Volume 26, Number 1, March 2025 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

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

CONTENTS

STATISTICS IN TRANSITION new series, *March 2025*
Vol. 26, No. 1, pp. III

Submission information for Authors

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

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

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

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

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

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

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

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

STATISTICS IN TRANSITION new series, *March 2025*
Vol. 26, No. 1, pp. V–VI

Policy Statement

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

***

Abstracting and Indexing Databases

Statistics in Transition new series is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

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

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Electronic Journals Library

Elsevier – Scopus

Genamics JournalSeek

Google Scholar

Index Copernicus

J-Gate

JournalGuide

JournalTOCs

Keepers Registry

MIAR

Microsoft Academic

OpenAIRE

ProQuest – Summon

Publons

QOAM (Quality Open Access Market)

ReadCube

RePec

SCImago Journal & Country Rank

TDNet

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

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. VII–X

From the Editor

The March issue of *Statistics in Transition new series*, presented to our readers hereby, opens this year's series of our quarterly. It contains a set of twelve articles by twenty-eight authors from nine countries (in order of appearance): Poland, USA, Sri Lanka, India, Botswana, Pakistan, Malaysia, Nigeria, and France. The wide spectrum of issues discussed confirms that our journal consistently strives to cover a wide geographical scope while remaining open to a variety of statistical problems that are of interest to experts from different fields of research, world-wide.

It is with great satisfaction that I note that this issue opens with an article by a distinguished member of our Editorial Board, Professor Janusz Wywiał.

Invited papers

Janusz L. Wywiał's paper, ***Generalised spatial autocorrelation coefficients***, focuses on properties of coefficients of spatial correlation generalised to the multidimensional case. The main result of the work is the decomposition of the introduced generalised autocorrelation coefficients into the sum of ordinary autocorrelation coefficients, but calculated on the basis of the principal components of the originally observed multidimensional variable. The development is illustrated with an empirical example. The coefficients provide a more detailed and useful description of the spatial relationships of a set of variables characterizing a population.

Research papers

In the article entitled ***An expectation-maximization algorithm for logistic regression based on individual-level predictors and aggregate-level response*, Zheng Xu** proposes an Expectation-Maximization (EM) algorithm to avoid the direct maximization of the complicated likelihood function. Simulation studies have been conducted to evaluate the performance of the EM estimator compared to different estimators proposed in the literature. Two real data-based studies have been conducted to illustrate the use of the different estimators. The EM estimator proves efficient for the logistic regression problem with an aggregate-level response and individual-level predictors.

The next paper *by* **D. Dilshanie Deepawansa** and **Priyanga Dunusinghe *Selection criteria and targeting the poor for poverty reduction: the case of social safety nets in Sri Lanka*** discusses a multidimensional selection criterion for the leading social safety net for Sri Lanka, Multidimensional Deprivation Score Test (MDST). The method

used has been applied to the HIES-2019 data. It showed that exclusion error is less than existing selection criteria when compared with different targeted groups. According to the selection cut-off, Samurdhi/welfare beneficiaries can be identified. In addition, in order to impact poverty, the transfer schemes should be varied concerning the severity of poverty. Otherwise, if all the beneficiaries get same amount of money, the impact on poverty is unlikely to change significantly. In addition to identifying the suitable beneficiaries, MDST helps to compute the contribution of deprivation in every dimension, which is taken into consideration by household or family, community, or geographical level.

Hemani Sharma's and **Parmil Kumar's** article ***On survival estimation of Lomax distribution under adaptive progressive type-II censoring*** compares the maximum likelihood (ML) estimation and the Bayesian approach for parameter estimation of the Lomax distribution. Additionally, the study aims to determine the approximate intervals for the parameters and the survival function based on adaptive progressive type-II censored data. The ML estimators of the probability distribution parameters were calculated using the Newton-Raphson method, while the delta method was used to compute the approximate confidence intervals for the survival function. The Bayesian approach was also used to estimate the unknown parameters and survival function. This was achieved through the construction of Bayesian estimators under an informative and non-informative prior based on the squared error loss function (SELF) and approximate credible intervals. The Markov Chain Monte Carlo (MCMC) method was employed to test the efficiency of the proposed method in various situations based on different criteria such as mean-squared error, bias, coverage probability, and expected length-estimated criteria.

In the paper entitled ***A fuzzy hybrid MCDM approach to the evaluation of subjective household poverty,*** **Aleksandra Łuczak** and **Sławomir Kalinowski** propose a comprehensive procedure for constructing a synthetic measure of subjective poverty. This involves aggregating factors describing the present, future, and past, which makes it easier to grasp the feeling of deprivation over time. Methods such as fuzzy TOPSIS and fuzzy hierarchical analysis (FHA) based on the fuzzy sets theory were used for this purpose. This innovative procedure was applied to assess the level of subjective household poverty in Poland based on data from survey research carried out in three stages in 2020 using the CAWI method. The results show that the assessment of household's current level of living conditions is also influenced by past events as well as projections of future developments. Changes in the values of the synthetic index illustrate the trajectory of switching from panic to negation, and attempting to cope with the situation or, alternatively, switching to a state of irritation.

The next paper, ***Type I heavy-tailed family of generalized Burr III distributions: properties, actuarial measures, regression and applications,*** by **Wilbert Nkomo**, **Broderick Oluyede**, and **Fastel Chipepa,** presents a new family of distributions (FoD)

called type I heavy-tailed odd Burr III-G (TI-HT-OBIII-G) distribution. Several statistical properties of the family are derived along with actuarial risk measures. The maximum likelihood estimation (MLE) approach is adopted in the parameter estimation process. The estimates are evaluated centered on mean square errors and average bias via the Monte Carlo simulation framework. A regression model is formulated and the residual analysis is investigated. Members of the new FoD are applied to heavy-tailed data sets and compared to some well-known competing heavy-tailed distributions. The practicality, flexibility and importance of the new distribution in modeling are empirically proven using three data sets.

Laba Handique, **Farrukh Jamal**, and **Subrata Chakraborty** in their article ***On a family that unifies the generalized Marshall-Olkin and Poisson-G family of distributions*** propose a unification of the generalized Marshall-Olkin (GMO) and Poisson-G (P-G) distributions into a new family of distributions. The density and survival function are expressed as infinite mixtures of an exponentiated-P-G family. The quantile function, asymptotes, shapes, stochastic ordering and Rényi entropy are derived. The paper presents a maximum likelihood estimation with large sample properties. A Monte Carlo simulation is used to examine the pattern of the bias and the mean square error of the maximum likelihood estimators. The utility of the proposed family is illustrated through its comparison with some important models and sub models of the family in terms of modeling real data.

Iwona Skrodzka's paper, ***Impact of human capital on the innovation performance of EU economies*** attempts to empirically determine the impact of human capital on the innovation performance of EU economies, given a gap in the literature regarding this issue. There are difficulties associated with the measurement as well as the limited number of methods to study the relationships between unobservable variables. In order to fills this gap, the partial least squares structural equation modelling (PLS-SEM) was used, covering the years 2014-2020.

The next article, ***Improving detectability of the indicator saturation approach through winsorization: an empirical study in the cryptocurrency market,*** by **Suleiman Dahir Mohamed**, **Mohd Tahir Ismail**, and **Majid Khan Bin Majahar Ali,** presents a hybrid approach called the Win-IS strategy, focusing on the influence of extreme outliers in the tail and subsequently identify breaks, trend breaks and outliers in crypto-currencies. The study uses cryptocurrencies like Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), Tether (USDT), and Ripple (XRP). This article improves the detectability of the IS approach by combining it with the winsorization strategy and hence proposes a technique known as Win-IS. The performance of Win-IS is then empirically compared to IS in five cryptocurrency markets. The Win-IS strategy outperformed the IS technique, as demonstrated by BIC scores. Furthermore, the Win-IS technique reduced severe outliers in four coins while revealing new outliers, breaks, and trend breaks, some of which were duplicated from the IS results. The repeated

outliers, breaks, and trend breaks show their importance in this market because they remained constant in both winsored and original returns.

M. Dileepkumar's, R. Anand's, and **P. G Sankaran's** paper, ***Reliability properties and applications of proportional reversed hazards in reversed relevation transform,*** describes important reliability properties of the reversed relevation transform under the proportional reversed hazards assumption. The results of research on information measures are presented. Various ageing concepts and stochastic orders are discussed. A new flexible generalization of the Fréchet distribution is introduced using the proposed transformation, and reliability properties and applications are discussed. The ageing and stochastic ordering properties of the model were derived.

In the paper ***Analytical modelling for COVID-19 data (fatality): a case study of Nigeria for the period of February 2020 – April 2022*****, E. Torsen**, **U. M. Modibbo**, **M. Mijinyawa**, **L. L. Seknewna,** and **I. Ali** used univariate time series models to analyze the confirmed cases of COVID-19 fatalities (count data and having zero inflation) due to COVID-19 in Nigeria. Specifically, the Autoregressive Integrated Moving Average (ARIMA), Zero-Inflated Poison Autoregressive (ZIPAR), and Zero-Inflated Negative Binomial Autoregressive (ZINBAR) models were employed. The findings indicate that ZINBAR having the lowest Root Mean Square Error (RMSE), the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) outperforms the other two models: hence, the ZINBAR performs better than the ZIPAR and the ARIMA. This demonstrates and emphasized the fact that for count time series data, count time series models should be used, with indication to the ZINBAR to be used to predict and forecast COVID-19 in Nigeria.

Research Communicates and Letters

Stanisław Jaworski's article ***Optimal sample size in a triangular model for sensitive questions*** considers the nonrandomized response model (proposed by Tian et al., 2007) and introduces a novel CI for the fraction of sensitive questions in the triangular model. Unlike the widely used asymptotic CI, the new approach maintains the prescribed confidence level. The minimum sample size satisfying two criteria was considered: average length and almost sure length. To obtain such sample sizes, the restrictions on privacy protection were imposed, specifically the probability of discovering a YES answer to the sensitive question. This probability should be sufficiently small to ensure the interviewee's comfort in answering the questionnaire.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 1–8, https://doi.org/10.59139/stattrans-2025-001
Received – 10.07.2023; accepted – 25.11.2024

Generalised spatial autocorrelation coefficients

Janusz L. Wywiał[1]

Abstract

The article focuses on properties generalised to the multidimensional case of known coefficients of spatial correlation. The main result of the work is the decomposition of the introduced generalised autocorrelation coefficients into the sum of ordinary autocorrelation coefficients, but calculated on the basis of the principal components of the originally observed multidimensional variable. The development is illustrated with an empirical example. The coefficients provide a more detailed description of the spatial relationships of a set of variables defined in a population.

Key words: Moran coefficient, Geary coefficient, spatial autocorrelation, Mahalanobis distance, principal components.

1. Introduction

Exploration of various phenomena in natural, social, economic and other populations requires an approach involving the analysis of relationships among observations of many features defined in these populations. This applies to populations in which a distance between pairs of its members is defined. This can lead to the division of the population into a set of homogeneous subpopulations. This was the inspiration for the preparation of this work. The known Moran (1950) and Geary (1954) spatial autocorrelation coefficients described below allow for the analysis of the spatial similarity in terms of single variables. The properties of autocorrelation coefficients were considered by, among others, Getis and Ord (1992), Griffith and Chun (2022). Recently, in the works of Krzyśko et al. (2023), Krzyśko et al. (2024), the autocorrelation coefficients were significantly generalised to the multivariate case. These generalizations use advanced functional analysis to simultaneously analyze the spatio-temporal autocorrelation of time-varying vector observations. Du (2012) generalised the Geary coefficient to a random vector. It could also be adapted to the Moran coefficient.

