange rates, while edge weights reflect the distances between series of returns for exchange rates. The edges in MSTs are undirected, while in Granger causality graphs they are directed. The networks were constructed for the entire period 3/01/2020 – 18/10/2024 as well as for 100-day rolling subsamples with a two-week step. 115 minimum spanning trees and 115 Granger causality graphs with significant edges were built for 115 sub-periods with 15 vertices each.

The MSTs were built on the basis of correlation graphs, in which the distance between nodes was calculated based on the values of the correlation coefficients for the return series (formula (2)). The Granger causality test led to the construction of directed graphs, in which the edge weights were the p-values from the causality test for currency pairs. The full Granger graphs in a further stage of the analysis were reduced to graphs in which the edge weights satisfy the condition p-value<0.05 (rejecting H_0 of non-causality in the Granger test).

In the construction of Granger causality networks, a significance level of α=0.05 was applied to identify valid edges. We acknowledge the potential risk of Type I errors associated with multiple testing. However, the use of more restrictive corrections, such as the Bonferroni correction, often leads to excessive sparsity in financial networks, potentially masking meaningful structural information. To mitigate this, in the remainder of our study, we focus on the evolution of topological metrics over time rather than the existence of individual edges.

Figure 1 presents graphs for three arbitrarily selected sub-periods: VIII–XII.2020, II–VI.2021, XI.2021–III.2022. The first is a period of uncertainty related to the COVID-19 pandemic and the US presidential election. The second is a period of monetary easing with numerous social transfers and money printing around the world. The third includes Russia's invasion of Ukraine and the energy crisis. The events observed in the world have significantly affected the structure of the currency network. In particular, it can be seen that greater uncertainty in the markets has resulted in lower density in the causality graph.

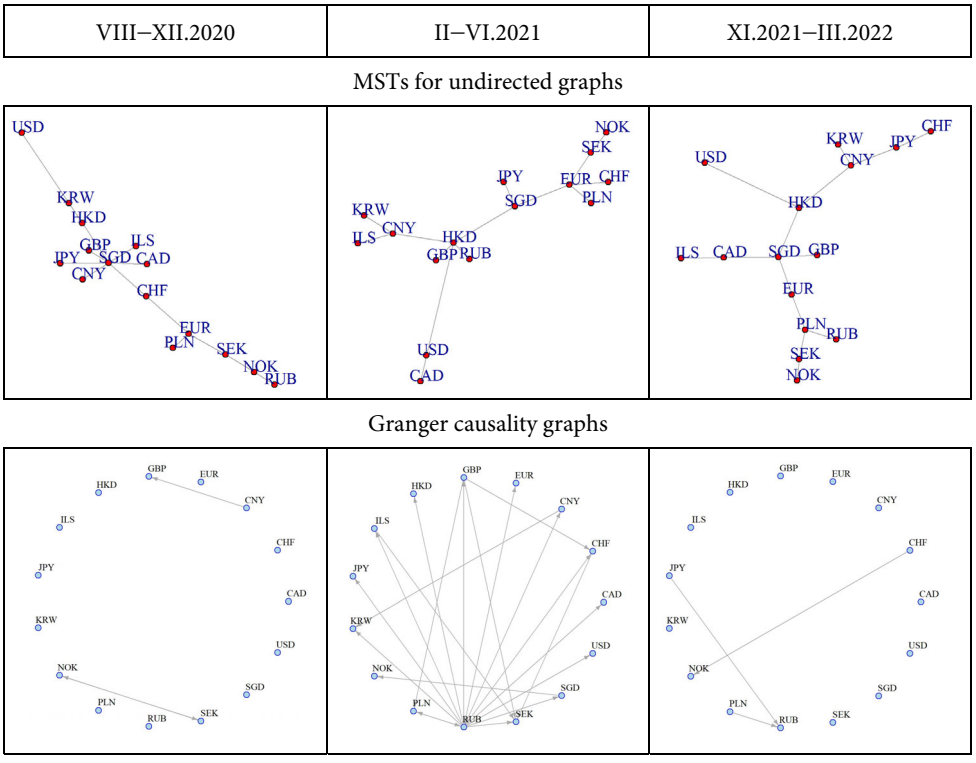


Figure 1. Structure of the currency network in three selected sub-periods: VIII–XII.2020, II–VI.2021, XI.2021–III.2022

Source: authors' work based on data from stooq.com (accessed October 21, 2024).

For all 230 networks obtained using the rolling window technique, the values of such topological characteristics as density, mean distance, diameter, and degree centrality were determined and presented in graphs (see Figure 2 for MSTs and Figure 3 for Granger networks).

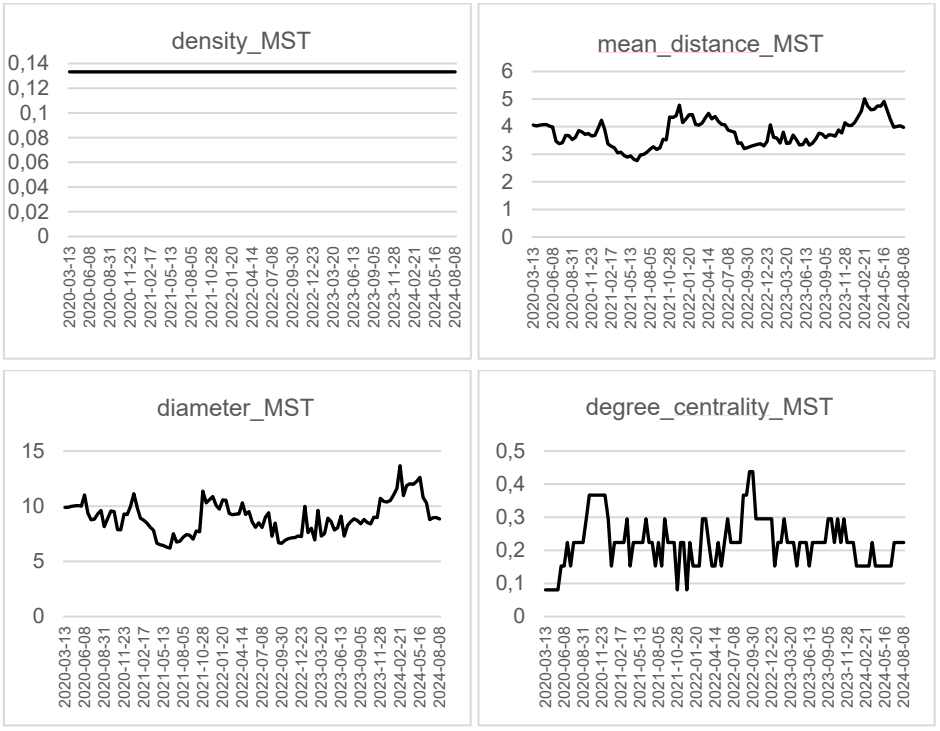


Figure 2. Selected characteristics of constructed minimum spanning trees (MST)
Source: authors' work based on data from stooq.com (accessed October 21, 2024).

It should be noted that MSTs for the currency market are characterized by a constant level of density (upper left panel of Figure 2). This follows from equation (6), where for MSTs we have $|E| = |V| - 1$. The remaining measures indicate different topological properties during the period considered. Lower average distance between nodes and lower graph diameter characterized the year 2021. Higher values for these measures occurred in 2020 (pandemic outbreak), 2022 (Russian invasion of Ukraine), and 2024. These years were also marked by lower graph centralization. It can be argued that during the difficult events of recent years, currencies were less interconnected (less correlated). Other authors have also noted that the correlation between currencies decreases during crises, while the length of the tree increases; see Jang et al. (2011), Feng and Wang (2010).

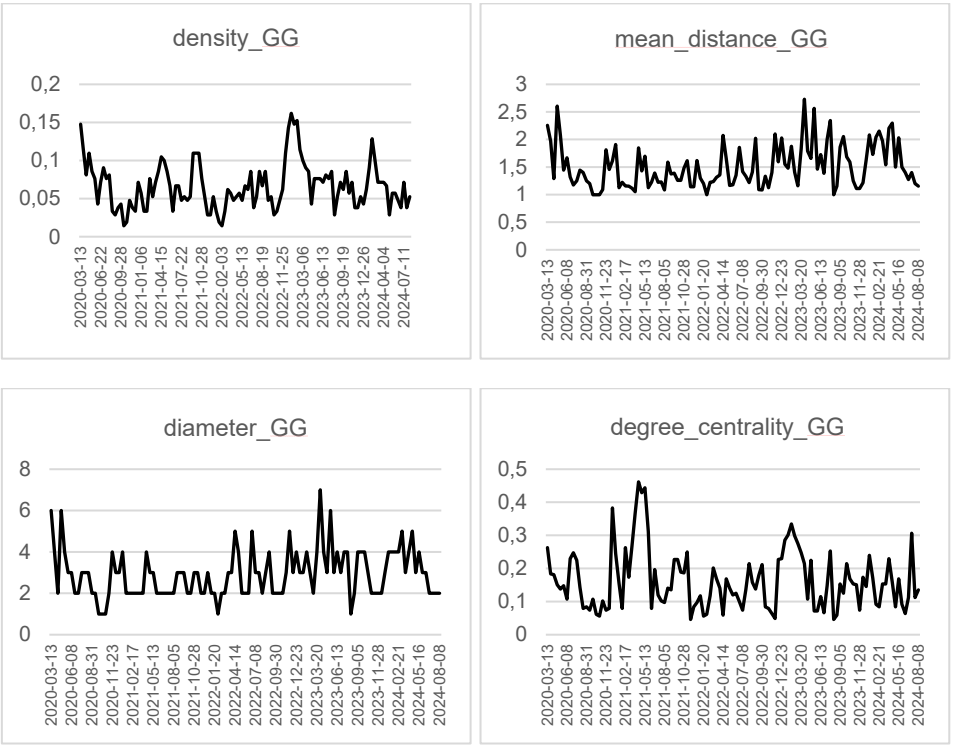


Figure 3. Selected characteristics of constructed Granger causality graphs (GG)
Source: authors' work based on data from stooq.com (accessed October 21, 2024).

In the case of Granger causality networks, their density is variable. The upper left panel of Figure 3 shows its successive decline caused by the freezing of economies and the fall in GDP in late 2020 (pandemic effect) and in February 2022 (Russian invasion). In contrast, peaks in density are visible in the spring and fall of 2021 (money transfers and commodity boom) and at the beginning of 2023 (growth of the money supply in the US, inflation). The density of the causality network appears to be lower during crisis periods. As a result of crises, the number of connections in networks decreases. A similar result was obtained by Park et al. (2020).

The values of the other network characteristics for MSTs and Granger graphs (GGs) are weakly correlated with each other. Table 2 presents Pearson correlations between the characteristics for the constructed graphs.

Analyzed separately, in both MSTs and GGs, the mean distance and longest shortest path (diameter) are strongly correlated. There is a positive correlation between density and mean distance (also between density and diameter, and centralization index) in Granger graphs.