Let $x_i, i = 1, ..., N$ be observations of x variable. Moran (1950) defined the coefficient of spatial autocorrelation in the following way:

$$I^M = \frac{1}{wv} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - \bar{x})(x_j - \bar{x}) w_{ij} = \frac{1}{v} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - \bar{x})(x_j - \bar{x}) q_{ij} \tag{1}$$

[1]Deaprtment of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Katowice, Poland. E-mail: wywial@ue.katowice.pl. ORCID:https://orcid.org/0000-0002-3392-1688.

where $w_{ij} \geq 0$, $w_{ii} = 0$, $w = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$, $v = \sum_{i=1}^{N} (x_i - \bar{x})^2/N$, $\bar{x} = \sum_{i=1}^{N} x_i/N$, $q_{ij} = w_{ij}/w$, $0 \leq q_{ij} \leq 1$. When neighbors are more similar (more different) than observations in general, then Moran's coefficient coefficient takes positive (negative) values. Values of this coefficient close to zero indicates absence of spatial similarity. Usually, $-1 \leq I^M \leq 1$, see Cliff and Ord (1981) or Overmars et al. (2003). However, this is not always the case. The range of coefficient variability may take into account such distribution features of the examined variable, as kurtosis or skewness.

Geary (1954) proposed the following coefficient:

$$I^G = \frac{N-1}{2Nvw} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2 w_{ij} = \frac{N-1}{2Nv} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)^2 q_{ij} \geq 0. \tag{2}$$

The value of the Geary coefficient greater (smaller) than one means large differences (similarity) of neighboring objects. The value of this coefficient close to one means the lack of substantial spatial autocorrelation in the sense described above.

Weight w_{ij}, $i \neq j = 1,...,N$ can be defined in several ways. For instance, the weights may indicate the economic relationship between sub-areas. They may, for example, indicate cooperative connections between economic regions, characterised by observations of a multidimensional variable. In particular, these connections may be financial flows between these companies. In this case, e.g. well-known input-output matrix of Leontief (1986) could be used to construct the weights. Getis and Ord (1992) suggested to set that $w_{ij} = 1$, when $|x_i - x_j| \geq d_0$ and $w_{ij} = 0$ in otherwise case, $i \neq j = 1,...,N$. For example, the constant d_0 could define the minimum flow of funds from one region to another or the maximum distance (in km) between them.

2. Generalization and decomposition of spatial autocorrelation coefficients

Let x_{it} be the i-th observation of the t-th variable, $i = 1,...,N$, $t = 1,...,k$. These data are elements of $X = [x_{it}]$ matrix of dimension $N \times k$, $k \leq N$, $X = [x_{*1}...x_{*t}...x_{*k}]$, where x_{*t} is the t-th column of X, $x_{*t}^T = [x_{1t}...x_{it}...x_{Nt}]$. The i-th row of X is denoted by $x_{i*} = [x_{i1}...x_{it}...x_{ik}]$. In particular, for $k = 1$, $X = [x_{11}\, x_{21}...x_{N1}]^T = [x_1\, x_2...x_N]^T$. The variance-covariance matrix is denoted by $V = [v_{jt}]$ where $v_{jt} = \frac{1}{N} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)(x_{it} - \bar{x}_t)$, $\bar{x}_t = \frac{1}{N} \sum_{i=1}^{N} x_{it}, t, j = 1,...,h$. We assume that V is nonsingular.

Du et al. (2012)) proposed the following generalization of Geary's coefficient:

$$I_k^G = \frac{N-1}{2kNw} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_{i*} - x_{j*}) V^{-1} (x_{i*} - x_{j*})^T w_{ij} = \frac{N-1}{2Nk} \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij} q_{ij}. \tag{3}$$

where q_{ij} is explained below the expression (1) and

$$d_{ij} = (x_{i*} - x_{j*}) V^{-1} (x_{i*} - x_{j*})^T \tag{4}$$

is the Mahalanobis distance between x_i* and x_j*. Values of I_k^G close to unity indicate lack of

similarity or differences of neighboring objects due to the multidimensional variable value vectors observed in them than in the case of all the objects (not necessary neighbours).

When $I_k^G > 1$, there is a tendency that neighboring objects are more dissimilar from each other in terms of Mahanalobis distance than in the case of $I_k^G < 1$. For instance, taking into account the aforementioned suggestion of Getis and Ord (1992) we can assume that $w_{ij} = 1$ if $d_{ij} \geq d_0$ and $w_{ij} = 0$ in otherwise case, $i \neq j = 1,...,N$, $d_0 > 0$.

Let us generalize Moran's coefficient for the case when $k \geq 1$ as follows:

$$I_k^M = \frac{1}{wk} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_{i*} - \bar{x}) V^{-1} (x_{j*} - \bar{x})^T w_{ij} = \frac{1}{k} \sum_{i=1}^{N} \sum_{j=1}^{N} b_{ij} q_{ij} \tag{5}$$

where

$$b_{ij} = (x_{i*} - \bar{x}) V^{-1} (x_{j*} - \bar{x})^T. \tag{6}$$

Positive values of I_k^M coefficient indicate that the observations of the vectors of the multivariate variable are similar in terms of the direction of their deviation from the vector of means. If the observation vectors of variables in neighboring objects deviate from the average vector in different directions, then we can expect that the autocorrelation coefficient is negative. Values of the autocorrelation coefficient close to zero indicate lack of similarity or dissimilarity of neighboring objects due to the multivariate variable. Just like it was in the case of I^G we can assume that $w_{ij} = 1$ if $b_{ij} \geq b_0$ and $w_{ij} = 0$ in otherwise case, $b_0 > 0$, $i \neq j = 1,...,N$.

In order to decompose the coefficients let us assume that C is such orthogonal matrix that $C^T C = U_k$ and $C^T V C = \lambda$ where U_k is $k \times k$ identity matrix, $\lambda = [\lambda_t]$ is the diagonal matrix consisting of the eigenvalues of V denoted by $\lambda_t \geq 0$, $t = 1,...,k$, see, e.g. Harville (1997) or Morrison (1976). Note that $V c_t = \lambda_t U_k$ where c_t is the t-th column of C, $c_t^T = [c_{1t}...c_{kt}]$ and it is the t-th eigenvector of V. Observations of the t-th principal component are determined by $z_t = X c_t$. The components of the vector $\lambda_t c_t$ are covariances between the t-th principal component z_t and the entire variables represented by the columns of X. The correlation coefficient between the t-th principal component and observations of the i-th original variable represented by the column x_{*i} is as follows:

$$r(z_t, x_{*i}) = c_{it} \sqrt{\frac{\lambda_t}{v_i}}, \quad i = 1,...,k. \tag{7}$$

In Appendix we show that the generalised Moran coefficient could be decomposed as follows:

$$I_k^M = \frac{1}{k} \sum_{t=1}^{k} I_{k,t}^M \tag{8}$$

where

$$I_{k,t}^M = \frac{1}{\lambda_t} \sum_{i=1}^{N} \sum_{j=1}^{N} (z_{it} - \bar{z}_t)(z_{jt} - \bar{z}_t) q_{ij} \tag{9}$$

is the ordinary Moran spatial autocorrelation coefficient calculated based for the t-th principal component of X. Hence, I_k^M is the average of the Moran autocorrelation coefficients

calculated for the principal components. If this average is equal to zero, the coefficients for the principal components may be non-zero – they happen to cancel each other out.

Similarly to (8) we derive (see Appendix) the following decomposition of Geary's coefficient:

$$I_k^G = \frac{1}{k}\sum_{t=1}^{k} I_{k,t}^G \tag{10}$$

where

$$I_{k,t}^G = \frac{N-1}{2N\lambda_t}\sum_{i=1}^{N}\sum_{j=1}^{N}(z_{it}-z_{jt})^2 q_{ij} \tag{11}$$

is the ordinary Geary's spatial autocorrelation coefficient calculated based on the t-th principal component of X. Thus, I_k^G is average of Geary's autocorrelation coefficients calculated for the principal components.

Example

We illustrate the generalised autocorrelation coefficient with an example of the following variables defined for Polish voivodships: revenues from total economic activity (x_1), sold production of industry (x_2), capital expenditures per capita (x_3), gross value of fixed assets per capita (x_4), average monthly gross salaries (x_5). Data are available at: https://bdl.stat.gov.pl/bdl/start. Variables have been scaled to have the value of each variable divided by the value assigned to the capital voivodship.

The values of the ordinary Moran autocorrelation coefficient (see expression (1)) for the listed variables $x_1,...,x_5$ are as follows: -0.2242, -0.3408, -0.2328, 0.2478, -0.2227. So, all Moran's coefficients are negative except x_4. The values of the ordinary Geary autocorrelation (given by expression (5)) for these variables are as follows: 2.7022, 2.7408, 2.4931, 1.8085, 2.5498.

Moran's and Geary's generalised coefficients take the following values -0.0332 and 1.0039, respectively. Thus, both coefficients would indicate that the spatial autocorrelation for all variables is very weak.

Now, let us consider the decomposition of the generalised coefficients. The eigenvalues (variances of principal components) of the considered $x_1,...,x_5$ are: 0.1633, 0.0318, 0.0082, 0.0060, 0.0018. The shares of these eigenvalues in their sum are as follows (%): 77.3, 15.1 3.9. 2.9, 0.8. The first two principal components explain 92.4% of the overall variation of $x_1,...,x_5$. Thus, the first two principal components explain almost all of the variability of $x_1,...,x_5$. So, the other three principal components can be ignored.

The Moran coefficient for the successive principal components are as follows: -0.2811, 0.3227, -0.0494, -0.7024 and -0.0856. The Geary coefficient for the successive principal components are as follows: 2.8019, 1.3663, 1.7604, 1.7323 and 2.3778.

The matrix of the ordinary correlation coefficients between the principal components and the original variables is as follows:

$$\begin{bmatrix} -0.9563 & -0.9167 & -0.8233 & 0.7170 & -0.8099 \\ 0.1825 & -0.3515 & -0.2944 & -0.6659 & -0.1934 \\ 0.0048 & -0.1692 & 0.4652 & 0.0167 & 0.2725 \\ 0.2270 & -0.8640 & -0.1176 & 0.2072 & 0.0776 \\ 0.0247 & -0.9779 & -0.0726 & 0.0137 & -0.4759 \end{bmatrix} \tag{12}$$

In the i-th row there are correlation coefficients between the i-th principal component and the original variable, $i = 1,...,5$. The first principal component representing the dispersion of all the original variables is strongly correlated with the original variables (see the first row of the matrix given by expression (12)). The second and last principal components are distinctly correlated with the original variables denoted by x_2, x_4 and x_2, x_5, respectively. The third and fourth principal components are rather clearly correlated with variables x_3 and x_2, respectively. The last three principal components explain less than 9% variability of the original variables. Therefore, it suffices to consider only spatial autocorrelation coefficient for the first and second component. Moran's and Geary's coefficients calculated on the basis of the first component are -0.2811 and 2.8019, respectively. Therefore, it can be concluded that neighboring Polish voivodeships differ in their observations of the first principal component. Moran's and Geary's coefficients calculated on the basis of the second component are 0.3227 and 1.3663, respectively. In this case, the coefficient indicated similarity and dissimilarity, respectively.

Note that the values of both the generalised Moran and Geary coefficients (calculated for the original vector observations) are close to zero and one, respectively. This means that there is no tendency to similarity or dissimilarity between the values of a multivariate variable observed on neighboring objects. In our case this is due to the fact that the generalised autocorrelation coefficients are the average values of the ordinary autocorrelation coefficients calculated for the individual principal components of a multivariate variable.

3. Conclusions

The results of considerations on the properties of generalised spatial autocorrelation coefficients of the population objects characterized by observation vectors of a multidimensional variable are as follows. For this purpose, a generalization of the Moran coefficient was defined in a similar way to the generalization of the Geary coefficient proposed by Du et al. (2012). Both generalised coefficients indicate the degree of similarity between neighboring objects due to the distance between the observation vectors of the multidimensional variable observed in them. The principal components of a multivariate variable allow for the presentation of each of the generalised coefficients as the arithmetic mean of the ordinary spatial autocorrelation coefficients, but calculated on the basis of the principal components. It was shown that the decomposition of the original variable into principal components can lead to a substantial simplification of the analysis of multivariate spatial autocorrelation. Moreover, it was concluded that the interpretation of the generalised autocorrelation coefficients may lead to misleading results and therefore must be carried out simultaneously with

the analysis of ordinary autocorrelation coefficients determined on the basis of individual principal components. Finally, we can say that the obtained results allow the use of principal component analysis to enrich the interpretation of generalised spatial autocorrelation coefficients.

Acknowledgements

The author would like to thank the reviewers for their valuable comments on this article.

References

Cliff, A. D., Ord, J. K., (1981). Spatial Processes: Models and Applications. Pion, London.

Du, Z., Jeong, J. S., Jeong, M. K. and Kong, S. G., (2012). Multidimensional local spatial autocorrelation measure for integrating spatial and spectral information in hyperspectral image band selection. *Applied Intelligence*, 36, pp. 542–552.

Geary, R. C., (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5 (3), pp. 115–145.

Getis, A., Ord, J. K., (1992). The analysis of spatial association by use of distance statistic. *Geographical Analysis*, 24(3), pp. 189–206.

Griffith, D. A., Chun, Y., (2022). Some useful details about Moran coefficient, Geary ratio and the joint count indices of spatial autocorrelation. *Journal of Spatial Econometric*, 3:12.

Harville, D. A., (1997). *Matrix Algebra from a Statistician's Perspective*, Springer New York, Berlin, Heidelberg, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo.

Krzyśko, M., Nijkamp, P., Ratajczak, W., Wołyński, W., Wojtyła, A. and Wenerska, B., (2023). A novel spatio-temporal principal component analysis based on Geary's contiguity ratio. *Computers, Environment and Urban Systems*, 103, pp. 1–8.

Krzyśko, M., Nijkamp, P., Ratajczak, W.,Wołyński, W., Wojtyła, A. and Wenerska, B., (2024). Spatio-temporal principal component analysis. *Spatial Economic Analysis*, 19:1, pp. 8–29. doi: 10.1080/17421772.2023.2237532.

Leontief, W. W., (1986). *Input - Output Economics*, Oxford University Press, New York.

Moran, P. A. P., (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, 37 (1), pp. 17–23. doi:10.2307/2332142.

Morrison, D. F., (1976). *Multidimensional Statistical Methods*, McGraw-Hill New York.

Overmars, K. P., de Koning, G. H. J. and Veldkamp, A., (2003). Spatial autocorrelation in multi-scale land use models. *Ecological Modelling*, 164, pp. 257–270.

Appendix

According to notation introduced in Section 2, the equation $CVC^T = \lambda$ is transformed to the following $V = C^T \lambda C$ because $C^{-1} = C^T$. The t-the principal component is determined by $z_{*t} = Xc_t$, $t = 1,...,k$ and $Z = [z_{*1}...z_{*k}] = XC$, $C = [c_1...c_k]$.

The equation $C^{-1} = C^T$ let us write $V^{-1} = (C\lambda C^T)^{-1} = (C^T)^{-1}(C\lambda)^{-1} = C\lambda^{-1}C^T$. Moreover: $ZC^T = X$, $z_{i*}C^T = x_{i*}$, $\bar{x} = U_N^T X/N = U_N^T ZC^T/N = \bar{z}C^T$, $i = 1,...,N$. These results let us rewrite the equation (6) as follows:

$$\begin{aligned} b_{ij} &= (z_{i*}C^T - \bar{z}C^T)V^{-1}(z_{j*}C^T - \bar{z}C^T)^T = (z_{i*} - \bar{z})C^T V^{-1} C(z_{j*} - \bar{z})^T = \\ &= (z_{i*} - \bar{z})C^T C\lambda^{-1}C^T C(z_{j*} - \bar{z})^T = (z_{i*} - \bar{z})\lambda^{-1}(z_{j*} - \bar{z})^T = \\ &= [(z_{i1} - \bar{z}_1)...(z_{it} - \bar{z}_t)...(z_{ik} - \bar{z}_k)][\lambda_t^{-1}][(z_{j1} - \bar{z}_1)...(z_{jt} - \bar{z}_t)...(z_{jk} - \bar{z}_k)]^T = \\ &= [(z_{i1} - \bar{z}_1)\lambda_1^{-1}...(z_{it} - \bar{z}_t)\lambda_t^{-1}...(z_{ik} - \bar{z}_1)\lambda_k^{-1}][(z_{j1} - \bar{z}_1)...(z_{jt} - \bar{z}_t)...(z_{jk} - \bar{z}_k)]^T = \\ &= \sum_{t=1}^{k}(z_{it} - \bar{z}_t)\lambda_t^{-1}(z_{jt} - \bar{z}_t) = \frac{1}{\lambda_t}\sum_{t=1}^{k}(z_{it} - \bar{z}_t)(z_{jt} - \bar{z}_t). \end{aligned}$$

This and equations (1) and (5) lead to the following:

$$\begin{aligned} I_k^M &= \sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij}q_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{1}{\lambda_t}\sum_{t=1}^{k}(z_{it} - \bar{z}_t)(z_{jt} - \bar{z}_t)q_{ij} = \\ &= \sum_{t=1}^{k}\frac{1}{\lambda_t}\sum_{i=1}^{N}\sum_{j=1}^{N}(z_{it} - \bar{z}_t)(z_{jt} - \bar{z}_t)q_{ij}. \end{aligned}$$

This directly leads to equation (8).

Similarly, equation (10) could be derived as follows:

$$\begin{aligned} d_{ij} &= (z_{i*}C^T - z_{j*}C^T)V^{-1}(z_{i*}C^T - z_{j*}C^T)^T = (z_{i*} - z_{j*})C^T V^{-1} C(z_{i*} - z_{j*})^T = \\ &= (z_{i*} - z_{j*})C^T C\lambda^{-1}C^T C(z_{i*} - z_{j*})^T = (z_{i*} - z_{j*})\lambda^{-1}(z_{i*} - z_{j*})^T = \\ &= [(z_{i1} - z_{j1})...(z_{it} - z_{jt})...(z_{ik} - z_{jk})][\lambda_t^{-1}][(z_{i1} - z_{j1})...(z_{it} - z_{jt})...(z_{ik} - z_{jk})]^T = \\ &= [(z_{i1} - z_{j1})\lambda_1^{-1}...(z_{it} - z_{jt})\lambda_t^{-1}...(z_{ik} - z_{jk})\lambda_k^{-1}][(z_{j1} - z_{j1})...(z_{jk} - z_{jk})]^T = \\ &= \frac{1}{\lambda_t}\sum_{t=1}^{k}(z_{it} - z_{jt})^2. \end{aligned}$$

This and equations (2), (3) lead to the following:

$$I_k^G = \frac{N-1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}q_{ij} = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{N-1}{2N\lambda_t}\sum_{t=1}^{k}(z_{it}-z_{jt})^2 q_{ij} =$$

$$= \sum_{t=1}^{k}\frac{N-1}{2N\lambda_t}\sum_{i=1}^{N}\sum_{j=1}^{N}(z_{it}-z_{jt})^2 q_{ij}.$$

This directly leads to equation (10).

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 9–28, https://doi.org/10.59139/stattrans-2025-002
Received – 10.07.2024; accepted – 04.12.2024

An expectation-maximization algorithm for logistic regression based on individual-level predictors and aggregate-level response

Zheng Xu[1]

Abstract

Logistic regression is widely used in complex data analysis. When predictors are at individual level and the response at aggregate level, logistic regression can be estimated using the Maximum Likelihood Estimation (MLE) method with the joint likelihood function formed by Poisson binomial distributions. When directly maximizing the complicated likelihood function, the performance of MLE will worsen as the number of predictors increases. In this article, we propose an expectation-maximization (EM) algorithm to avoid the direct maximization of the complicated likelihood function. Simulation studies have been conducted to evaluate the performance of our EM estimator compared to different estimators proposed in the literature. Two real data-based studies have been conducted to illustrate the use of the different estimators. Our EM estimator proves efficient for the logistic regression problem with an aggregate-level response and individual-level predictors.

Key words: expectation-maximization algorithm, missing values, Poisson binomial distribution, logistic regression, data aggregation, numerical optimization.

1. Introduction

With the fast development in technology, massive complex data have been collected from multiple sources. New methods have been proposed for complex data situations such as (1) how to deal with semi-structured data and structured data in the web (Zhai and Liu, 2006; Getdoor and Mihalkova, 2011), (2) analysis of graph-structured data (Geamsakul et al., 27 2005; Henaff et al., 2015) and (3) multi-level and mixed-level data analysis (Primo et al., 2007; Saramago et al., 2012).

Data can be collected, reported, and are available at different levels due to a range of reasons such as confidentiality, data collection difficulty, and cost saving. For example, the United State Department of Agriculture (USDA) National Agricultural Statistical Services (NASS) (`https://www.nass.usda.gov/`) reports agricultural crop yields at the county level instead of at the farm level, where county-level average or total is aggregated or estimated based on farm-level data in each county and farm-level data are confidential and unavailable to the public. Business data may only publish aggregated commodity purchase data at the store level and the month level to the public and keep individual-level data and

[1]Correspondence Author. Department of Mathematics and Statistics, Wright State University, Dayton, OH, USA. E-mail: zheng.xu@wright.edu. ORCID: https://orcid.org/0000-0003-0311-7004.

daily data confidential. Biological data, social-economic data, survey data, business data are often collected and reported at different levels.

Data can be aggregated in different ways. For example, a sequential two-stage testing method is used to study infectious diseases in epidemiology and bio-statistics. In the first stage, group testing is applied to the combined sample. In the second stage, individuals showing positive in the first stage are called for testing at the individual level. This group-testing strategy has been widely used in coronavirus disease 2019 (COVID-19) testing to increase efficiency and reduce cost (Mercer and Salit, 2021). Group-level Y in Group i can be calculated via the formula $Y_i = 1(\sum_{j=1}^{n_i} Y_{ij} > 0)$, where Y_{ij} is the response for the j-th person in group i, n_i is the number of individuals in group i, and $1(.)$ is the indicator function. The US Census Bureau reports household income as aggregate-level Y and individual income as individual-level Y, the aggregation method is summation, i.e. $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, where Y_{ij} is the income of the j-th person in the i-th household.