Table 2. Pearson correlations between characteristics for graphs

Specification	density_ MST	mean_distance_ MST	diameter_ MST	degree_centrality_ MST	density_ GG	mean_distance_ GG	diameter_ GG	degree_centrality_ GG
Density_MST	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Mean_distance_MST		1.00	0.90***	-0.46***	-0.13	0.20**	0.16*	-0.30***
Diameter_MST			1.00	-0.51***	-0.11	0.23**	0.16*	-0.23**
Degree_centrality_MST				1.00	-0.23**	-0.30***	-0.29***	-0.06
Density_GG					1.00	0.45***	0.45***	0.61***
Mean_distance_GG						1.00	0.94***	0.14
Diameter_GG							1.00	0.12
Degree_centrality_GG								1.00

*, **, *** denote statistical significance at the 10%, 5%, and 1% level, respectively

Source: authors' own calculations based on data from stooq.com (accessed October 21, 2024).

MSTs are characterized by a negative correlation between graph centralization and mean distance (as well as the longest shortest path). This is because the lack of connection between nodes lengthens the path. In causality graphs, there is no significant correlation between graph centralization and average distance (and also the diameter). Interesting and similar considerations on this topic can be found in Park et al. (2020).

There seems to be no relationship between degree centralization in MSTs and Granger graphs (the correlation coefficient is close to zero). Low density in GGs harmonizes with a high centralization index in MSTs. The relationship between density in GGs and mean distance in MSTs is also negative but not statistically significant.

From an economic perspective, the shift in the causality-based networks structure during the Ukraine war suggests that causality is not just a statistical artifact but reflects a reconfiguration of safe-haven flows. Unlike MST, which only shows that currencies move together, our causality analysis identifies which currencies triggered the movement, providing a 'early warning' signal for systemic risk.

It is worth looking at the degrees of vertices in the constructed graphs. Table 3 contains the mean vertex degrees and the estimates along with statistical significance information for the coefficient β in the $degree_t = \alpha + \beta \cdot t + \varepsilon_t$ model.

Bold values refer to vertices with the highest degree (SGD, EUR, HKD in MSTs; RUB, CAD, SGD, and JPY in GGs). The causality networks allowed us to obtain valuable additional information about the in/out-degrees of the vertices. Values in italics are for the most "influential" – CHF, ILS, RUB, SGD – and the least "susceptible to influence" – CHF, EUR, ILS – vertices. On the other hand, the most "influence-prone" and least "influential" currencies turned out to be NOK, CAD, GBP, JPY, and PLN. The

estimated trend models identified the weakening influence of CNY and USD, the strengthening position of GBP and NOK, and the weakening position of JPY in the network.

Table 3. Degrees of vertices in constructed graphs

Specification	Degree MST		Degree GG		In-degree GG		Out-degree GG	
	mean	**trend**	**mean**	**trend**	**mean**	**trend**	**mean**	**trend**
CAD	1.28	-0.002	**2.20**	0.007	1.19	0.006	1.01	0.001
CHF	1.53	-0.005**	1.83	-0.003	*0.64*	-0.001	*1.18*	-0.002
CNY	2.14	0.002	1.92	-0.010**	1.08	0.006*	0.84	-0.016***
EUR	**3.85**	0.011***	1.54	0.008	*0.48*	-0.003	1.06	0.011**
GBP	1.14	0.003***	1.48	-0.004	0.84	-0.008***	0.63	0.004
HKD	**2.64**	-0.013***	1.88	-0.002	0.87	0.001	1.01	-0.003
ILS	1.03	-0.001***	1.74	0.010	*0.63*	-0.001	*1.11*	0.011*
JPY	1.17	-0.005***	**2.10**	0.018***	1.14	0.009***	0.96	0.009**
KRW	1.03	0.000	1.93	0.004	1.10	0.006*	0.83	-0.002
NOK	1.07	-0.003***	1.94	-0.007	1.38	-0.010***	0.56	0.003
PLN	1.34	0.003	1.09	-0.001	0.77	0.000	0.31	-0.001
RUB	1.02	0.000	**2.45**	-0.022**	0.99	-0.004	*1.46*	-0.018**
SEK	2.03	0.000	1.79	0.009*	0.96	-0.002	0.83	0.011***
SGD	**4.50**	0.000	**2.12**	0.017***	0.97	0.007***	*1.16*	0.010**
USD	2.23	0.009***	1.84	-0.011**	0.89	0.001	0.96	-0.012***

*, **, *** denote statistical significance at the 10%, 5%, and 1% level, respectively

Source: authors' own calculations based on data from stooq.com (accessed October 21, 2024).

The observed fluctuations in currency positions within the MST and causality networks correspond to major shifts in the global macroeconomic landscape. The weakening centrality of the USD during certain sub-periods of the pandemic may reflect the extraordinary monetary easing by the Federal Reserve and a temporary shift toward regional safe-haven assets. Similarly, the declining influence of the CNY can be attributed to China's 'Zero-COVID' policy and subsequent property market crises, which disrupted international trade structures and dampened the currency's role as a regional anchor. In contrast, the strengthening roles of the GBP and NOK during the 2022–2023 period are closely tied to the divergence in monetary policy and commodity market shocks. The GBP's increased centrality coincided with the Bank of England's early and aggressive stance against surging inflation. Meanwhile, the NOK's elevated position in the causality network following the outbreak of the war in Ukraine reflects its status as a key energy-linked currency. As European energy markets faced unprecedented volatility, the NOK became a primary channel for shock transmission, acting as a barometer for regional energy security and capital flows. These findings suggest that topological centrality is not merely a statistical artifact but a reflection of a currency's susceptibility to - and influence over - global macroeconomic shocks.

Unfortunately, our study did not confirm the result of Výrost et al. (2015) that currencies with low in-degree tend to have higher out-degree and vice-versa. In the equation *in-degree* =1.1018 − 0.1871·*out-degree*, the coefficient of the out-degree variable was statistically insignificant. Perhaps the reason for this is that the sample range was too broad and further research would require sub-period analyses.

The attempt to reduce the Granger causality networks to spanning arborescences of minimum weight (MSAs) cannot be considered successful, as the Chu-Liu/Edmonds algorithm resulted in multiple branches for each sub-period, depending on the starting vertex. The inability to construct stable and meaningful MSAs in our study constitutes a significant empirical finding in its own right. As suggested by the nature of the foreign exchange market, this outcome may stem from the absence of a single, natural 'root' currency in the global causal system. Unlike correlation networks, which can be effectively compressed into a MST, the causal structure of the FX market appears to be inherently non-hierarchical and decentralized. The failure of the MSA algorithm suggests that causal information transmission in the FX market cannot be simplified into a single directed tree without losing essential information about the complexity of the system. Furthermore, the use of p-values as edge weights, while statistically sound for identifying links, measures the significance rather than the economic strength of relationships, which may further complicate the identification of a stable arborescence. This supports the view that the FX market is a complex network of multi-directional flows rather than a simple hierarchical structure.

5. Conclusions

The purpose of the study was to compare two types of topological networks for the foreign exchange market: those based on correlation coefficients and those based on Granger's concept of causality. The study also aimed to demonstrate that causality networks act as a complementary framework to traditional correlation models. Accordingly, currency networks were built taking into account the strength of relationships between currency pairs and the directionality of these links.

The networks were constructed for the exchange rates of the world's 15 major currencies against the NZD, using a rolling window technique. The characteristics of the networks were analyzed.

This study set out to answer three research questions (Q1 – Q3) regarding the comparative advantages of causal versus correlation-based currency networks. Based on the empirical analysis, we formulate the following conclusions:

Ad Q1: The topological dynamics of causal networks exhibit significantly higher sensitivity to market shocks compared to correlation-based structures. While the cor-

relation network (MST) remained relatively stable, the causality-based network responded more dynamically to the outbreak of the COVID-19 pandemic and the conflict in Ukraine, showing rapid changes in connectivity and density.

Ad Q2: The inclusion of directionality changes the identification of the market's most influential nodes. Unlike the correlation approach, which only identifies pairs of co-moving currencies, the causality-based analysis allowed us to distinguish 'source' currencies (drivers) from 'sink' currencies (followers), effectively highlighting the dominant role of the CHF during periods of instability.

Ad Q3: The causality-based approach provides substantial 'informational richness' that cannot be obtained from correlation analysis alone. It acts as a complementary framework by uncovering the hidden directional architecture of the market and identifying the specific pathways of shock transmission (contagion channels), which are essential for effective risk monitoring.

Table 4 compares the two network approaches used and is the main result of the analysis.

Table 4. Comparison of correlational and causal approaches to network construction

Specification	Correlation networks	Causality networks
Idea	reflect the strength of relationships between pairs of currencies	reflect the strength and directionality of relationships between currency pairs
Preliminary assumptions	-	stationary time series
Consequences of the adopted distance measure	networks constructed based on Pearson's linear correlation coefficient are dependent on the assumed base currency; alternative: partial correlation coefficient	large choice of causality tests not necessarily leading to networks with identical topology
A way to simplify the relationship	minimum spanning tree (MST)	- reduction to graphs with edges with p-value<0.05, - minimum-cost arborescence (MSA) [but here: sensitivity to initial vertex = root]
Density and average distance	constant density in MSTs; high average distance in crises	during periods of uncertainty density decreases, and with expansionary monetary policy increases
Degrees of vertices	degree of vertex determined by the total number of incoming and outgoing edges	possible identification of in/out degrees of vertices => indication of “influence-prone” and “influential” nodes
Further applications	allow the construction of a hierarchical structure of the market	helpful in studying the dynamic propagation of shocks in the currency system

Source: author's own investigation.

The foreign exchange market is a complex system. Its complexity is accompanied by specific interdependence. In this article, we have proposed to quantify this interdependence using correlation networks and Granger causality networks. Correlation networks, and especially minimum spanning trees, provide broad insights into the linkages between currencies, while Granger causality networks capture the complex web of statistical relations between them. The linkages between currencies during the period under review were highly dynamic and changed over time depending on market and political conditions. The use of a wide range of tools to assess the topology of the networks allowed a better understanding of the phenomena taking place in the foreign exchange market.

Despite the insights provided by this comparative analysis, certain methodological limitations should be acknowledged. First, the results remain conditional on the chosen reference currency (numéraire). While the choice of NZD was a strategic decision to minimize artificial centralization, future research could employ numéraire-independent approaches. These might include valuing individual currencies against a weighted basket of currencies or utilizing partial correlations to further mitigate common-factor bias.

Second, the use of linear Granger causality focuses on dependencies in the conditional mean. While this provides a robust and interpretable baseline for identifying directional information flows, it may not capture nonlinear dynamics, threshold effects, or dependencies in higher moments of the distribution (e.g. volatility spillovers and tail behavior), which are characteristic of foreign exchange markets during periods of extreme stress. With nonlinear extensions of Granger causality, a higher degree of interconnectedness between currencies could potentially be uncovered. Consequently, the causality networks reported here should be interpreted as representations of local, short-term predictive relationships.

Finally, while we adopted a uniform lag length of k=1 - motivated by the high informational efficiency of global FX markets and statistical parsimony - interactions in less liquid markets or during specific regime changes might materialize with longer delays. Although our rolling window approach explicitly accounts for local shifts in market regimes (such as the pandemic or the conflict in Ukraine), future studies could explore the use of frequency-domain causality or time-varying lag structures to further refine the map of information transmission in the global financial system.

Treating the constructed Granger causality networks as a starting point for further studies, we intend to focus on tracking the evolution of vertex degrees for specific currencies and drawing conclusions in this regard.

References

Albert, R., Jeong, H., Barabási, A.-L., (2000). Error and attack tolerance of complex networks. *Nature*, Vol. 406, pp. 378–382.

Andrzejak, J., Chmielewski, L. J., Landmesser-Rusek, J. and Orłowski, A., (2024). The impact of the measure used to calculate the distance between exchange rate time series on the topological structure of the currency network. *Entropy*, Vol. 26, pp.1–17.

Barabási, A.-L., Bonabeau, E., (2003). Scale-free networks. *Scientific American*, Vol. 288, pp. 60–69.

Basnarkov, L., Stojkoski, V., Utkovski, Z. and Kocarev, L., (2019). Correlation Patterns in Foreign Exchange Markets. *Physica A: Statistical Mechanics and its Applications*, Vol. 525, pp. 1026–1037.

Billio, M., Getmansky, M., Lo, A. W. and Pelizzon, L., (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, Vol. 104, pp. 535–559.

Bullmore, E., Sporns, O., (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, Vol. 10, pp. 186–198.

Cao, H., Guo, Z., Li, Y. and Ran, Z., (2020). The Relationship Structure of Global Exchange Rate Based on Network Analysis. *Journal of Mathematical Finance*, Vol. 10, pp. 58–76.

Cohen, R., Erez, K., ben-Avraham and D., Havlin, S., (2000). Resilience of the Internet to Random Breakdowns. *Physical Review Letters*, Vol. 85, pp. 4626–4628.

Feng, X., Wang, X., (2010). Evolutionary topology of a currency network in Asia. *International Journal of Modern Physics C*, Vol. 21, pp. 471–480.

Fortunato, S., (2010). Community detection in graphs. *Physics Reports*, Vol. 486, pp. 75–174.

Górski, A. Z., Drożdż, S., Kwapień, J. nad Oświęcimka, P., (2008). Minimal Spanning Tree graphs and power like scaling in FOREX networks. *Acta Physica Polonica A*, Vol. 114, pp. 531–538.

Granger, C. W. J., (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, Vol. 37, pp. 424–438.

Granger, C. W. J., (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, Vol. 2, pp. 329–352.

Gupta, K., Chatterjee, N., (2020). *Examining Lead-Lag Relationships In-Depth, with Focus on FX Market as COVID-19 Crises Unfolds*, arXiv, arXiv:2004.10560.

Hidalgo, C. A., Rodriguez-Sickert, C., (2008). The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, Vol. 387, pp. 3017–3024.

Jang, W., Lee, J. and Chang, W., (2011). Currency crises and the evolution of foreign exchange market: Evidence from minimum spanning tree. *Physica A: Statistical Mechanics and its Applications*, Vol. 390, pp. 707–718.

Jiang, Z., Arreola Hernandez, J., McIver, R. P. and Yoon, S.-M., (2022). Nonlinear Dependence and Spillovers between Currency Markets and Global Economic Variables. *Systems*, Vol. 10, 80.

Kazemilari, M., Mohamadi, A., (2018). Topological Network Analysis Based on Dissimilarity Measure of Multivariate Time Series Evolution in the Subprime Crisis, International Journal of Financial Studies, Vol. 6, 47.

Kenett, D. Y., Tumminello, M., Madi, A., Gur-Gershgoren, G., Mantegna, R. N. and Ben-Jacob, E., (2010). Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLoS ONE*, Vol. 5, e15032.

Kruskal, J. B., (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, Vol. 7, pp. 48–50.

Kwapień, J., Gworek, S. and Drożdż, S., (2009). Structure and Evolution of the Foreign Exchange Networks. *Acta Physica Polonica B*, Vol. 40, pp. 175–194.

Liu, J., Wu, J. and Yang, Z. R., (2004). The spread of infectious disease on complex networks with household-structure. *Physica A: Statistical Mechanics and its Applications*, Vol. 341, pp. 273–280.

Liu, J., Xiong, Q., Shi, W., Shi, X. and Wang, K., (2016). Evaluating the importance of nodes in complex networks. *Physica A: Statistical Mechanics and its Applications*, Vol. 452, pp. 209–219.

Mantegna, R. N., (1999). Hierarchical structure in financial markets. *The European Physical Journal B: Condensed Matter and Complex Systems*, Vol. 11, pp. 193–197.

Mantegna, R. N., Stanley, H. E., (1999). *Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, Cambridge.

Marti, G., Nielsen, F., Bińkowski, M. and Donnat, P., (2021). *A Review of Two Decades of Correlations, Hierarchies, Networks and Clustering in Financial Markets*. In: Progress in Information Geometry. Signals and Communication Technology; Nielsen, F., (Ed); Springer: Cham, Switzerland.

McDonald, M., Suleman, O., Williams, S., Howison, S. and Johnson, N., (2005). Detecting a Currency's Dominance or Dependence using Foreign Exchange Network Trees. *Physical Review E*, Vol. 72, 046106.

Miśkiewicz, J., (2021). Network Analysis of Cross-Correlations on Forex Market during Crises. Globalisation on Forex Market. *Entropy*, Vol. 23, pp. 1–19.

Mizuno, T., Takayasu, H. and Takayasu, M., (2006). Correlation networks among currencies. *Physica A: Statistical Mechanics and its Applications*, Vol. 364, pp. 336–342.

Naylor, M. J., Rose, L. C. and Moyle, B. J., (2007). Topology of foreign exchange markets using hierarchical structure methods. *Physica A: Statistical Mechanics and its Applications*, Vol. 382, pp. 199–208.

Ortega, G. J., Matesanz, D., (2006). Cross-country hierarchical structure and currency crises. *International Journal of Modern Physics C*, Vol. 17, pp. 333–341.

Palla, G., Barabási, A. and Vicsek, T., (2007). Quantifying social group evolution, Nature, Vol. 446, pp. 664–667.

Park, J. H., Chang, W. and Song, J. W., (2020). Link prediction in the Granger causality network of the global currency market. *Physica A: Statistical Mechanics and its Applications*, Vol. 553, 124668.

Porter, M. A.; Onnela, J.-P. and Mucha, P. J., (2009). Communities in Networks. *Notices of the American Mathematical Society*, Vol. 56, pp. 1082–1097, 1164–1166.

Výrost, T., Lyócsa, Š. and Baumöhl, E., (2015). Granger causality stock market networks: Temporal proximity and preferential attachment. *Physica A: Statistical Mechanics and its Applications*, Vol. 427, pp. 262–276.

Wang, G. J., Xie, C., Han, F. and Sun, B., (2012). Similarity measure and topology evolution of foreign exchange markets using dynamic time warping method: Evidence from minimal spanning tree. *Physica A: Statistical Mechanics and its Applications*, Vol. 391, 4136–4146.

Watts, D., Strogatz, S., (1998). Collective dynamics of 'small-world' networks. *Nature*, Vol. 393, pp. 440–442.

Wu, J.-J., Gao, Z.-Y. and Sun, H.-J., (2008). Optimal traffic networks topology: A complex networks perspective. *Physica A: Statistical Mechanics and its Applications*, Vol. 387, pp. 1025–1032.

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 205–218, https://doi.org/10.59139/stattrans-2026-024
Received – 20.07.2024; accepted – 13.05.2025

R-optimal design strategies for logistic regression models with complementary log-log link

Tofan Kumar Biswal[1]

Abstract

The manuscript explores optimal experimental design strategies, specifically R-optimality, for two-parameter logistic regression (2PLR) models using the complementary log-log (c-loglog) link function. The study seeks to establish efficient designs that minimize the average width of confidence bands across the range of predictor variables. The general equivalence theorem validates the necessary and sufficient conditions of this optimality criterion.

Key words: logistic regression model, link function, R-optimality, equivalence theorem.

1. Introduction

Nelder and Wedderburn (1972) proposed the generalized linear model (GLM), which is a generalized version of the ordinary linear regression model. It has several uses in a variety of industries, including clinical trials, engineering, agriculture, economics, insurance, and many more. One can refer to the articles by Bailey et al. (1960), Myers and Montgomery (1997), de Jong and Heller (2008), Fox (2015), and Goldburd (2016) for further information on the uses of GLM. When conducting studies using categorical response types, Generalized Linear Models (GLMs) are typically employed. These models are widely used in many kinds of research when the researcher seeks to: (i) to discover the relationship between the number of encounters with other partners and explanatory factors and the risk of contracting HIV (Human Immunodeficiency Virus) (refer to Jewell and Shiboski, 1992); (ii) to look at the distribution pattern of important tree species; and (iii) to estimate the benefits of each particular treatment in a multicenter clinical trial (refer to Lee and Nelder, 2002). Agresti (2002) and McCullagh and Nelder (1989) gave a thorough explanation of GLM data analysis and its applications in several multidisciplinary fields.

The fundamental theoretical work of optimal designs was developed by Kiefer (1959), and Kiefer and Wolfowitz (1959). For further details, one can refer to the work of Atkinson et al. (2007). The main objective of obtaining an optimal design is to discuss statistical inference about the quantities of interest by selecting the control variable wisely. The values of the control variables are chosen to minimize the variability of the estimators of the unknown parameters involved in the regression model. It becomes difficult to discover the best design for the GLM since the information matrix depends on the unknown parameters;

[1]Department of Statistics, Central University of Odisha, Sunabeda 763004, India.
E-mail: tofankumarbiswal100@gmail.com. ORCID: https://orcid.org/0000-0003-4379-1298.

that is, one must know the parameters in order to find the best design, which requires estimating the unknown values. The standard approach of obtaining non-Bayesian optimal designs for generalized linear models is to use the best guess of the parameter values and derive the locally optimal designs [see Chernoff (1953)]. In many real-life problems, the prior information is often available in the form of historical data, expert opinion, etc. Therefore, an experimenter may utilize this prior information to obtain a Bayesian optimal design. The Bayesian optimal design problem is a statistical decision problem in which the design space, utility function, distribution of the random variables is generally involved. The major advantage of using the Bayesian optimal designs is that it helps an experimenter not only to guess the initial value of the parameter but also to incorporate the associated uncertainty.

Chaloner and Larntz (1989) discussed Bayesian D-optimal designs for the logistic regression model. Ford et al. (1992) found C- and D-optimal designs for the same model by using a geometric method. Sitter and Wu (1993) obtained D-, A-, and F-optimal designs whereas Dette and Haines (1994) discovered E-optimal designs for the same model with two parameters. Mathew and Sinha (2001) proposed a unified technique of D-, A-optimal design for same model. The best designs for two variable binary logistic models with interaction were reported by Dror and Steinberg (2006), and Haine et al. (2018). Numerical techniques were employed in the construction of these designs.

Recently, many authors have obtained R-optimal designs for different types of regression models, e.g. multi-response regression models with multiple factors (Liu et al., 2022), models with mixture experiments (Panda and Sahoo, 2022), gamma model with two parameters (Panda et al., 2024), logistic model with two variables (Panda and Biswal, 2024), Poisson model using log link (Biswal, 2024), and linear regression model with two variables (Biswal, 2024), etc. In this context, the present article aims to construct locally R-optimal designs for the Logistic Model with two parameters including the intercept parameter through complementary loglog link function.