When individual-level X and individual-level Y are modeled by logistic regression, individual-level Y follows a Binomial Distribution with success probability as a function of individual-level X, denoted as $\pi(X) = \exp(X^T\beta)/(1+\exp(X^T\beta))$. Then aggregate-level Y, as the sum of individual-level Y, follows a Poisson-Binomial distribution (Hong, 2013; Xu, 2023). A complicated likelihood function $L(\beta)$ is derived and we previously proposed MLE estimator $\hat{\beta}_{MLE} = \text{argmax}_\beta L(\beta)$ with satisfactory statistical performance (Xu, 2023).

Because the maximization of the complicated likelihood function $L(\beta)$ is with respect to $\beta \in \mathscr{R}^p$, the performance of $\hat{\beta}_{MLE}$ will decrease when the dimension p increases (Xu, 2023). Different optimization methods to maximize the likelihood function have been considered and compared in our previous study (Xu, 2023).

We noticed that the limitation of $\hat{\beta}_{MLE}$ is mainly due to the direct optimization of the complicated likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$, formed by Poisson binomial distributions. This optimization can be avoided when an expectation-maximization (EM) method is adopted. As stated in Hastie et al (2009) and Givens and Hoeting (2012), the EM algorithm is a popular tool for simplifying difficult maximum likelihood problems for which maximization of the likelihood function is difficult, but made easier by enlarging the sample with latent data, i.e. a data-argumentation process. For our logistic regression problem with individual-level X and aggregate-level Y, we can enlarge the sample with the latent individual-level Y. One reason for using latent individual-level Y is that the usual logistic regression can be easily conducted when both X and Y are at the same level. Under mild conditions, this usual logistic regression has a unique MLE solution as a convex optimization problem with a convex objective function (Agresti, 2013; Hilbe, 2009). The unique solution can be obtained numerically via Newton's method, which uses the observed second derivative or the Fisher scoring method, which uses the expectation of this second derivative, and the Fisher scoring method is an application of the method of iteratively reweighted least squares (IRLS) (Agresti, 2013; Hilbe, 2009). Our EM algorithm conducts the usual logistic regression using IRLS method with stable performance and avoids the difficult optimization of the complicated likelihood function. Another reason to propose our estimator as an EM algorithm is that the EM algorithm view our problem in the perspective of missing values and data augmentation. This different perspective, compared with our previously proposed MLE estimator, makes our problem easily adapted and extensible to more complicated but

similar problems including (1) the problem that predictors X themselves are at mixed levels, (2) the problem that both X and Y contain missing values and (3) the problem that X and Y are modeled via a generalized linear model (GLM). Both the EM algorithm and our previously proposed MLE estimator in Xu (2023) have their own advantage in model extension to solve more complicated data situations, and choosing which is better depends on specific data situations. Therefore, both EM estimator and MLE estimator are necessary in methodological development of logistic regression.

The aim of this article is to study the performance of EM estimator in logistic regression based on aggregate-level Y and individual-level X. We proposed our EM estimator in Section 2. We conducted simulation studies to evaluate the performance and compare our EM estimator with literature estimators in Section 3. We illustrated the use of different estimators in real data-based studies in Section 4. We provided discussion in Section 5 and made conclusions in Section 6.

2. Materials and Methods

2.1. Data and Model Specification

Suppose there are N independent individuals aggregated into M groups, with group i having n_i individuals, i.e. $N = \sum_{i=1}^{M} n_i$. Denote (X_{ij}, Y_{ij}), $X_{ij} \in \mathscr{R}^p$, $Y_{ij} \in \mathscr{R}$, $i = 1,2,\ldots,M$, $j = 1,2,\ldots,n_i$, as the predictor vector and the response for the j-th individual in the i-th group. Thus, X_{ij} and Y_{ij} are individual-level predictor vector (X) and individual-level response (Y). Aggregate-level Y is obtained by summation within a group, i.e. $Y_i = \sum_{j=1}^{n_i} Y_{ij}$. Suppose there is a logistic regression model at the individual level, i.e.

$$\ln\left(\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right) = X_{ij}^T\beta,\ i = 1,2,...,M,\ j = 1,2,...,n_i. \tag{1}$$

Then $Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$, where $\pi_{ij} = P(Y_{ij}=1) = \frac{\exp(X_{ij}^T\beta)}{1+\exp(X_{ij}^T\beta)}$. When individual-level X and individual-level Y are both available, the logistic model as a generalized linear model (GLM) can be estimated using a range of methods including the Newton-Raphson method and the Fisher scoring method and the Fisher scoring method is an application of the method of iteratively reweighted least squares (IRLS) (Agresti, 2013; Givens and Hoeting, 2012). We name the logistic regression based on X and Y at the same level as the "usual" logistic regression (Agresti, 2013; Givens and Hoeting, 2012), to be compared with our problem of conducting logistic regression based on individual-level X and aggregate-level Y, which was considered in Xu (2023) and this article.

2.2. Joint Likelihood and MLE Method

Then the distribution of aggregate-level response, $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, as the sum of n_i independent Bernoulli random variables $Y_{ij} \sim \text{Bernoulli}(\pi_{ij})$, $j = 1,2,\ldots,n_i$, is a Poisson

binomial distribution, i.e.

$$Y_i \sim \text{PoissonBinomial}(n_i, (\pi_{i1}, \pi_{i2}, \cdots, \pi_{in_i})), \tag{2}$$

where $\pi_{ij} = P(Y_{ij} = 1) = \frac{\exp(X_{ij}^T \beta)}{1+\exp(X_{ij}^T \beta)}$ (Wang, 1993; Hong, 2013; Xu, 2023).

The joint likelihood function is

$$L(\beta) = \prod_{i=1}^{M} P(Y_i | X_{i1}, ..., X_{in_i}; \beta), \tag{3}$$

where $P(Y_i | X_{i1}, ..., X_{in_i}; \beta)$ is the probability of Y_i, as specified in Equation 2.

The calculation of probability for a Poisson binomial distribution is complicated. In general, for a variable $Y \sim \text{PoissonBinomial}(n, (\pi_1, \pi_2, \ldots, \pi_n))$, the probability mass function is $P(Y = y) = \sum_{A \in F_y} \prod_{i \in A} \pi_i \prod_{j \in A^c} (1 - \pi_j)$, where F_y is the set of all subsets of y integers that can be selected from $\{1, 2, 3, ..., n\}$ and A^c is the complement of A (Wang, 1993). The set F_k contains $\binom{n}{k}$ elements so the sum over it is computationally intensive and even infeasible for large n. Instead, more efficient ways were proposed, including the use of a recursive formula to calculate $P(Y = y)$ based on $P(Y = k)$, $k = 0, ..., y-1$, which is numerically unstable for large n (Chen et al., 1994), and the inverse Fourier transform method (Fernandez and Williams, 2010). Hong (2013) further developed it by proposing an algorithm that efficiently implements the exact formula with a closed expression for the Poisson binomial distribution (Hong, 2013). We adopted Hong's algorithm and exact formula in calculating the likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$ in Equation 3 since they are more precise and numerically stable (Xu, 2023). Three optimization methods (Nelder and Mead's simplex method (NM) (Nelder and Mead, 1965), the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (Fletcher, 1970), and the conjugate gradient (CG) method (Fletcher and Reeves, 1964)) to maximize the joint likelihood function $L(\beta)$ were compared in Xu (2023) and the three methods show similar performance with NM method slightly better as our recommended method, and NM method is derivative free (Xu, 2023; Givens and Hoeting, 2012). We note that along the category of methods of directly optimizing the likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$, there can be a range of potential methods including evolutional algorithm and simulated annealing which may have similar or even better performance compared with our recommended directly optimization method (Givens and Hoeting, 2012; Xu, 2023). The search of optimization methods which directly maximizes $L(\beta)$ is not the objective of this article. Instead, we intend to develop methods not in this category of methods, i.e. methods not directly maximizing $L(\beta)$.

2.3. Expectation Maximization Algorithm

As an optimization problem $\max_\beta L(\beta)$, $\beta \in \mathscr{R}^p$, its performance will become worse when the number of predictors p increases. The objective function $L(\beta)$ is a complicated likelihood function so that we consider whether it is possible to circumvent or avoid the direct optimization of $L(\beta)$.

We noticed that the "usual" logistic regression, i.e. logistic regression when X and Y are

at the "same" level, is numerically stable and relatively easy to calculate. However, for our data situation, the usual logistic regression is infeasible because individual-level X is not available. To address this issue, we view our problem as a missing value problem where the latent variable is individual-level Y and we adopt an EM algorithm to substitute it. In each iteration of the EM algorithm, the usual logistic regression is conducted with individual-level Y, i.e. Y_{ij}, substituted with its expectation given current-iteration parameter estimate, i.e. $E(Y_{ij}|Y_i, \beta^{(k)})$, where $\beta^{(k)}$ is the estimated value of parameter β in iteration k.

Illuminated by the materials of EM algorithm in Hastie et al. (2009) and Givens and Hoeting (2012), we developed the EM algorithm for our problem. The EM algorithm is described as Algorithm 1 in the following page. The estimator obtained via the EM algorithm is named as the EM estimator.

One advantage of EM estimator is that it can avoid the direct optimization of the complicated nonlinear likelihood function $L(\beta)$. EM algorithm conducts the usual weighted logistic regression in each iteration. EM estimator is expected to have similar performance or even potentially slightly better performance compared with our MLE estimator in Xu (2023), which directly maximizes $L(\beta)$.

Another advantage of EM estimator is that it views our problem in a different perspective, i.e. missing values and data augmentation. This makes our method easily adapted and extensible for some applications. Potential applications which our EM algorithm may solve after modifications include (1) the situation where X are at mixed levels, i.e. different predictors are at levels, (2) the situation where there are missing values in X and Y, (3) the situation where individual-level X and Y is described by a generalized linear model (GLM), and (4) the situation where the objective is to use a variational Bayes to find a posteriori estimation (MAP) and make use of prior information (Bernardo et al., 2003).

3. Simulation Studies

3.1. Simulation Setups

We conducted simulation studies to evaluate the performance of the following four estimators. Estimator 1, named as "individual-LR", is the logistic regression estimator based on individual-level X and individual-level Y. This estimator is infeasible in our data situation where only aggregate-level Y is available. Because aggregate-level Y contains less information compared to individual-level Y, we expect that this infeasible estimator can provide an upper bound for the performance of feasible estimators based on aggregate-level Y. Estimator 2, named as "naive–LR", is the logistic regression estimator based on the aggregate-level X, which is the mean of X in each group, and the aggregate-level Y, i.e. $Y_i \sim \text{Binomial}(n_i, \sum_{j=1}^{n_i} X_{ij}/n_i)$, $i = 1, 2, \ldots, M$. This estimator can provide a rough approximate estimation. Estimators 3 is our previously recommended MLE estimator via Nelder-Mead optimization (Xu, 2023). Estimator 4 is our proposed EM estimator as described above. The performances of these estimators were compared under three scenarios. In each scenario, simulations were conducted with the number of groups ($M = 300, 500, 1000$), and equal group sizes ($n_i = 5, 10$, $i = 1, 2, \ldots, M$). The setup of data generation is specified as follows:

Algorithm 1 EM Algorithm for Logistic Regression Based on Individual-Level X and Aggregate-Level Y

1. Start with initial value for the parameter β, i.e. $\hat{\beta}^{(0)}$, where the initial value is obtained from the following values: (1) estimated value by the usual logistic regression of aggregate-level Y on aggregate-level X, (2) MLE estimate in Xu (2023), (3) the zero vector $(0,0,\ldots,0) \in \mathscr{R}^p$, (4) the unit vector $(1,1,\ldots,1) \in \mathscr{R}^p$, and (5) the vector $(-1,-1,\ldots,-1) \in \mathscr{R}^p$.