The article is organized as follows. Section 2 presents the locally R-optimal designs and associated nomenclature. In Section 3, R-optimal strategies for the logistic regression model with two parameters are covered. In Section 4, the article comes to a close with some last thoughts and discussions. Finally, Section 5 concludes the article with concluding remarks.

2. Notation and R-optimal design

Consider a binary response variable Y which follows a Bernoulli distribution, that has a probability of "success" given by π_i and probability of "failure" given by $(1-\pi_i)$. The response Y is related to the predictor through binary logistic model i.e.

$$g(\pi_i) = \eta_i = \mathbf{f}^T(\mathbf{x}_i)\beta, \quad (i = 1,2,\ldots,q). \tag{1}$$

The aim is to select:

(i) the support points represented by the $s \times 1$ vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$ from a set, $\mathscr{S} \in \mathbb{R}^s$, of possible points.

(ii) the associated design weights $\omega_1, \omega_2, \ldots, \omega_q$.

So, the approximate design $\xi \in \Xi$ (Ξ is the set of all approximate designs) is defined by

$$\xi = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_q \\ \omega_1 & \omega_2 & \cdots & \omega_q \end{pmatrix}, \quad \omega_i > 0, \quad \sum_{i=1}^{q} \omega_i = 1. \tag{2}$$

where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q \in \mathscr{S}$ are the q distinct points and ω_i is the weight associated with the point $\mathbf{x}_i$ for $i = 1, 2, \ldots, q$.

For the model Equation (1), the Fisher information matrix of a design ξ at parameter vector β is defined as

$$M(\mathbf{x}_i, \beta) = \frac{\exp[2\eta_i - \exp(\eta_i)]}{1 - \exp[-\exp(\eta_i)]} \mathbf{f}(\mathbf{x}_i)\mathbf{f}^T(\mathbf{x}_i) \tag{3}$$

where

$$\frac{\exp[2\eta_i - \exp(\eta_i)]}{1 - \exp[-\exp(\eta_i)]}$$

is the complementary log-log link function. For more details, one can refer to Russell (2018, p. 92, section 4.3.3).

The information matrix evaluated at the approximate design follows immediately as

$$M(\xi, \beta) = \sum_{i=1}^{q} \omega_i M(\mathbf{x}_i, \beta) = \sum_{i=1}^{q} \omega_i \frac{\exp[2\eta_i - \exp(\eta_i)]}{1 - \exp[-\exp(\eta_i)]} \mathbf{f}(\mathbf{x}_i)\mathbf{f}^T(\mathbf{x}_i) \tag{4}$$

R-optimal Design: A design $\xi \in \Xi$ with a non-singular information matrix $M(\xi)$ is called R-optimal for the model Equation (1) if it minimizes

$$H(\xi) = \prod_{i=1}^{q} \left(M^{-1}(\xi)\right)_{ii} = \prod_{i=1}^{q} \mathbf{e}_i^T M^{-1}(\xi)\mathbf{e}_i \tag{5}$$

for all $\xi \in \Xi$. Here, $\mathbf{e}_i$ denotes the i^{th} unit vector in $\mathbb{R}^q$, where q is the number of unknown parameters associated with the model Equation (1).

The necessary and sufficient conditions for the R-optimality will be verified through the following equivalence theorem. For more details, one can refer to Dette (1997).

Equivalence theorem: For model Equation (1), let

$$E(\mathbf{x}, \xi) = \mathbf{f}^T(\mathbf{x})M^{-1}(\xi) \left(\sum_{i=1}^{q} \frac{\mathbf{e}_i \mathbf{e}_i^T}{\mathbf{e}_i^T M^{-1}(\xi)\mathbf{e}_i} \right) M^{-1}(\xi)\mathbf{f}(\mathbf{x}) \tag{6}$$

A design $\xi^* \in \Xi$ is R-optimal if and only if

$$\sup_{\mathbf{x} \in \mathscr{S}} E(\mathbf{x}, \xi^*) = q$$

with equality attained at the support points ξ^*.

3. R-optimal designs for two parameters

In this section, locally R-optimal designs for the model Equation (1) that involves two unknown parameters including the intercept parameter, i.e. $\mathbf{f}^T(\mathbf{x})\beta = \beta_0 + \beta_1 \mathrm{x} > 0$, for all $\mathrm{x} \in \mathrm{R}$. Here, we restrict our search by considering discrete values of β_0 and β_1 arbitrarily chosen intervals, i.e. $\beta_0 \in [1,5]$ and $\beta_1 \in [1,10]$.

3.1. Designs derived from two support points

Let us consider a 2-point design ξ of the form

$$\xi = \left\{ \begin{array}{cc} m & n \\ \omega & 1-\omega \end{array} \right\} \text{where } \ 0 < \omega < 1. \tag{7}$$

Theorem 3.1.1. The design ξ^* that assigns a weight of ω^* to the point m^* and $(1-\omega)^*$ to the point n^* in $\mathscr{S}$ is an R-optimal design where m^*, n^*, and ω^* are given in Table 2 (Appendix-II).

Proof. The information matrix for the model Equation (7) at the two-point design ξ defined in Equation (4) is given by

$$\boldsymbol{M}(\xi) = \left[\begin{array}{cc} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{array} \right] \tag{8}$$

with

$$\alpha_{11} = \frac{e^{-e^{\beta_0+\beta_1 n}+2(\beta_0+\beta_1 n)}(1-\omega)}{1-e^{-e^{\beta_0+\beta_1 n}}} + \frac{e^{-e^{\beta_0+\beta_1 m}+2(\beta_0+\beta_1 m)}\omega}{1-e^{-e^{\beta_0+\beta_1 m}}}$$

$$\alpha_{12} = \alpha_{21} = \frac{e^{-e^{\beta_0+\beta_1 n}+2(\beta_0+\beta_1 n)}n(1-\omega)}{1-e^{-e^{\beta_0+\beta_1 n}}} + \frac{e^{-e^{\beta_0+\beta_1 m}+2(\beta_0+\beta_1 m)}m\omega}{1-e^{-e^{\beta_0+\beta_1 m}}}$$

$$\alpha_{22} = \frac{e^{-e^{\beta_0+\beta_1 n}+2(\beta_0+\beta_1 n)}n^2(1-\omega)}{1-e^{-e^{\beta_0+\beta_1 n}}} + \frac{e^{-e^{\beta_0+\beta_1 m}+2(\beta_0+\beta_1 m)}m^2\omega}{1-e^{-e^{\beta_0+\beta_1 m}}}$$

The inverse of the above Fisher-information matrix is given by

$$\boldsymbol{M}^{-1}(\xi) = \left[\begin{array}{cc} \alpha^*_{11} & \alpha^*_{12} \\ \alpha^*_{21} & \alpha^*_{22} \end{array} \right] \tag{9}$$

with

$$\alpha^*_{11} = \frac{e^{-2(\beta_0+\beta_1(m+n))}\left(e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)n^2(\omega-1) - e^{2\beta_1 m}(-1+e^{e^{\beta_0+\beta_1 n}})m^2\omega\right)}{(m-n)^2(\omega-1)\omega}$$

$$\alpha^*_{12} = \alpha^*_{21} = \frac{e^{-2(\beta_0+\beta_1(m+n))}\left(-e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)n(1-\omega) + e^{2\beta_1 m}(-1+e^{e^{\beta_0+\beta_1 n}})m\omega\right)}{(m-n)^2(\omega-1)\omega}$$

$$\alpha_{22}^{*}=\frac{e^{-2(\beta_0+\beta_1(m+n))}\left(e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)(\omega-1)-e^{2\beta_1 m}(-1+e^{e^{\beta_0+\beta_1 n}})\omega\right)}{(m-n)^2(\omega-1)\omega}$$

Using Equation (5), we obtain the function

$$H(\xi)=\frac{\begin{array}{c}e^{-4(\beta_0+\beta_1(m+n))}\left(e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)(\omega-1)-e^{2\beta_1 m}(e^{e^{\beta_0+\beta_1 n}}-1)\omega\right)\\ \times\left(e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)n^2(\omega-1)-e^{2\beta_1 m}(e^{e^{\beta_0+\beta_1 n}}-1)m^2\omega\right)\end{array}}{(m-n)^4(\omega-1)^2\omega^2} \tag{10}$$

Now, the problem is to minimize the function $H(\xi)$ with respect to m, n, and ω for given values of β_0 and β_1. This is accomplished by utilizing Mathematica's "NMinimize" function and getting the optimal values m^*, n^*, and ω^*. The numerical values of m^*, n^*, and ω^* are given in Table 2 (Appendix-II).

Next, by using Equation (9) we derive the quadratic form as specified in Equation (6) which is as follows:

$$E(\mathbf{x},\xi^*)=j\times\left\{\alpha_{11}^{*}+\alpha_{12}^{*}x+\left(\frac{\lambda_1\left(\alpha_{12}^{*}+\alpha_{22}^{*}x\right)}{\lambda_2}\right)+x\left(\alpha_{12}^{*}+\alpha_{22}^{*}x+\frac{\lambda_1\left(\alpha_{11}^{*}+\alpha_{12}^{*}x\right)}{\lambda_3}\right)\right\} \tag{11}$$

with

$$j=\frac{\exp\left[2\eta-\exp(\eta)\right]}{1-\exp\left[-\exp(\eta)\right]}$$

$$\lambda_1=\left(-e^{2\beta_1 n}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)n(\omega-1)+e^{2\beta_1 m}\left(e^{e^{\beta_0+\beta_1 n}}-1\right)m\omega\right)$$

$$\lambda_2=\left(e^{2\beta_1 n}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)(\omega-1)+e^{2\beta_1 m}\left(e^{e^{\beta_0+\beta_1 n}}-1\right)\omega\right)$$

$$\lambda_3=\left(e^{2\beta_1 n}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)n^2(\omega-1)-e^{2\beta_1 m}\left(e^{e^{\beta_0+\beta_1 n}}-1\right)m^2\omega\right)$$

Replacing the numerical values of m^*, n^*, and ω^* in Equation (11) and using the Mathematica software, we obtain

$$\sup_{\mathbf{x}\in\mathscr{S}}E(\mathbf{x},\xi^*)=2,$$

which is nothing but the number of unknown parameters. This provides the necessary and sufficient condition of the equivalence theorem. This proves Theorem 3.1.1.

3.2. Designs derived from three support points

Let us consider a 3-point design ξ of the form

$$\xi=\left\{\begin{array}{ccc} m & n & o\\ \omega/2 & 1-\omega & \omega/2\end{array}\right\}\ \text{where } 0<\omega<1. \tag{12}$$

Theorem 3.2.1. The design ξ^* that assigns a weight of $(\omega^*/2)$ to the point m^* , $(1-\omega)^*$

to the point n^* and $(\omega^*/2)$ to the point o^* in $\mathscr{S}$ is an R-optimal design where m^*, n^*, o^*, and ω^* are given in Table 3 (Appendix-II).

Proof. Using Equation (4), the information matrix for the model Equation (12) at the three-point design ξ will be

$$\boldsymbol{M}(\xi) = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \tag{13}$$

with

$$a = \frac{e^{-e^{\beta_0+\beta_1 n}+2(\beta_0+\beta_1 n)}(1-\omega)}{1-e^{-e^{\beta_0+\beta_1 n}}} + \frac{e^{-e^{\beta_0+\beta_1 m}+2(\beta_0+\beta_1 m)}\,\omega}{2\left(1-e^{-e^{\beta_0+\beta_1 m}}\right)} + \frac{e^{-e^{\beta_0+\beta_1 o}+2(\beta_0+\beta_1 o)}\,\omega}{2\left(1-e^{-e^{\beta_0+\beta_1 o}}\right)}$$

$$b = \frac{e^{-e^{\beta_0+\beta_1 n}+2(\beta_0+\beta_1 n)}\,n(1-\omega)}{1-e^{-e^{\beta_0+\beta_1 n}}} + \frac{e^{-e^{\beta_0+\beta_1 m}+2(\beta_0+\beta_1 m)}\,m\omega}{2\left(1-e^{-e^{\beta_0+\beta_1 m}}\right)} + \frac{e^{-e^{\beta_0+\beta_1 o}+2(\beta_0+\beta_1 o)}\,o\omega}{2\left(1-e^{-e^{\beta_0+\beta_1 o}}\right)}$$

$$c = \frac{e^{-e^{\beta_0+\beta_1 n}+2(\beta_0+\beta_1 n)}\,n^2(1-\omega)}{1-e^{-e^{\beta_0+\beta_1 n}}} + \frac{e^{-e^{\beta_0+\beta_1 m}+2(\beta_0+\beta_1 m)}\,m^2\omega}{2\left(1-e^{-e^{\beta_0+\beta_1 m}}\right)} + \frac{e^{-e^{\beta_0+\beta_1 o}+2(\beta_0+\beta_1 o)}\,o^2\omega}{2\left(1-e^{-e^{\beta_0+\beta_1 o}}\right)}$$

The inverse of the above Fisher-information matrix is given by

$$\boldsymbol{M}^{-1}(\xi) = \begin{bmatrix} a^* & b^* \\ b^* & c^* \end{bmatrix} \tag{14}$$

with

$$a^* = \frac{1}{t}\Bigg\{2e^{-2\beta_0}\Big(2e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)(e^{e^{\beta_0+\beta_1 o}}-1)n^2(\omega-1) + (e^{e^{\beta_0+\beta_1 n}}-1)\big(-e^{2\beta_1 m}(e^{e^{\beta_0+\beta_1 o}}-1)m^2 - e^{2\beta_1 o}(e^{e^{\beta_0+\beta_1 m}}-1)o^2\big)\omega\Big)\Bigg\}$$

$$\begin{aligned} t = \omega\Big(&-2e^{2\beta_1(m+n)}(m-n)^2(\omega-1) + 2e^{e^{\beta_0+\beta_1 o}+2\beta_1(m+n)}(m-n)^2(\omega-1) \\ &-2e^{2\beta_1(n+o)}(n-o)^2(\omega-1) + 2e^{e^{\beta_0+\beta_1 m}+2\beta_1(n+o)}(n-o)^2(\omega-1) \\ &+2e^{2\beta_1(m+o)}(m-o)^2\omega - 2e^{e^{\beta_0+\beta_1 n}+2\beta_1(m+o)}(m-o)^2\omega\Big) \end{aligned}$$

$$b^* = -\frac{1}{t}\Bigg\{2e^{-2\beta_0}\Big(2e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)(e^{e^{\beta_0+\beta_1 o}}-1)n(\omega-1) - (e^{e^{\beta_0+\beta_1 n}}-1)\big(e^{2\beta_1 m}(e^{e^{\beta_0+\beta_1 o}}-1)m + e^{2\beta_1 o}(e^{e^{\beta_0+\beta_1 m}}-1)o\big)\omega\Big)\Bigg\}$$

$$c^* = \frac{1}{t}\Bigg\{2e^{-2\beta_0}\Big(2e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)(e^{e^{\beta_0+\beta_1 o}}-1)(\omega-1) - (e^{e^{\beta_0+\beta_1 n}}-1)\big(e^{2\beta_1 o}(e^{e^{\beta_0+\beta_1 m}}-1)+e^{2\beta_1 m}(e^{e^{\beta_0+\beta_1 o}}-1)\big)\omega\Big)\Bigg\}$$

Using Equation (5), we get the function

$$\begin{aligned} H(\xi) = \frac{1}{v}\Bigg\{ & 4e^{-4\beta_0}\Big(2e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)(e^{e^{\beta_0+\beta_1 o}}-1)(\omega-1) \\ & - (e^{e^{\beta_0+\beta_1 n}}-1)\big(e^{2\beta_1 o}(e^{e^{\beta_0+\beta_1 m}}-1)+e^{2\beta_1 m}(e^{e^{\beta_0+\beta_1 o}}-1)\big)\omega\Big) \\ & \times\Big(2e^{2\beta_1 n}(e^{e^{\beta_0+\beta_1 m}}-1)(e^{e^{\beta_0+\beta_1 o}}-1)n^2(\omega-1) \\ & + (e^{e^{\beta_0+\beta_1 n}}-1)\big(-e^{2\beta_1 m}(e^{e^{\beta_0+\beta_1 o}}-1)m^2 - e^{2\beta_1 o}(e^{e^{\beta_0+\beta_1 m}}-1)o^2\big)\omega\Big)\Bigg\} \end{aligned} \tag{15}$$

with

$$\begin{aligned} v = \Bigg\{ & \omega^2\Big(2e^{2\beta_1(m+n)}(m-n)^2(\omega-1) - 2e^{e^{\beta_0+\beta_1 o}+2\beta_1(m+n)}(m-n)^2(\omega-1) \\ & + 2e^{2\beta_1(n+o)}(n-o)^2(\omega-1) - 2e^{e^{\beta_0+\beta_1 m}+2\beta_1(n+o)}(n-o)^2(\omega-1) \\ & - 2e^{2\beta_1(m+o)}(m-o)^2\omega + e^{e^{\beta_0+\beta_1 n}+2\beta_1(m+o)}(m-o)^2\omega\Big)^2\Bigg\} \end{aligned}$$

Now, the problem is to minimize the function $H(\xi)$ with respect to m, n, o, and ω for given values of β_0 and β_1. This is achieved by using the "NMinimize" function in Mathematica software and getting the optimal values m^*, n^*, o^*, and ω^*. The numerical values of m^*, n^*, o^*, and ω^* are given in Table 3 (Appendix-II). Next, by using Equation (14) we derive the quadratic form as specified in Equation (6) which is as follows:

$$E(\mathbf{x},\xi^*) = j\left\{a^* + b^* x + \left(\frac{\delta_1\,(b^*+c^* x)}{\delta_2}\right) + x\left(b^* + c^* x + \frac{\delta_1\,(a^*+b^* x)}{\delta_3}\right)\right\} \tag{16}$$

with

$$\begin{aligned} \delta_1 = & \left(2e^{2\beta_1 n}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)\left(e^{e^{\beta_0+\beta_1 o}}-1\right)n(\omega-1)\right. \\ & \left. -\left(e^{e^{\beta_0+\beta_1 n}}-1\right)\left(e^{2\beta_1 m}\left(e^{e^{\beta_0+\beta_1 o}}-1\right)m + e^{2\beta_1 o}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)o\right)\omega\right) \end{aligned}$$

$$\begin{aligned} \delta_2 = & \left(2e^{2\beta_1 n}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)\left(e^{e^{\beta_0+\beta_1 o}}-1\right)(\omega-1)\right. \\ & \left. -\left(e^{e^{\beta_0+\beta_1 n}}-1\right)\left(e^{2\beta_1 o}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)+e^{2\beta_1 m}\left(e^{e^{\beta_0+\beta_1 o}}-1\right)\right)\omega\right) \end{aligned}$$

$$\delta_3 = \left(2e^{2\beta_1 n}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)\left(e^{e^{\beta_0+\beta_1 o}}-1\right)n^2(\omega-1) + \left(e^{e^{\beta_0+\beta_1 n}}-1\right)\left(-e^{2\beta_1 m}\left(e^{e^{\beta_0+\beta_1 o}}-1\right)m^2 - e^{2\beta_1 o}\left(e^{e^{\beta_0+\beta_1 m}}-1\right)o^2\right)\omega\right)$$

Replacing the numerical values of m^*, n^*, o^*, and ω^* in Equation (16) and using the Mathematica software, we obtain

$$\sup_{\mathbf{x}\in\mathscr{S}} E(\mathbf{x},\xi^*) = 2,$$

which is nothing but the number of unknown parameters. Thus, the necessary and sufficient condition of the equivalence theorem is established. This proves Theorem 3.2.1.

4. Discussion

This study introduced R-optimal design strategies for logistic regression models with the complementary log-log link, targeting applications in reliability and accelerated life testing. The key objective of the R-optimal criterion is to minimize the volume of the parameter confidence region, thereby improving estimation precision and inferential performance in binary outcome models.

To demonstrate the practical utility of the proposed design approach, we considered a simulation study inspired by an accelerated life testing scenario. In this example, electronic components (LED Device Failure Under Voltage Stress) were tested under varying voltage stress levels, with the failure probability modeled as a function of voltage using a complementary log log link. The covariate (voltage) was defined over the interval 1.5V to 3.0V, and regression parameters were varied across $\beta_0 \in [1,5]$ and $\beta_1 \in [1,10]$ to reflect different levels of baseline risk and stress effect.

For a representative case with $\beta_0 = 2$ and $\beta_1 = 1.5$, we compared the R-optimal design with both D-optimal and uniform designs under a fixed sample size. Simulation results showed that the R-optimal design produced lower standard errors for the slope estimate, narrower 95% confidence intervals, and better predictive accuracy. These improvements were especially notable when the covariate effect was strong, a common condition in reliability experiments where stress factors sharply influence failure behavior.

Table 1. Performance Comparison of Experimental Designs

Design	SE($\hat{\beta}_1$)	95% CI	Prediction Accuracy	Coverage
Uniform	0.21	0.82	83%	94%
D-optimal	0.17	0.68	86%	95%
R-optimal	0.14	0.56	88%	95%

The findings confirm that R-optimal designs provide tangible advantages in estimating parameters efficiently and in making reliable predictions, particularly in high-gradient regions of the covariate space. This underscores their relevance for practical reliability

studies, especially in industrial testing scenarios where experimental runs are expensive and precision is critical.

Future research could explore extensions to multi-covariate models, sequential R-optimal designs, and the incorporation of prior uncertainty through Bayesian R-optimal design frameworks.

5. Conclusions

The manuscript investigates optimal experimental design strategies, specifically R-optimality, for two-parameter logistic regression (2PLR) models using the complementary log-log link function based on two- and three-support point designs. Furthermore, at the support points of the R-optimal design, the equivalence theorem (i.e., the necessary and sufficient condition of R-optimality) is established using the *Mathematica* software. This program is employed to quantitatively determine the support points of the optimal designs and the corresponding weights assigned to these points. A catalog of support points and the weights allocated to each of them in accordance with R-optimal designs is provided in Tables 2 and 3 (Appendix-II).