2. Expectation Step: at the j-th step, compute

$$\begin{aligned} Q(\beta', \hat{\beta}^{(j)}) &= E(l(\beta'; \{Y_{ij}\}) | \{Y_i\}, \hat{\beta}^{(j)}) \qquad (4) \\ &= \sum_{i=1}^{M} \sum_{j=1}^{n_i} \{E(Y_{ij}|Y_i, \hat{\beta}^{(j)}) \ln(\pi_{ij}) + (1 - E(Y_{ij}|Y_i, \hat{\beta}^{(j)})) \ln(1-\pi_{ij})\} \end{aligned}$$

   as a function of the dummy argument β'. The expected value of latent value Y_{ij} is computed via the formula

$$\begin{aligned} & E(Y_{ij}|Y_i = y, \hat{\beta}^{(j)}) \qquad (5) \\ = \; & P(Y_{ij} = 1 | Y_i = y, \hat{\beta}^{(j)}) = \frac{P(Y_{ij}=1) P(Y_i - Y_{ij} = y-1)}{P(Y_i = y)} \\ = \; & \frac{\pi_{ij} \times \text{PoissonBinomial}(y-1; n_i - 1, \pi_{i1}, \ldots, \pi_{i,j-1}, \pi_{i,j+1}, \ldots, \pi_{in_i})}{\text{PoissonBinomial}(y-1; n_i, \pi_{i1}, \pi_{i2}, \ldots, \pi_{in_i})}, \end{aligned}$$

   where PoissonBinomial(.) is the probability mass function of a Poisson binomial distribution, and $\pi_{ij} = \exp(X_{ij}^T \beta')/(1+\exp(X_{ij}^T \beta'))$. As the convention in regression analysis, we can treat X as fixed. For random X, we can use the argument of conditioning Y on X and this conditioning is equivalent to treating X as fixed (Hastie et al., 2009; Givens and Hoeting, 2012).

3. Maximization Step: determine the new estimate $\hat{\beta}^{(j+1)}$ as the maximizer of $Q(\beta', \hat{\beta}^{(j)})$ over β'. This step is obtained by conducting weighted logistic regression with the likelihood function specified in Equation 4. To be more specific, our dataset has N observations of individual-level X, i.e. X_{ij}, $i = 1,2,\ldots,M$, $j = 1,2,\ldots,n_i$, $N = \sum_{i=1}^{M} n_i$. A pseudo-dataset of $2N$ pseudo-observations is created with the pseudo-observation represented as $(\tilde{X}_{ijk}, \tilde{Y}_{ijk}, \tilde{W}_{ijk})$, $i = 1,2,\ldots,M$, $j = 1,2,\ldots,n_i$, $k = 0,1$, where $\tilde{X}$, $\tilde{Y}$ and $\tilde{W}$ are respectively the predictor vector, response and weight in the pseudo-dataset. For each observation X_{ij}, two pseudo-observations, i.e. $(\tilde{X}_{ij0}, \tilde{Y}_{ij0}, \tilde{W}_{ij0})$ and $(\tilde{X}_{ij1}, \tilde{Y}_{ij1}, \tilde{W}_{ij1})$, are created as follows:

$$\begin{aligned} & \tilde{X}_{ij0} = X_{ij},\ \tilde{Y}_{ij0} = 0,\ \tilde{W}_{ij0} = 1 - E(Y_{ij}|Y_i, \hat{\beta}^{(j)}) \\ & \tilde{X}_{ij1} = X_{ij},\ \tilde{Y}_{ij1} = 1,\ \tilde{W}_{ij1} = E(Y_{ij}|Y_i, \hat{\beta}^{(j)}). \end{aligned}$$

   Weighted logistic regression is conducted based on the pseudo-dataset.

4. Iterate steps 2 and 3 until convergence.

- In Scenario 1, $p = 5$, $(X_{i1}, X_{i2}) \sim \text{multinormal}_2(0_{2\times1}, \Sigma_{2\times2})$, $0_{2\times1} = (0,0)^T$, $\Sigma_{2\times2} = (\sigma_{ij})$ where $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = 0.6$ for $i \neq j$. $X_{i3} \sim \text{t(df} = 2)$, $X_{i4} \sim \text{Bernoulli}(0.5)$, $X_i = (1, X_{i1}, X_{i2}, X_{i3}, X_{i4})^T \in \mathscr{R}^p$, $Y_i \sim \text{Bernoulli}(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-0.5, 1, -0.5, 2, -1.6)^T$.
- In Scenario 2, $p = 10$, $(X_{i1}, X_{i2}, X_{i3}, X_{i4}) \sim \text{multinormal}_4(0_{4\times1}, \Sigma_{4\times4})$, $0_{4\times1} = (0,0,0,0)^T$, $\Sigma_{4\times4} = (\sigma_{ij})$ where $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = 0.6$ for $i \neq j$. $X_{i5} \sim \text{t(df} = 2)$, $X_{i6} \sim \text{t(df} = 4)$, $X_{i7} \sim \text{chi-square(df} = 2)$, $X_{i8} \sim \text{chi-square(df} = 3)$, $X_{i9} \sim \text{Bernoulli}(0.5)$, $X_i = (1, X_{i1}, X_{i2}, \ldots, X_{i9})^T \in \mathscr{R}^p$, $Y_i \sim \text{Bernoulli}(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-0.5, 1, -2.5, 2, -1.6, 0.7, 0.9, -2.4, 0.5, -1.3)^T$.
- In Scenario 3, $p = 20$, $(X_{i1}, X_{i2}, \ldots, X_{i10}) \sim \text{multinormal}_{10}(0_{10\times1}, \Sigma_{10\times10})$, $0_{10\times1} = (0,0,\ldots,0)^T$, $\Sigma_{10\times10} = (\sigma_{ij})$ where $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = 0.6$ for $i \neq j$. $X_{i11} \sim \text{t(df} = 2)$, $X_{i12} \sim \text{t(df} = 4)$, $X_{i13} \sim \text{t(df} = 6)$, $X_{i14} \sim \text{chi-square(df} = 2)$, $X_{i15} \sim \text{chi-square(df} = 3)$, $X_{i16} \sim \text{chi-square(df} = 4)$, $X_{i17} \sim \text{Bernoulli}(0.3)$, $X_{i18} \sim \text{Bernoulli}(0.5)$, $X_{i19} \sim \text{Bernoulli}(0.7)$, $X_i = (1, X_{i1}, X_{i2}, \ldots, X_{i19})^T \in \mathscr{R}^p$, $Y_i \sim \text{Bernoulli}(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-0.5, -0.8969, 0.1848, 1.5878, -1.1304, -0.0803, 0.1324, 0.7080, -0.2397, 1.9845, -0.1388, 0.4177, 0.9818, -0.3927, -1.0397, 1.7822, -2.311, 0.8786, 0.036, 1.013)^T$. Note that the values of the 19 slope coefficients, i.e. $\beta_1, \ldots, \beta_{19}$, were generated as standard normal random variables, and the generation of the values of the 19 slope coefficients was implemented in R language using the command: "set.seed(2); rnorm(19)". The value of the intercept parameter, i.e. β_0, was fixed at -0.5.

3.2. Performance Evaluation Metrics

The squared bias, variance, mean square error (MSE), and mean absolute deviation (MAD) of each of the four estimators' (E1 to E4) model parameters $(\beta_0, ..., \beta_{p-1}) \in \mathscr{R}^p$ were calculated. Denote the bias, variance, MSE, and MAD of the q-th estimator of β_j as $\text{Bias}(\hat{\beta}_{j,E_q})$,$\text{Var}(\hat{\beta}_{j,E_q})$, $\text{MSE}(\hat{\beta}_{j,E_q})$, and $\text{MAD}(\hat{\beta}_{j,E_q})$. The average squared bias, variance, MSE, and MAD of the q-th estimator were calculated as $\overline{\text{Bias}^2}(E_q) = (1/p)\sum_{j=0}^{p-1}\text{Bias}^2(\hat{\beta}_{j,E_q})$, $\overline{\text{Var}}(E_q) = (1/p)\sum_{j=0}^{p-1}\text{Var}(\hat{\beta}_{j,E_q})$, $\overline{\text{MSE}}(E_q) = (1/p)\sum_{j=0}^{p-1}\text{MSE}(\hat{\beta}_{j,E_q})$, and $\overline{\text{MAD}}(E_q) = (1/p)\sum_{j=0}^{p-1}\text{MAD}(\hat{\beta}_{j,E_q})$.

3.3. Simulation Results

In Table 1, we report the average squared biases and variances for the four estimators (Individual-LR, Naive-LR, MLE and EM) under different scenarios, sample sizes, and group sizes. Regarding bias, the infeasible individual-LR shows smallest bias and the naive-LR shows biggest bias. The reason for individual-LR to have smallest bias is that individual-LR conducts the usual logistic regression based on individual-level X and individual-level Y which makes use of more information than available in our data situation where individual-level Y is not available. Naive-LR is found to have the biggest bias, which was explained by the fact that logistic regression model uses a "non-linear" logit link function and Naive-LR conducts a naive rough approximate using the mean of X, which ignores the nonlinearity in

the link function, so that Naive-LR can induce a big bias. The biases of MLE estimator and EM estimator are found to be between the two extremes, i.e. individual-LR and naive-LR.

Regarding variance, individual-LR and naive-LR have relatively smaller variance, compared with MLE estimator and EM estimator. We explained that the smaller variance in individual-LR is because it uses more information than available in our data situation where individual-level Y is not available. The smaller variance in naive-LR is also reasonable. As a poor rough approximate estimator, naive-LR can have big bias and small variance. For example, suppose that a toy estimator always reports a constant value as its estimate. This toy estimator will have zero variance and a big bias. Thus, we put more focus on mean square error (MSE) and mean absolute deviation (MAD) instead of bias and variance in evaluating estimators.

Next, we check MSE and MAD of the four estimators. In Table 2, we report average MSE and average MAD. The infeasible individual-LR estimator shows the best performance in terms of both MSE and MAD. This is because individual-LR estimators makes use of more information than available in our data situation where individual Y is latent. The naive-LR estimator shows the worst performance in terms of both MSE and MAD. This is because naive-LR is a naive rough approximate estimator which can lead to a big bias due to non-linearity in link function. In terms of MSE and MAD, we found our MLE estimator and EM estimator are between the two extremes (individual-LR and naive-LR). Our MLE and EM estimator show similar performance with EM estimator having potentially slightly better performance.

We add a cautionary note that simulation studies cannot substitute theoretical verification. Simulation studies cannot fully assess theoretical properties of estimators. Theoretical properties of MLE estimators and EM estimators have to be inferred based on theoretical literature on MLE and EM.

4. Real Data-Based Studies

We used real data to illustrate the use of our EM estimator and compare it with different estimators in the literature. Two real data-based studies are shown. One study is wine quality modeling based on physico-chemical tests. The other study is maternal health risk modeling. Both studies used the datasets from UC Irvine machine learning repository (`https://archive.ics.uci.edu/`).