Acknowledgements

The author would like to sincerely thank the anonymous reviewers for their insightful comments and constructive suggestions, which greatly helped me, improve the quality of this work. The author is also grateful to Prof. (Dr.) Mahesh Kumar Panda for providing the software that was essential for carrying out this research.

References

Agresti, A., (2002). Logistic regression. *Categorical data analysis*.

Atkinson, A., Donev, A., and Tobias, R., (2007). *Optimum Experimental Designs, with SAS*. 34, Oxford University Press, Oxford.

Bailey, R. A., and Simon, L. J., (1960). Two studies in automobile insurance ratemaking. *ASTIN Bulletin: The Journal of the IAA*, 1(4), pp. 192–217.

Biswal, T. K., (2024). R-optimal designs for Poisson regression model with two parameters. *Pakistan journal of statistics*, 40(3), pp. 285–300.

Biswal, T. K., (2024). R-optimal designs for Linear regression model in two explanatory variables. *International Journal of Statistics and Reliability Engineering*, pp. 1–12.

Chaloner, K., Larntz, K., (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21(2), pp. 191–208.

Chernoff, H., (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, pp. 586–602.

De Jong, P., Heller, G. Z., (2008). *Generalized linear models for insurance data*. Cambridge Books.

Dette, H., Haines, L. M., (1994). E-optimal designs for linear and nonlinear models with two parameters. *Biometrika*, 81(4), pp. 739–754.

Dette, H., (1997). Designing experiments with respect to some 'standardized' optimality criteria. *Journal of Royal Statistical Society*, 59(B), pp. 97–110.

Dror, H. A., and Steinberg, D. M., (2006). Robust experimental design for multivariate generalized linear models. *Technometrics*, 48(4), pp. 520–529.

Ford, I., Torsney, B. and Wu, C. J., (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(2), pp. 569–583.

Fox, J., (2015). *Applied Regression Analysis and Generalized Linear Models*. Sage Publications.

Goldburd, M., Khare, A., Tevet, D. and Guller, D., (2016). Generalized linear models for insurance rating. *Casualty Actuarial Society*, Virginia.

Haines, L. M., Kabera, G. M., (2018). D-optimal designs for the two-variable binary logistic regression model with interaction. *Journal of Statistical Planning and Inference*, 193, pp. 136–150.

Kiefer, J., (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2), pp. 272–304.

Kiefer, J., Wolfowitz, J., (1959). Optimal designs in regression problems. *Annals of Mathematical Statistics*, 30, pp. 271–294.

Lee, Y., and Nelder, J. A., (2002). Analysis of ulcer data using hierarchical generalized linear models. *Statistics in Medicine*, 21(2), pp. 191–202.

Liu, P., Gao, L. L. and Zhou, J., (2022). R-optimal designs for multi-response regression models with multi-factors. *Communications in Statistics-Theory and Methods*, 51(2), pp. 340–355.

Mathew, T., Sinha, B. K., (2001). Optimal designs for binary data under logistic regression. *Journal of Statistical Planning and Inference*, 93(1-2), pp. 295–307.

McCullough, P., Nelder, J. A., (1989). *Generalized linear models*. Chapman and hall. New York.

Myers, R. H., and Montgomery, D. C., (1997). A tutorial on generalized linear models. *Journal of Quality Technology*, 29(3), pp. 274–291.

Nelder, J. A., Wedderburn, R. W., (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3), pp. 370–384.

Panda, M. K., Sahoo, R. P., (2022). R-optimal designs for linear log contrast model with mixture experiments. *Communications in Statistics-Theory and Methods*, *53*(7), pp. 2355–2368.

Panda, M. K., Biswal, T. K., (2024). R-optimal designs for logistic regression model with two variables, *Statistics and Applications*, 23(1), pp. 1–12 .

Panda, M. K., Biswal, T. K. and Gupta V. K., (2024). R-Optimal Designs for Gamma Regression Model with Two Parameters. *Statistics and Applications*, 23(1), pp. 33–53.

Russell, K. G., (2018). *Design of experiments for generalized linear models*, CRC Press.

Shiboski, S. C. and Jewell, N. P., (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association*, 87(418), pp. 360–372.

Silvey, S. D., (1980). *Optimal Design*, London: Chapman and Hall.

Sitter, R. R. and Wu, C. F. J., (1993). Optimal designs for binary response experiments: Fieller, D, and A criteria. *Scandinavian Journal of Statistics*, pp. 329–341.

Appendix-I

"Mathematica" codes to get R-optimal designs for two-parameter logistic regression (2PLR) models using the complementary log-log (c-loglog) link function: for two-point design. Steps:

1. $$\eta = \beta_0 + \beta_1 x$$

2. $$j = \frac{\exp[2\eta - \exp[\eta]]}{1 - \exp[-\exp[\eta]]}$$

3. $$A_1 = j \cdot \{\{1,x\},\{x,x^2\}\}\Big|_{x\to m}, \quad A_2 = j \cdot \{\{1,x\},\{x,x^2\}\}\Big|_{x\to n}$$

4. $$M = \omega \cdot A_1 + (1-\omega)\cdot A_2$$

5. $$M_1 = \text{FullSimplify}[M], \quad \text{MatrixForm}[M]$$

6. $$M_2 = \text{FullSimplify}[\text{Inverse}[M_1]], \quad \text{MatrixForm}[M_1]$$

7. $$H = \text{FullSimplify}[\alpha_{11}^*.\alpha_{22}^*]$$
   (Input β_0, β_1 values from Table 2)

8. $$\text{NMinimize}[\{H, 0<\omega<1\}, \{m,n,\omega\}]$$

9. $$M_3 = \left\{\left\{\frac{1}{\alpha_{11}^*}, 0\right\}, \left\{0, \frac{1}{\alpha_{22}^*}\right\}\right\}, \quad \text{MatrixForm}[M_3]$$

10. $$Z = \{\{1,x\}\}, \quad Z^T = \text{Transpose}[\{\{1,x\}\}]$$

11. $$V = j \cdot Z \cdot M_2 \cdot M_3 \cdot M_2 \cdot Z^T$$

12. $$V_1 = V \Big/ .\{m \to \text{Input}, n \to \text{Input}, \omega \to \text{Input}, x \to \text{Input}\}$$
    (Input $m, n, \omega,$ and x values from Table 2)

Note: *In a similar way one can get the R-optimal design for three support points by taking the values of m, n, o, and ω values from Table 3.

Appendix-II

Table-2 & Table-3 provides locally R-optimal designs is for vectors $\beta = (\beta_0, \beta_1)^T$ with $\beta_0 \in [1,5]$ and $\beta_1 \in [1,10]$.

Table 2. Two support points design

β	$\beta_0 = 1, \beta_1 = 1$	$\beta_0 = 1, \beta_1 = 2$	$\beta_0 = 1, \beta_1 = 3$	$\beta_0 = 1, \beta_1 = 4$
$\mathbf{x}$	$\begin{pmatrix} 0.0944 & -2.6045 \\ 0.5993 & 0.4007 \end{pmatrix}$	$\begin{pmatrix} 0.0472 & -1.3022 \\ 0.5993 & 0.4007 \end{pmatrix}$	$\begin{pmatrix} 0.0314 & -0.8681 \\ 0.5993 & 0.4007 \end{pmatrix}$	$\begin{pmatrix} 0.0236 & -0.6511 \\ 0.5993 & 0.4007 \end{pmatrix}$
ω				
β	$\beta_0 = 1, \beta_1 = 5$	$\beta_0 = 1, \beta_1 = 6$	$\beta_0 = 1, \beta_1 = 7$	$\beta_0 = 1, \beta_1 = 8$
$\mathbf{x}$	$\begin{pmatrix} 0.0188 & -0.5209 \\ 0.5993 & 0.4007 \end{pmatrix}$	$\begin{pmatrix} 0.0157 & -0.4340 \\ 0.5993 & 0.4007 \end{pmatrix}$	$\begin{pmatrix} 0.0134 & -0.3720 \\ 0.5993 & 0.4007 \end{pmatrix}$	$\begin{pmatrix} 0.0118 & -0.3255 \\ 0.5993 & 0.4007 \end{pmatrix}$
ω				
β	$\beta_0 = 1, \beta_1 = 9$	$\beta_0 = 1, \beta_1 = 10$	$\beta_0 = 2, \beta_1 = 1$	$\beta_0 = 2, \beta_1 = 2$
$\mathbf{x}$	$\begin{pmatrix} 0.0104 & -0.2833 \\ 0.5993 & 0.4007 \end{pmatrix}$	$\begin{pmatrix} 0.0094 & -0.2604 \\ 0.5993 & 0.4007 \end{pmatrix}$	$\begin{pmatrix} -0.7911 & -3.9028 \\ 0.5461 & 0.4539 \end{pmatrix}$	$\begin{pmatrix} -0.3951 & -1.9511 \\ 0.5461 & 0.4539 \end{pmatrix}$
ω				
β	$\beta_0 = 2, \beta_1 = 3$	$\beta_0 = 2, \beta_1 = 4$	$\beta_0 = 2, \beta_1 = 5$	$\beta_0 = 2, \beta_1 = 6$
$\mathbf{x}$	$\begin{pmatrix} -0.2637 & -1.3009 \\ 0.5461 & 0.4539 \end{pmatrix}$	$\begin{pmatrix} -0.1977 & -0.9747 \\ 0.5461 & 0.4539 \end{pmatrix}$	$\begin{pmatrix} -0.1582 & -0.7805 \\ 0.5461 & 0.4539 \end{pmatrix}$	$\begin{pmatrix} -0.1318 & -0.6504 \\ 0.5461 & 0.4539 \end{pmatrix}$
ω				
β	$\beta_0 = 2, \beta_1 = 7$	$\beta_0 = 2, \beta_1 = 8$	$\beta_0 = 2, \beta_1 = 9$	$\beta_0 = 2, \beta_1 = 10$
$\mathbf{x}$	$\begin{pmatrix} -0.1130 & -0.5575 \\ 0.5461 & 0.4539 \end{pmatrix}$	$\begin{pmatrix} -0.0988 & -0.4878 \\ 0.5461 & 0.4539 \end{pmatrix}$	$\begin{pmatrix} -0.0879 & -0.4336 \\ 0.5461 & 0.4539 \end{pmatrix}$	$\begin{pmatrix} -0.0791 & -0.3902 \\ 0.5461 & 0.4539 \end{pmatrix}$
ω				
β	$\beta_0 = 3, \beta_1 = 1$	$\beta_0 = 3, \beta_1 = 2$	$\beta_0 = 3, \beta_1 = 3$	$\beta_0 = 3, \beta_1 = 4$
$\mathbf{x}$	$\begin{pmatrix} -1.7588 & -4.9906 \\ 0.4919 & 0.5081 \end{pmatrix}$	$\begin{pmatrix} -0.8799 & -2.4953 \\ 0.4919 & 0.5081 \end{pmatrix}$	$\begin{pmatrix} -0.5866 & -1.6635 \\ 0.4919 & 0.5081 \end{pmatrix}$	$\begin{pmatrix} -0.4391 & -1.2476 \\ 0.4919 & 0.5081 \end{pmatrix}$
ω				
β	$\beta_0 = 3, \beta_1 = 5$	$\beta_0 = 3, \beta_1 = 6$	$\beta_0 = 3, \beta_1 = 7$	$\beta_0 = 3, \beta_1 = 8$
$\mathbf{x}$	$\begin{pmatrix} -0.3519 & -0.9981 \\ 0.4919 & 0.5081 \end{pmatrix}$	$\begin{pmatrix} -0.2933 & -0.8317 \\ 0.4919 & 0.5081 \end{pmatrix}$	$\begin{pmatrix} -0.7129 & -0.2514 \\ 0.4920 & 0.5080 \end{pmatrix}$	$\begin{pmatrix} -0.2199 & -0.6238 \\ 0.4919 & 0.5081 \end{pmatrix}$
ω				
β	$\beta_0 = 3, \beta_1 = 9$	$\beta_0 = 3, \beta_1 = 10$	$\beta_0 = 4, \beta_1 = 1$	$\beta_0 = 4, \beta_1 = 2$
$\mathbf{x}$	$\begin{pmatrix} -0.1955 & -0.5545 \\ 0.4919 & 0.5081 \end{pmatrix}$	$\begin{pmatrix} -0.1759 & -0.4990 \\ 0.4919 & 0.5081 \end{pmatrix}$	$\begin{pmatrix} -2.7481 & -6.0244 \\ 0.4063 & 0.5937 \end{pmatrix}$	$\begin{pmatrix} -1.3740 & -3.0122 \\ 0.4063 & 0.5937 \end{pmatrix}$
ω				
β	$\beta_0 = 4, \beta_1 = 3$	$\beta_0 = 4, \beta_1 = 4$	$\beta_0 = 4, \beta_1 = 5$	$\beta_0 = 4, \beta_1 = 6$
$\mathbf{x}$	$\begin{pmatrix} -0.9160 & -2.0081 \\ 0.4603 & 0.5397 \end{pmatrix}$	$\begin{pmatrix} -0.6870 & -1.5061 \\ 0.4603 & 0.5397 \end{pmatrix}$	$\begin{pmatrix} -0.5496 & -1.2048 \\ 0.4603 & 0.5397 \end{pmatrix}$	$\begin{pmatrix} -0.4580 & -1.0004 \\ 0.4603 & 0.5397 \end{pmatrix}$
ω				
β	$\beta_0 = 4, \beta_1 = 7$	$\beta_0 = 4, \beta_1 = 8$	$\beta_0 = 4, \beta_1 = 9$	$\beta_0 = 4, \beta_1 = 10$
$\mathbf{x}$	$\begin{pmatrix} -0.3925 & -0.8606 \\ 0.4603 & 0.5397 \end{pmatrix}$	$\begin{pmatrix} -0.3435 & -0.7530 \\ 0.4603 & 0.5397 \end{pmatrix}$	$\begin{pmatrix} -0.6893 & -0.3053 \\ 0.4064 & 0.5936 \end{pmatrix}$	$\begin{pmatrix} -0.6024 & -0.2748 \\ 0.4063 & 0.5397 \end{pmatrix}$
ω				
β	$\beta_0 = 5, \beta_1 = 1$	$\beta_0 = 5, \beta_1 = 2$	$\beta_0 = 5, \beta_1 = 3$	$\beta_0 = 5, \beta_1 = 4$
$\mathbf{x}$	$\begin{pmatrix} -3.7424 & -7.0409 \\ 0.4404 & 0.5596 \end{pmatrix}$	$\begin{pmatrix} -1.8712 & -3.5204 \\ 0.4404 & 0.5596 \end{pmatrix}$	$\begin{pmatrix} -1.2474 & -2.3469 \\ 0.4404 & 0.5596 \end{pmatrix}$	$\begin{pmatrix} -0.9355 & -1.7606 \\ 0.4404 & 0.5596 \end{pmatrix}$
ω				
β	$\beta_0 = 5, \beta_1 = 5$	$\beta_0 = 5, \beta_1 = 6$	$\beta_0 = 5, \beta_1 = 7$	$\beta_0 = 5, \beta_1 = 8$
$\mathbf{x}$	$\begin{pmatrix} -0.7484 & -1.4081 \\ 0.4404 & 0.5596 \end{pmatrix}$	$\begin{pmatrix} -0.6237 & -1.1734 \\ 0.4404 & 0.5596 \end{pmatrix}$	$\begin{pmatrix} -1.0058 & -0.5346 \\ 0.4404 & 0.5596 \end{pmatrix}$	$\begin{pmatrix} -0.8801 & -0.4678 \\ 0.4404 & 0.5596 \end{pmatrix}$
ω				
β	$\beta_0 = 5, \beta_1 = 9$	$\beta_0 = 5, \beta_1 = 10$	-	-
$\mathbf{x}$	$\begin{pmatrix} -0.4158 & -0.7823 \\ 0.4404 & 0.5596 \end{pmatrix}$	$\begin{pmatrix} -0.3742 & -0.7040 \\ 0.4404 & 0.5596 \end{pmatrix}$	-	-
ω				

Table 3. Three support points design

β	$\beta_0 = 1, \beta_1 = 1$	$\beta_0 = 1, \beta_1 = 2$	$\beta_0 = 1, \beta_1 = 3$
$\mathbf{x}$	$\begin{pmatrix} 0.0944 & 0.0944 & -2.6045 \\ 0.4006 & 0.1988 & 0.4006 \end{pmatrix}$	$\begin{pmatrix} 0.0472 & -1.3022 & 0.0472 \\ 0.2996 & 0.4007 & 0.2996 \end{pmatrix}$	$\begin{pmatrix} 0.0314 & -0.8681 & 0.0314 \\ 0.2996 & 0.4007 & 0.2996 \end{pmatrix}$
ω			
β	$\beta_0 = 1, \beta_1 = 4$	$\beta_0 = 1, \beta_1 = 5$	$\beta_0 = 1, \beta_1 = 6$
$\mathbf{x}$	$\begin{pmatrix} 0.0236 & -0.6511 & 0.0236 \\ 0.2996 & 0.4007 & 0.2996 \end{pmatrix}$	$\begin{pmatrix} 0.0188 & 0.0188 & -0.5209 \\ 0.4006 & 0.1988 & 0.4006 \end{pmatrix}$	$\begin{pmatrix} 0.0157 & 0.0157 & -0.4340 \\ 0.4006 & 0.1988 & 0.4006 \end{pmatrix}$
ω			
β	$\beta_0 = 2, \beta_1 = 1$	$\beta_0 = 3, \beta_1 = 3$	$\beta_0 = 3, \beta_1 = 5$
$\mathbf{x}$	$\begin{pmatrix} -0.7911 & -3.9028 & -0.7911 \\ 0.2730 & 0.4539 & 0.2730 \end{pmatrix}$	$\begin{pmatrix} -0.5866 & -1.6635 & -0.5866 \\ 0.2459 & 0.5081 & 0.2459 \end{pmatrix}$	$\begin{pmatrix} -0.9981 & -0.9981 & -1.3534 \\ 0.4919 & 0.0161 & 0.4919 \end{pmatrix}$
ω			
β	$\beta_0 = 3, \beta_1 = 6$	$\beta_0 = 3, \beta_1 = 7$	$\beta_0 = 3, \beta_1 = 8$
$\mathbf{x}$	$\begin{pmatrix} -0.8317 & -0.8317 & -0.2933 \\ 0.4919 & 0.0161 & 0.4919 \end{pmatrix}$	$\begin{pmatrix} -0.7129 & -0.7129 & -0.2514 \\ 0.4919 & 0.0161 & 0.4919 \end{pmatrix}$	$\begin{pmatrix} -0.6238 & -0.6238 & -0.2199 \\ 0.4919 & 0.0161 & 0.4919 \end{pmatrix}$
ω			
β	$\beta_0 = 3, \beta_1 = 9$	$\beta_0 = 4, \beta_1 = 1$	$\beta_0 = 4, \beta_1 = 2$
$\mathbf{x}$	$\begin{pmatrix} -0.1955 & -0.5545 & -0.5545 \\ 0.4919 & 0.0161 & 0.4919 \end{pmatrix}$	$\begin{pmatrix} -6.0244 & -6.0244 & -2.7481 \\ 0.4603 & 0.0794 & 0.4603 \end{pmatrix}$	$\begin{pmatrix} -3.0122 & -3.0122 & -1.3740 \\ 0.4603 & 0.0794 & 0.4603 \end{pmatrix}$
ω			
β	$\beta_0 = 4, \beta_1 = 4$	$\beta_0 = 4, \beta_1 = 5$	$\beta_0 = 4, \beta_1 = 7$
$\mathbf{x}$	$\begin{pmatrix} -0.6870 & -1.5061 & -0.6870 \\ 0.2301 & 0.5397 & 0.2301 \end{pmatrix}$	$\begin{pmatrix} -1.2048 & -1.2048 & -0.5496 \\ 0.4603 & 0.0794 & 0.4603 \end{pmatrix}$	$\begin{pmatrix} -0.3925 & -0.8606 & -0.3925 \\ 0.2301 & 0.5397 & 0.2301 \end{pmatrix}$
ω			
β	$\beta_0 = 4, \beta_1 = 9$	$\beta_0 = 5, \beta_1 = 4$	$\beta_0 = 5, \beta_1 = 5$
$\mathbf{x}$	$\begin{pmatrix} -0.6693 & -0.6693 & -0.3053 \\ 0.4603 & 0.0794 & 0.4603 \end{pmatrix}$	$\begin{pmatrix} -1.7602 & -0.9355 & -1.7602 \\ 0.4404 & 0.1191 & 0.4404 \end{pmatrix}$	$\begin{pmatrix} -1.4081 & -1.4081 & -0.7484 \\ 0.4404 & 0.1191 & 0.4404 \end{pmatrix}$
ω			
β	$\beta_0 = 5, \beta_1 = 6$	$\beta_0 = 5, \beta_1 = 7$	$\beta_0 = 5, \beta_1 = 9$
$\mathbf{x}$	$\begin{pmatrix} -1.1734 & -1.1734 & -0.6237 \\ 0.4404 & 0.1191 & 0.4404 \end{pmatrix}$	$\begin{pmatrix} -1.0058 & -1.0058 & -0.5346 \\ 0.4404 & 0.1191 & 0.4404 \end{pmatrix}$	$\begin{pmatrix} -0.7823 & -0.7823 & -0.4158 \\ 0.4404 & 0.1191 & 0.4404 \end{pmatrix}$
ω			

STATISTICS IN TRANSITION new series, June 2026
Vol. 27, No. 2, pp. 219–224

About the Authors

Abu Awwad Raed R. has been an Associate Professor in the Department of Mathematics, University of Petra, Jordan, since 2013. His research interests include mathematical statistics, Bayesian analytics, progressive censoring, and reliability analysis. He has published his research in international journals.

Abufoudeh Ghassan K. has been an Associate Professor at the University of Petra since 2017. He received his PhD from Jordan University in 2013. His research areas include mathematical statistics, probability theory, and entropy measures.