4.1. Wine Quality Modeling

We obtained two datasets of wine quality from UC Irvine machine learning repository (Cortez and Reis, 2009). The two datasets are related to red and white verde wine samples, from the north of Portugal. Due to privacy and logistic issues, only physicochemical (inputs, i.e. X) and sensory (the output, i.e. Y) variables are available. The output variable sensory wine quality score is a score between 1 and 10. This wine quality score was dichotomized into a binary variable, which takes the value of 1 (high-quality) or 0 (low-quality) depending on whether the score is between 6 and 10, or between 1 and 5. Thus, as specified in UC Irvine machine learning repository, the wine quality datasets can be used for both clas-

Table 1: Average Squared Bias and Variance of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM). In the columns for average squared bias and average variance, the unit is 0.001.

			Average Squared Bias				**Average Variance**			
Scenario	**M**	n_i	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1	300	5	0.08	718	0.68	0.37	15.8	21.1	35.3	34.0
1	300	10	0.11	788	3.16	0.74	7.4	16.5	48.3	34.5
1	500	5	0.20	717	0.36	0.90	9.0	11.3	30.7	27.6
1	500	10	0.03	788	0.96	1.12	4.8	13.3	36.4	25.7
1	1000	5	0.03	729	0.29	0.31	4.8	8.3	19.3	12.8
1	1000	10	0.01	799	3.57	0.05	2.3	5.9	23.0	11.2
2	300	5	0.72	1007	6.48	4.39	34.3	24.3	91.2	57.5
2	300	10	0.43	1064	25.23	5.08	13.6	23.0	134.9	46.2
2	500	5	0.43	1021	19.14	0.25	14.6	14.3	63.2	26.4
2	500	10	0.18	1063	49.26	1.54	7.8	11.6	96.0	27.0
2	1000	5	0.18	1018	17.67	0.57	7.8	7.7	57.2	15.1
2	1000	10	0.06	1078	49.59	0.37	4.0	6.8	81.5	11.7
3	300	5	6.25	658	178.2	14.39	48.0	27.6	86.3	78.6
3	300	10	2.15	683	282.9	10.05	22.3	25.3	63.8	65.3
3	500	5	1.08	667	200.2	3.23	28.7	15.2	70.0	43.6
3	500	10	0.40	693	300.0	2.46	13.1	14.4	47.5	33.5
3	1000	5	0.40	668	192.5	0.96	13.1	7.5	59.3	20.3
3	1000	10	0.13	689	306.2	1.03	6.4	6.5	35.1	16.0

sification problem (Y is the binary wine quality variable) and regression problem (Y is the wine quality score which is between 1 and 10). There are 11 continuous features/predictors in X. They are (1) fixed acidity, (2) volatile acidity, (3) citric acid, (4) residual sugar, (5) chlorides, (6) free sulfur dioxide, (7) total sulfur dioxide, (8) density, (9) pH, (10) sulphates and (11) alcohol. For more details in the wine quality datasets, please refer to Cortez and Reis (2009).

We used the wine quality datasets to illustrate the use of logistic regression under the data situation of aggregate-level Y and individual-level X. In practice, there are multiple reasons which can contribute to the situation that Y is available at aggregate level instead of individual level. One reason is confidentiality. For example, suppose the objective is to predict or model wine quality provided by some specific wine association or agency. However, the wine association or agency wants to keep its evaluation in confidentiality and do not want its evaluation to be easily modeled or predicted. In addition, the wine

Table 2: Average Mean Squared Error (MSE) and Mean Absolute Deviation (MAD) of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM). In the columns for average MSE and average MAD, the unit is 0.001.

			Average MSE				**Average MAD**			
Scenario	**M**	n_i	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1	300	5	15.9	739	36.0	34.4	95.5	680.7	146.5	140.1
1	300	10	7.5	804	51.5	35.2	63.7	714.5	169.4	146.4
1	500	5	9.2	729	31.0	28.5	74.6	676.7	137.1	130.7
1	500	10	4.9	801	37.4	26.8	53.6	708.0	140.4	125.3
1	1000	5	4.9	738	19.6	13.1	53.6	685.0	100.2	89.0
1	1000	10	2.3	805	26.6	11.3	37.4	715.0	113.7	82.4
2	300	5	35.0	1032	97.7	61.9	136.9	865.6	228.7	188.2
2	300	10	14.0	1087	160.2	51.3	87.5	885.9	290.2	166.8
2	500	5	15.1	1035	82.3	26.7	90.5	869.1	205.7	122.5
2	500	10	8.0	1075	145.2	28.5	65.2	886.7	264.9	126.0
2	1000	5	8.0	1026	74.9	15.6	65.2	868.5	186.5	91.5
2	1000	10	4.1	1084	131.1	12.0	46.2	897.6	248.6	84.0
3	300	5	54.2	685	264.5	93.0	176.7	645.6	401.8	233.7
3	300	10	24.5	709	346.7	75.3	119.1	658.1	465.2	207.6
3	500	5	29.8	682	270.2	46.8	127.9	643.7	413.7	163.8
3	500	10	13.5	708	347.6	36.0	89.6	658.2	472.1	143.3
3	1000	5	13.5	675	251.8	21.3	89.6	639.1	401.4	112.1
3	1000	10	6.6	695	341.3	17.0	61.2	647.9	466.3	98.3

association is interested in ranking wineries or wine firms based on multiple wine samples submitted by each winery or firm. The rule is that each winery or firm is allowed to submit samples from multiple brands the winery or firm owns. The wine association will only specify how many samples are in high-quality in their submission and does not disclose wine quality of each individual wine sample. In this way, the firms will compete with aggregate-level Y available instead of individual-level Y, and the wine association or agency keep its evaluation result of individual samples to be confidential.

We illustrated the use of our EM estimator and other estimators (infeasible individual-LR, naive aggregate-LR, and MLE estimator in Xu (2023)) in the literature for wine quality modeling. There are 4898 observations in white wine data, and 1599 observations in red wine data. We conducted random aggregation with equal group size $n_i = 5$ and 10. For white wine data, there are $979 = 4895/5$ groups of size $n_i = 5$ formed, and $489 = 4890/10$ groups of size $n_i = 10$ formed. Thus, there are 4895 and 4890 observations used in our data

situation $n_i = 5$ and $n_i = 10$ for white wine data. For red wine data, there are $319 = 1595/5$ groups of size $n_i = 5$ formed, and $159 = 1590/10$ groups of size $n_i = 10$ formed. Thus, there are 1595 and 1590 observations used in our data situation $n_i = 5$ and $n_i = 10$ for red wine data.

We showed the estimated values of the estimators based on our data sets with random aggregation. The estimators are: (1) individual-LR, which conducts logistic regression based on individual-level X and individual-level Y. Individual-LR is considered to be the best estimator since it uses more information (individual-level Y) than the information available in our data situation where aggregate-level Y instead of individual-level Y is available. Thus, individual-LR is infeasible. (2) naive-LR, which conducts logistic regression based on aggregate-level X and aggregate-level Y. (3) our previously proposed MLE in Xu (2023). (4) our EM estimator proposed in this article. We illustrated the use of each estimator based on wine quality data and report the estimated values of parameters for white wine data in Table 3 and the estimated values of parameters for red wine data in Table 4. As shown in Table 3 and 4, these estimators reported numerically different values. We recommend the use of EM estimator and MLE estimator, since individual-LR is infeasible and naive-LR can induce a big bias. Because there is no ground truth (true values) of logistic model parameters known in the real data, no statistical performances (such as bias and variance) were evaluated based on the real data.

Table 3: Estimated Values of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On White Wine Quality Data.

	$n_i = 5$				$n_i = 10$			
Variable	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
β_0	0.920	0.719	0.910	0.923	0.920	0.703	0.913	0.917
β_1	0.032	0.048	0.075	0.018	0.033	0.026	0.065	0.042
β_2	-0.651	-0.456	-0.627	-0.636	-0.650	-0.459	-0.645	-0.658
β_3	0.015	0.066	0.075	0.090	0.015	-0.028	0.010	-0.001
β_4	0.865	0.395	0.599	0.451	0.866	0.984	1.435	1.359
β_5	0.020	-0.066	-0.058	-0.075	0.019	-0.131	-0.078	-0.082
β_6	0.163	0.170	0.170	0.223	0.164	0.214	0.264	0.285
β_7	-0.056	-0.066	-0.019	-0.045	-0.057	-0.141	-0.101	-0.112
β_8	-0.812	-0.267	-0.473	-0.212	-0.814	-0.789	-1.202	-1.086
β_9	0.166	0.131	0.172	0.139	0.167	0.217	0.347	0.323
β_{10}	0.205	0.159	0.209	0.191	0.206	0.182	0.397	0.387
β_{11}	0.915	0.840	1.056	1.210	0.911	0.536	0.681	0.749

Table 4: Estimated Values of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Red Wine Quality Data.

	$n_i = 5$				$n_i = 10$			
Variable	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
β_0	0.239	0.143	0.239	0.226	0.237	0.141	0.291	0.295
β_1	0.247	0.367	0.382	0.693	0.251	0.853	0.880	1.000
β_2	-0.585	-0.337	-0.500	-0.571	-0.588	-0.454	-0.702	-0.771
β_3	-0.248	-0.379	-0.371	-0.532	-0.252	-0.282	-0.356	-0.413
β_4	0.079	-0.043	-0.003	0.108	0.080	0.028	-0.137	-0.296
β_5	-0.183	0.083	-0.023	0.052	-0.180	-0.114	-0.144	-0.161
β_6	0.233	0.337	0.260	0.309	0.230	0.273	0.352	0.083
β_7	-0.539	-0.574	-0.684	-0.721	-0.535	-0.288	-0.317	-0.110
β_8	-0.104	0.051	-0.020	-0.197	-0.104	-0.491	-0.478	-0.544
β_9	-0.052	-0.101	-0.217	-0.117	-0.053	0.159	0.079	0.190
β_{10}	0.475	0.224	0.296	0.326	0.469	0.550	0.643	0.642
β_{11}	0.917	0.688	0.993	0.973	0.917	0.451	0.805	0.876

4.2. Maternal Health Risk Modeling

We obtained the dataset of maternal health risk from UC Irvine machine learning repository (Ahmed, 2023; Ahmed et al., 2020). The data were collected through the IoT-based risk monitoring system from a range of hospitals, community clinics, maternal health care in the rural areas of Bangladesh (Ahmed, 2023). The response variable is the binary maternal health risk level (low risk or high risk). The predictors are (1) age, (2) systolic blood pressure, (3) diastolic blood pressure, (4) blood glucose, (5) body temperature, and (6) heart rate. All these predictors are the responsible and significant risk factors for maternal mortality (Ahmed et al., 2020). UC Irvine machine learning repository specify it as a classification problem since the response variable is binary. There are 1013 individual observations in the dataset. For more details in the maternal health risk data, please refer to Ahmed et al. (2020) and Ahmed (2023).

We conducted random aggregation on the data. There are 202=1010/5 groups of size $n_i = 5$ and 101=1010/10 groups of size $n_i = 10$ formed. Thus, there are 1010 observations in our study of maternal health risk modeling.

Based on the maternal health risk data with random aggregation, we conducted individual-LR, naive-LR, MLE in Xu (2023) and EM estimator proposed in the article. The estimated values of these estimators are shown in Table 5. There is numerical difference in the estimated values of different estimators. We recommend the use of our proposed EM estimator and our previously proposed MLE estimator in the study.

Table 5: Estimated Values of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Maternal Health Risk Data.

	$n_i = 5$				$n_i = 10$			
Variable	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
β_0	0.913	0.544	0.785	0.784	0.913	0.546	0.592	0.669
β_1	-0.079	-0.062	-0.075	-0.079	-0.079	-0.075	-0.019	-0.113
β_2	1.116	2.014	1.060	1.059	1.116	3.413	1.138	1.102
β_3	-0.365	-0.551	-0.346	-0.345	-0.365	-1.205	-0.357	-0.433
β_4	1.631	0.717	1.333	1.329	1.631	0.460	0.640	1.032
β_5	0.668	0.998	0.652	0.650	0.668	1.388	0.746	0.594
β_6	0.272	0.177	0.213	0.214	0.272	-0.104	0.433	0.228

5. Discussion

There are at least two categories of methods to solve the problem of logistic regression based on individual-level X and aggregate-level Y. The first category is directly maximizing the complicated likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$ to find MLE as we previously proposed in Xu (2023). The second category is to avoid the direct optimization of the likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$ using the EM algorithm as we propose in this article. In theory, both categories of methods are valid. Similar but slightly different performances are expected theoretically. We note that the two categories of methods are generic so that there are a range of ways in each category. Along the first category, i.e. obtaining MLE by directly maximizing $L(\beta)$, $\beta \in \mathscr{R}^p$, there can be a range of optimization methods with slightly better or worse performance. A non-exhaustive list of these methods includes: (1) Nelder and Mead's simplex method (NM) (Nelder and Mead, 1965), (2) the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (Fletcher, 1970), (3) the conjugate gradient (CG) method (Fletcher and Reeves, 1964), (4) simulated annealing (Brooks and Morgan, 2018), and (5) evolutional algorithm (Lambora et al., 2019), and their combinations or variants such as Generalized simulated annealing (GSA) and variable step size generalized simulated annealing (VGSA) (Kalivas, 1992). Along the second category avoiding directly maximization of likelihood function $L(\beta)$, $\beta \in \mathscr{R}^p$, there can be a range of methods including (1) the standard EM (McLachlan and Krishnan, 2007), (2) Monte Carlo EM (Wei and Tanner, 1990), and (3) variational Bayes EM (Bernardo et al., 2003).

Both categories of methods have their own advantages and are necessary for the logistic regression based on individual-level X and aggregate-level Y. Which category of methods to use in practice depends on the specific problem. The advantages of the second category of methods, including EM algorithms, Bayes methods and their variants, are the convenience in solving a range of data situations including missing values. Along this category of methods, methods can be potentially adapted to solve data situations such as (1) the sit-

uation where X and Y are at mixed-levels, (2) the situation where X and Y contain missing values, (3) the situation where prior information is preferred to use or consider, (4) the situation where individual-level X and individual-level Y is described by a generalized linear model, which can be a linear model, logistic model, Poisson model or other generalized linear models. In comparison, the advantages of the first category of methods include (1) there are a range of optimization methods to try, (2) the potential further extension of methods to penalized likelihood functions which will add a penalty term such as L_p norm of model parameter β with $1 \leq p \leq 2$ to the current complex likelihood function (Hastie et al., 2009), and (3) likelihood-based statistical inferences such as likelihood ratio test, score test, standard errors, and confidence intervals. Studies of these extensions are beyond the scope of this article and are under development as future work.

The dimension p, i.e. the number of predictors, influences the performance of EM estimator and MLE estimator. Given the sample size n, both EM performance and MLE performance are expected to decrease when p increases. The deterioration of both performances with the increase in p is as expected since the optimization problem $\max L(\beta)$, $\beta \in \mathscr{R}^p$ in theory will decrease when p increases, given a fixed sample size n. Both EM and MLE will maximize $L(\beta)$, either indirectly or directly.

However, we need to note although EM algorithms always have likelihood non-decreasing in each step, EM may converge to a local maximum of the observed likelihood function for some starting values instead of a global maximum so that EM estimators may not converge to MLE (Givens and Hoeting, 2012). Our EM estimator is a standard EM estimator, suffering from the (common) limitations of EM estimators while enjoying the (common) benefits and advantages of EM estimators.

In logistic regression, both continuous predictors and categorical predictors can be included. Our simulation studies used both types of predictors. However, for categorical predictors, we only used binary predictors. A categorical predictor with K levels can lead to or amount to $K-1$ binary predictors, which will increase the number of predictors, i.e. p. As the number of levels K increases, the number of predictors, i.e. p, increases, which will make estimation performance become worse. Thus, a categorical predictor with multiple levels may decrease estimation performance of our estimators. Future studies can be on the influence of categorical predictors with more than two levels.

There are some assumptions in our model setup. Firstly, we only consider independent individual-level data, i.e. (X_i, Y_i), $i = 1, 2, \cdots, n$, in this article. In practice, individual-level observations can be correlated or dependent. Secondly, we only consider the situation of "grouping completely at random", which means that the grouping mechanism is completely random, and is not influenced by the values of X and Y. In practice, grouping may not be completely random such as the situation where individuals with similar values in X or Y are more likely to be grouped together. Further studies can be conducted for grouping not completely at random. Thirdly, only summation aggregation $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ is studied. Other aggregations, such as $Y_i = 1(\sum_{j=1}^{n_i} Y_{ij} > 0)$ used in group testing of infectious disease, are not studied in this article, since the group-testing problem with $Y_i = 1(\sum_{j=1}^{n_i} Y_{ij} > 0)$ for logistic regression has been well studied in bio-statistics and epidemiology.

Although the method is proposed for a logistic regression (logistic link function) to deal with binary response variable Y, other link functions can also be used to handle the binary

response. For example, the tobit regression which uses a probit link function can also be used to analyze individual-level X and aggregate-level Y. In addition, the current article is based on the binary response variable Y. A follow-up study to extend our method to handle responses with more than two levels are under development.

6. Conclusions

We proposed an EM estimator for logistic regression based on individual-level predictors (X) and aggregate-level response (Y). We conducted simulation studies to evaluate the performance of the EM estimator and compare it with estimators in the literature (individual-LR, naive-LR and MLE). We then conducted two real data-based studies, i.e. wine quality modeling and maternal health risk modeling, to illustrate the use of different estimators. Both the simulation studies and real data-based studies have shown the use of our EM estimator in conducting logistic regression based on individual-level X and aggregate-level Y. We think both categories of methods (MLE category of methods or EM category of methods) work and are necessary for the problem of logistic regression based on individual-level X and aggregate-level Y. Similar and slightly different performances are expected for estimators along the two categories of methods.

Declarations

Availability of code and data: R functions implementing EM algorithm as described in the manuscript are available in Github repository via the link
`https://github.com/zhengxu0459/EM-Algorithm-Logistics-Regression`.
All data used in the study are publicly available.
Funding: This study receives no funding.

References

Agresti, A., (2013). Categorical Data Analysis, Wiley Series in Probability and Statistics. *Wiley*, Hoboken, NJ, USA.

Ahmed, M., (2023). Maternal Health Risk. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DP5D.

Ahmed, M., Kashem, M. A., Rahman, M. and Khatun, S., (2020). Review and analysis of risk factor of maternal health in remote area using the internet of things (iot). URL https://api.semanticscholar.org/CorpusID:214577407.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., West, M., et al., (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7, (pp. 453–464):210.

Brooks, S. P., Morgan, B. J. T., (2018). Optimization Using Simulated Annealing. *Journal of the Royal Statistical Society Series D: The Statistician*, 44(2), pp. 241–257, 12. ISSN: 2515–7884.

Chen, X., Dempster, A. and Liu, J., (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, pp. 457–469.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J., (2009). Wine Quality. *UCI Machine Learning Repository*. DOI: https://doi.org/10.24432/C56S3T.

Fernandez, M., Williams, S., (2010). Closed-formexpression for the Poisson-binomial probability density function. *IEEE Trans. Aerosp. Electron. Syst.*, 46, pp. 803–817.

Fletcher, R., (1970). A new approach to variable metric algorithms. *Comput. J.*, 13, pp. 317–322.

Fletcher, R., Reeves, C., (1964). Function minimization by conjugate gradient. *Comput. J.*, 7, pp. 149–154.

Geamsakul W., Yoshida T., Ohara K., Motoda H., Yokoi H., and Takabayashi K., (2005). Constructing a decision tree for graph-structured data and its applications. *Fundamenta Informaticae*, 66(1–2), pp. 131–160.

Getoor, L., Mihalkova, L., (2011). Learning statistical models from relational data. *In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 1195–1198.

Givens, G., Hoeting, J., (2012). Computational Statistics, Wiley Series in Probability and Statistics. *Wiley*, Hoboken, NJ, USA.

Hastie, T., Tibshirani, R. and Friedman, J., (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics. *Springer*, Berlin, Germany. ISBN 9780387848846.

Henaff, M., Bruna, J. and LeCun, Y., (2015). Deep convolutional networks on graph-structured data. arXiv preprint arXiv:1506.05163.

Hilbe, J., (2009). Logistic Regression Models. Chapman & Hall/CRC Texts in Statistical Science. *CRC Press*, Boca Ration, Florida, USA. ISBN 9781420075779.

Hong, Y., (2013). On computing the distribution function for the Poisson binomial distribution. *Comput. Stat. Data Anal.*, 59, pp. 41–51.

Kalivas, J. H., (1992). Optimization using variations of simulated annealing. *Chemometrics and Intelligent Laboratory Systems*, 15(1), pp. 1–12. ISSN 0169-7439.

Lambora, A., Gupta, K. and Chopra, K., (2019). Genetic algorithm-a literature review. In 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon), pp. 380–384. *IEEE*.

McLachlan, G. J. and Krishnan, T., (2007). The EM algorithm and extensions. John Wiley & Sons, New York City, USA.

Mercer, T. R., Salit, M.. (2021). Testing at scale during the covid-19 pandemic. *Nature Reviews Genetics*, 22(7), pp. 415–426.

Nelder, J., Mead, R., (1965). A simplex method for function minimization. *Comput. J.*, 7, pp. 308–313.

Primo, D. M., Jacobsmeier, M. L. and Milyo, J., (2007). Estimating the impact of state policies and institutions with mixed-level data. *State Politics & Policy Quarterly*, 7(4), pp. 446–459.

Saramago, P., Sutton, A. J., Cooper, N. J. and Manca, A., (2012). Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in medicine*, 31(28), pp. 3516–3536.

Wang, Y., (1993). On the number of successes in independent trials. *Stat. Sin.*, 3, pp. 295–312.

Wei, G. C., Tanner, M. A., (1990). A Monte Carlo implementation of the em algorithmand the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411), pp. 699–704,

Xu, Z., (2023). Logistic regression based on individual-level predictors and aggregate-level responses. *Mathematics*, 11(3), p.746.

Zhai, Y., Liu, B., (2006). Structured data extraction from the web based on partial tree alignment. *IEEE Transactions on Knowledge and Data Engineering*, 18(12), pp. 1614–1628.