Aleshinloye Yusuf Abass is a Lecturer and Researcher in the Department of Computer Science and Software Engineering, Kampala International University, Kampala, Uganda. He holds a PhD from the Nile University of Nigeria and is a member of the Computer Professionals Registration Council of Nigeria. His main research interests include artificial intelligence, machine learning, deep learning, data science, computer vision, educational technology and applied computational modelling. He has contributed to research in AI-driven analytics, intelligent systems, medical imaging, and digital transformation in education. He is also involved in academic collaborations within and outside Nigeria, including research and knowledge-exchange activities with international institutions. His professional work focuses on advancing practical AI applications, interdisciplinary research, and technology-driven innovation in higher education.

Almheidat Maaleee works at the Department of Mathematics, University of Jordan. Her research interests include statistical inference, Bayesian analysis, and divergence measures.

Alokaily Samer is an Associate Professor at the University of Petra. He received his PhD from the Michigan Technological University in 2017. He received Outstanding Graduate Student Awards. His research focuses on computational fluid dynamics.

Bahi Yassine is a PhD candidate in the Department of Computer Science at the Faculté des Sciences de Rabat and a Data Science Engineer. He has graduated from the Higher School of Information Sciences (ESI). His research and professional activities focus on statistics, data science, cloud computing, and advanced data analytics technologies, with contributions to several academic and applied projects in these fields.

Baraka Achraf Chakir holds a PhD in Statistics and Data Science from the Faculty of Sciences of Kenitra and a State Engineering degree in Computer Science from the

National Institute of Statistics and Applied Economics. He is currently a Professor in the Department of Computer Science at the Faculté des Sciences de Rabat. His teaching and research activities focus on statistics, data science, business intelligence, and artificial intelligence. He has supervised and co-supervised numerous undergraduate, master's, and doctoral research projects. He is the author and co-author of several scientific publications in internationally recognized journals and conference proceedings.

Baraka Kaoutar is a PhD candidate in the Department of Computer Science at the Faculté des Sciences de Rabat and a Data Science Engineer. She graduated from the Higher School of Information Sciences (ESI). Her research focuses on statistics, data science, and advanced data analysis methods, with contributions to several academic and applied research projects in these fields.

Biswal Tofan Kumar is affiliated with the Department of Statistics, Central University of Odisha, Sunabeda, India. His main research interests include optimal experimental design, Bayesian optimal design, generalized linear models, and statistical inference. His current research focuses on locally optimal and Bayesian optimal designs for generalized linear models under different link functions. He has contributed research articles in the area of design of experiments and statistical modelling. He is actively engaged in academic research and publication activities in statistics and related interdisciplinary applications.

Boratyńska Agata is an Associate Professor at the Institute of Econometrics, Collegium of Economic Analysis, Warsaw School of Economics SGH. She has a PhD in mathematical sciences and habilitation degree in economics. Her main areas of interest include mathematical statistics, Bayesian statistical models, robustness of statistical models with respect to a prior distribution, application of Bayesian statistical models in insurance, prediction of reserves and credibility theory.

Gaire Arjun Kumar is a Senior Lecturer at the Department of Science and Humanities, Khwopa Engineering College, Purbanchal University. He is currently a PhD scholar at the Central Department of Population Studies, Tribhuvan University. He has a strong academic background in demography and statistics, with research expertise in statistical modeling, probability distributions, and applied data analysis. He has published more than a dozen research papers and contributed to academic textbooks in statistics. His main research interests include demographic analysis, reproductive health modeling, and applications of statistical machine learning. Currently, he is engaged in research on probabilistic modeling of reproductive life-course patterns among Nepalese women and the development of data-driven computational tools for applied statistics. He also serves as a member of the editorial board of the Journal of Science and Engineering (JScE), Nepal, published by Khwopa Engineering College.

Khalifi Hamid, who holds a PhD in Engineering, has a PhD in Data Science and Artificial Intelligence from the Faculty of Science in Meknes, a State Engineering Degree in Computer Science from the National Institute of Statistics and Applied Economics in Rabat, a Bachelor's degree in Computer Science and a University Diploma in Software Engineering from Moulay Ismail University in Meknes. Currently, he is a Professor in the Department of Computer Science and the coordinator of the Master's program in Information Technology, Governance and Digital Transformation at the Faculty of Science, Mohammed V University, Rabat. He has supervised and co-supervised several final-year projects, dissertations and theses at Bachelor's, Master's and PhD levels. He is the author and co-author of numerous publications in internationally renowned journals and conference proceedings.

Kokczyński Bernard, PhD, is an Assistant Professor at the Department of Corporate Finance Management, University of Lodz. His research interests focus on bankruptcy prediction, credit risk assessment, financial analysis, and quantitative methods in economics and finance. His professional experience combines academic research with practical work in the financial and energy sectors, including credit and market risk analysis.

Kot Stanisław Maciej is a Professor Emeritus at the Department of Statistics and Econometrics, Gdansk University of Technology. His research area concerns the distributional analysis with special attention to income inequality, poverty, and equivalence scales. Recently, his scientific activity has concentrated on measuring unobservable normative parameters, such as inequality aversion, which determine social welfare functions.

Krajčíková Lívia is a PhD. student at the Department of Statistics, Faculty of Economic Informatics, Bratislava University of Economics and Business. Her main areas of interest are poverty, social exclusion and severe material and social deprivation.

Landmesser-Rusek Joanna is an Associate Professor and Head of the Department of Econometrics and Statistics at WULS-SGGW, Warsaw. She earned her PhD from Bundeswehr University Munich and habilitation from Nicolaus Copernicus University. Her academic research focuses on microeconometric modeling, financial market network analysis, and survival analysis. She has authored 97 peer-reviewed publications throughout her career. Additionally, she serves as a member of the Council of the Data Classification and Analysis Section of the Polish Statistical Association.

Nkpordee Lekia is a Lecturer and Researcher in the Department of Mathematics and Statistics at the Kampala International University. His research interests include time series analysis, econometrics, statistical computing, machine learning applications, and computational statistics. He has experience in AI aided statistical modelling, multivariate analysis, and applied data science. Dr. Nkpordee has supervised undergraduate and

postgraduate research projects and has published in areas related to statistical modelling, forecasting, and predictive analytics. He is actively involved in interdisciplinary research applying statistical and machine learning techniques to environmental, economic, and health related problems.

Olugbenga Ejidokun Adekunle is a Researcher in the Department of Computer Science at Kampala International University. His research focuses on machine learning, artificial intelligence, data science, computational modelling, and intelligent systems development. He has strong interest in deep learning applications, predictive analytics, and hybrid modelling techniques for solving real world problems. His academic work integrates computer science methods with statistical learning approaches for environmental and socio-economic data analysis. He is also involved in interdisciplinary research aimed at advancing data-driven decision making and innovative computing solutions.

Osayomore Ikpotokin is a Researcher in the Department of Mathematics and Statistics at the Kampala International University. His areas of interest include applied statistics, statistical modelling, time series forecasting, and quantitative data analysis. He has participated in research involving predictive modelling, environmental statistics, and computational data analysis. His work focuses on the application of modern statistical techniques to real world datasets for evidence-based decision making. He is particularly interested in the integration of classical statistical methods with emerging analytical approaches for improving forecasting accuracy and model performance.

Pinsky Eugene is an Associate Professor in the Department of Computer Science at Boston University Metropolitan College. His research interests are the design and analysis of computationally simple and explainable prediction and classification methods in machine learning and statistics, with a particular focus on mean absolute deviations and computational finance.

Prodhani Hosenur Rahman has completed his PhD in Distribution Theory from Assam University, Silchar, Assam, India under the supervision of Professor Rama Shanker and presently working as an Assistant Professor in the Department of Statistics, B.N. College (Autonomous), Dhubri, Assam, India. His research interest includes distribution theory, reliability theory and Bayesian inference. He has published more than 20 research papers in national and international journals of statistics.

Rahmaoui Mehdi holds a PhD in Biology, Nutrition and Health. His research interests include the analysis and interpretation of biological data, with a particular focus on nutrition, health sciences, and biomedical research. His work primarily involves biological data analysis and its application to support scientific research and improve health-related knowledge.

Shanker Rama is a Professor of Statistics in the Department of Statistics, Assam University, Silchar, Assam, India. His research interests are distribution theory, reliability theory, statistical inference, modeling lifetime data, and mathematical programming. Professor Shanker has published 235 research papers in international/national journals and conferences. He has successfully guided four PhD scholars. He has also published a book entitled "Introduction to Calculus for Business and Economics". Professor Shanker is an active member of many scientific professional bodies in India and worldwide.

Vojtková Mária is an associate professor and head of the Department of Statistics at the Faculty of Economic Informatics of the Bratislava University of Economics and Business. Simultaneously she holds the position of Vice-President of the Slovak Statistical and Demographic Society for Academic Statistics. In her research, she focuses on the application of quantitative methods in the analysis of socio-economic phenomena, focusing mainly on multicriteria evaluation, dimensionality reduction methods, classification and modeling of dependencies. She is actively involved in solving domestic and international research projects.

Wang Zibo has graduated from Boston University Metropolitan College. His research interests are primarily in statistics. He currently works in asset management at China Rida Investment Development Group Co., Ltd. He has long focused on data analysis and asset management and has completed several papers in statistics, including "On the Use of Mean Absolute Deviations in Beta-PERT", "Mean Absolute Deviation for the Weibull Distribution with Applications in Survival Analysis", and "Comparative Analysis of Temperature Prediction Models: Simple Models vs. Deep Learning Models".

Witkowska Dorota is a Full Professor of Economics and Finance at the Department of Corporate Finance Management, University of Lodz. Her research interests are multivariate statistical analysis, modeling of socio-economic phenomena and financial markets. She is the author or co-author of 13 monographs, 17 textbooks, over 265 reviewed articles and chapters (in Polish and international journals also listed on the ISI/WoS/Scopus), over 200 papers at national and international conferences in Poland and abroad. She serves on the board of editors of eight international scientific journals, is a member of four international scientific societies, and a member of the scientific and program committees of many scientific international conferences.

Yamoul Nada, an architect and holder of a PhD in energy efficiency, is a graduate of the Faculty of Science at Ibn Tofail University in Kenitra. Her doctoral thesis has been completed under a co-supervision agreement between the Doctoral Studies Centre at Ibn Tofail University in Kenitra and the Doctoral Studies Center at the National School of Architecture in Rabat. She has contributed to several scientific publications in the

fields of building physics, energy, sustainable architecture, the thermal performance of buildings and building materials.

Zhang Weiqi has graduated from Boston University and currently serves as a Developer and Financial Analyst at Ecesis Investments LLC. Her main areas of research include stock market analysis, power company operations, energy development, income distribution, and statistical theory. She has authored and co-authored several academic papers. Her publications include, as the first author, "Pareto Distribution of the Forbes Billionaires", published in Computational Economics (SCIE), and a study on the mean absolute deviation of hyperexponential and hypoexponential distributions.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* https://sit.stat.gov.pl/ForAuthors.

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **Bold**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper.
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, (**1.1.**, **1.2.** …), **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see https://anglia.libguides.com/harvardctr. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).