Appendix

A. Additional Real Data Study Using Data in Xu (2023)

We conducted additional real data study based on the same data as used in Xu (2023). The dataset is "Social-Network-Ads" data in Kaggle Machine Learning Forum (`https://www.kaggle.com`). The dataset is a categorical dataset to determine whether a user purchases a particular product. It contains 400 observations. Two predictors are age and salary, after data standardization. The same as in Xu (2023), we impose data aggregation on this dataset with the group size equal to 3, 5 and 7. We conducted (1) infeasible individual-level logistic regression, (2) naive logistic regression, (3) MLE estimator in Xu (2023), and (4) our proposed EM estimator in this manuscript. Because true parameter values are unknown, we illustrate the use of different estimators and report estimated values using different estimators in Table 6.

Table 6: Estimated Values of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Social Network Ads Data.

	$n_i = 3$				$n_i = 5$				$n_i = 7$			
Var	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
β_0	-1.14	-0.71	-1.17	-1.17	-1.14	-0.68	-1.25	-1.24	-1.13	-0.64	-1.27	-1.27
β_1	2.45	1.67	2.53	2.53	2.45	1.61	2.79	2.79	2.44	1.59	2.97	2.97
β_2	1.22	0.79	1.47	1.47	1.22	0.64	1.54	1.54	1.22	0.53	1.26	1.26

B. Additional Simulation Study Using Xu (2023)'s Setup

We conducted additional simulation study using Xu (203)'s simulation setup as follows. In each scenario, simulations were conducted with sample sizes ($K = 300, 500, 100$), equal group sizes ($n_g = 7, 30$), and different parameter values. Data were generated as follows:

- In Scenario 1, $X_{i1} \sim N(0,1)$, $X_i = (1, X_{i1})^T$, $Y_i \sim$ Bernoulli$(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (1,-2)^T$ (Scenario 1A) or $(1,3)$ (Scenario 1B).
- In Scenario 2, $X_{i1} \sim N(0,1)$, $X_{i2} \sim t(df = 5)$, $X_i = (1, X_{i1}, X_{i2})^T$, $Y_i \sim$ Bernoulli$(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-1,1,2)^T$ (Scenario 2A) or $(0,-2,1)$ (Scenario 2B).
- In Scenario 3, $(X_{i1}, X_{i2}) \sim$ BivariateNormal$(0,2,1,4,\rho = 0.5)$, $X_{i3} \sim$ Cauchy$(0,1)$, $X_i = (1, X_{i1}, X_{i2}, X_{i3})^T$, $Y_i \sim$ Bernoulli$(e^{X_i^T\beta}/(1+e^{X_i^T\beta}))$, $\beta = (-1,1,0,-1)^T$ (Scenario 3A) or $(0,-2,1,1)$ (Scenario 3B).

We reported squared bias and variance of four estimators (E1: Individual-LR, E2: Naive-LR, E3: MLE and E4: EM) in Table 7. We reported MSE and MAD of the four estimators in Table 8. Results obtained from additional simulation studies confirm our findings based on simulation studies. The same findings were obtained.

Table 7: Average Squared Bias and Variance of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Simulation Setup in Xu (2023). In the columns for average squared bias and average variance, the unit is 0.001.

			Average Squared Bias				**Average Variance**			
Scenario	**M**	n_i	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1A	300	7	0.28	310.34	0.76	0.86	7.69	6.84	23.74	22.58
1A	300	30	0.09	343.26	0.53	0.82	1.43	6.81	30.52	24.37
1A	500	7	0.02	317.16	0.00	0.00	3.56	3.49	11.45	10.60
1A	500	30	0.01	356.36	0.01	0.07	0.91	2.70	11.68	11.33
1A	1000	7	0.08	302.01	0.20	0.26	2.14	1.83	6.90	6.99
1A	1000	30	0.01	352.87	0.24	0.22	0.42	1.76	6.29	6.21
1B	300	7	0.01	1256.06	0.36	0.54	10.90	6.91	37.38	35.02
1B	300	30	0.17	1376.61	1.90	3.24	2.96	5.12	38.64	30.50
1B	500	7	0.16	1264.86	0.35	0.51	6.86	3.30	16.90	16.43
1B	500	30	0.00	1401.36	0.00	0.02	1.58	2.32	23.51	19.81
1B	1000	7	0.02	1267.42	0.05	0.07	3.00	1.74	10.07	10.04
1B	1000	30	0.00	1400.68	0.37	0.07	0.69	1.23	13.19	6.68
2A	300	7	0.00	485.55	0.12	0.14	6.09	6.33	17.56	17.33
2A	300	30	0.02	547.58	0.32	0.23	1.37	4.72	27.61	26.45
2A	500	7	0.09	487.78	0.05	0.08	4.19	3.95	12.23	12.35
2A	500	30	0.01	540.78	0.07	0.13	0.89	3.48	13.58	11.83
2A	1000	7	0.04	484.78	0.04	0.04	1.86	1.70	6.49	6.33
2A	1000	30	0.00	540.90	0.02	0.00	0.47	1.65	7.32	6.88
2B	300	7	0.04	304.21	0.35	0.40	5.37	5.82	17.33	17.08
2B	300	30	0.03	339.23	0.35	0.39	1.25	4.09	18.77	18.82
2B	500	7	0.04	304.27	0.11	0.18	3.21	3.18	10.89	10.67
2B	500	30	0.00	334.51	0.16	0.23	0.80	2.48	12.16	11.68
2B	1000	7	0.06	304.24	0.06	0.09	1.62	1.44	5.31	4.99
2B	1000	30	0.00	333.05	0.12	0.16	0.37	1.38	6.51	6.09
3A	300	7	0.06	336.15	0.55	0.58	4.46	6.19	13.40	12.67
3A	300	30	0.02	336.61	0.60	0.91	0.83	6.34	14.10	13.88
3A	500	7	0.03	342.89	0.12	0.08	2.04	3.22	7.70	7.67
3A	500	30	0.00	345.45	0.74	0.55	0.64	3.39	8.91	7.21
3A	1000	7	0.02	343.02	0.27	0.21	1.24	2.40	4.98	4.38
3A	1000	30	0.00	350.84	0.18	0.01	0.27	1.75	4.83	3.26
3B	300	7	0.27	587.43	0.48	0.62	6.67	4.84	17.43	16.45
3B	300	30	0.04	605.00	0.27	1.11	1.24	3.65	17.57	13.91
3B	500	7	0.10	588.20	0.15	0.18	3.12	2.85	11.54	8.89
3B	500	30	0.01	611.21	0.12	0.36	0.67	2.20	17.26	13.13
3B	1000	7	0.01	592.02	0.03	0.14	1.67	1.46	6.01	4.82
3B	1000	30	0.01	615.74	0.28	0.06	0.33	1.22	6.56	4.26

Table 8: Average Mean Squared Error (MSE) and Average Mean Absolute Deviation (MAD) of Estimator E1 (Individual-LR), E2 (Naive-LR), E3 (MLE) and E4 (EM) Based On Simulation Setup in Xu (2023). In the columns for average MSE and average MAD, the unit is 0.001.

			Average MSE				Average MAD			
Scenario	**M**	n_i	**E1**	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**
1A	300	7	7.98	317.18	24.50	23.44	67.78	529.08	116.88	113.94
1A	300	30	1.53	350.07	31.05	25.19	30.39	558.73	118.68	107.82
1A	500	7	3.58	320.65	11.45	10.60	48.31	534.84	80.78	77.99
1A	500	30	0.92	359.06	11.69	11.40	24.30	567.55	76.76	76.65
1A	1000	7	2.22	303.84	7.10	7.25	35.76	523.73	60.91	61.97
1A	1000	30	0.43	354.63	6.53	6.42	16.33	565.64	56.66	55.27
1B	300	7	10.91	1262.97	37.74	35.56	79.25	1003.77	138.60	135.89
1B	300	30	3.13	1381.73	40.54	33.74	42.20	1051.67	139.06	129.85
1B	500	7	7.03	1268.16	17.25	16.95	66.24	1006.76	97.60	96.12
1B	500	30	1.58	1403.67	23.52	19.84	30.15	1059.46	106.15	99.14
1B	1000	7	3.02	1269.16	10.12	10.11	43.39	1007.77	74.67	73.13
1B	1000	30	0.69	1401.91	13.55	6.75	19.30	1059.49	73.50	59.47
2A	300	7	6.09	491.88	17.68	17.47	62.82	634.87	102.26	101.07
2A	300	30	1.38	552.30	27.93	26.68	29.40	675.59	121.24	117.04
2A	500	7	4.28	491.72	12.27	12.43	51.44	635.26	83.97	83.86
2A	500	30	0.90	544.26	13.65	11.96	23.51	673.15	87.50	82.35
2A	1000	7	1.90	486.48	6.53	6.37	34.90	636.25	62.12	61.98
2A	1000	30	0.47	542.56	7.35	6.89	17.05	671.85	64.59	62.78
2B	300	7	5.40	310.03	17.68	17.48	58.39	444.51	101.07	100.09
2B	300	30	1.28	343.31	19.11	19.21	27.81	460.20	93.87	93.62
2B	500	7	3.26	307.45	11.00	10.85	45.17	441.17	75.70	75.21
2B	500	30	0.80	336.99	12.32	11.91	22.86	458.86	78.42	77.51
2B	1000	7	1.68	305.68	5.37	5.07	31.61	437.72	55.62	53.87
2B	1000	30	0.37	334.43	6.63	6.25	14.96	455.67	56.98	55.07
3A	300	7	4.52	342.34	13.95	13.26	51.25	475.14	89.68	87.77
3A	300	30	0.85	342.95	14.70	14.79	22.81	471.88	90.96	91.01
3A	500	7	2.06	346.11	7.81	7.75	36.19	478.33	66.02	66.15
3A	500	30	0.64	348.84	9.65	7.76	19.24	479.12	72.57	66.37
3A	1000	7	1.27	345.42	5.25	4.59	26.67	474.11	53.40	51.13
3A	1000	30	0.28	352.59	5.00	3.27	12.73	482.34	50.64	43.76
3B	300	7	6.94	592.27	17.91	17.07	62.77	648.17	98.15	96.75
3B	300	30	1.28	608.65	17.84	15.02	27.56	653.18	93.76	88.13
3B	500	7	3.23	591.06	11.70	9.07	41.68	646.40	78.30	70.34
3B	500	30	0.68	613.41	17.38	13.48	20.27	656.09	86.77	79.63
3B	1000	7	1.68	593.48	6.04	4.97	31.43	646.77	55.45	51.70
3B	1000	30	0.34	616.96	6.84	4.32	13.78	657.31	54.93	45.06

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 29–49, https://doi.org/10.59139/stattrans-2025-003
Received – 31.07.2024; accepted – 23.12.2024

Selection criteria and targeting the poor for poverty reduction: the case of social safety nets in Sri Lanka

Diana Dilshanie Deepawansa[1], **Priyanga Dunusinghe**[2]

Abstract

Reducing poverty and improving the living standards of the poor and vulnerable populations in Sri Lanka have been one of the country's key goals. The government has designed poverty-targeting programs with relevant government agencies working to support low-income families. The programs include cash transfers, microfinancing and various community-based and livelihood development activities, including the "Aswasuma" program, which is the primary safety net initiative. Although safety net programs have been receiving significant financial support for decades, many people still remain excluded as a result of mistargeting, lack of transparency and poor beneficiary selection methods. To address these challenges, the selection criteria have to be redesigned to effectively target poverty. This article explores the Multidimensional Deprivation Score Test (MDST), which assesses the multiple dimensions of household deprivation by weighting each deprivation through a data-driven approach. This methodology aims to identify the poorest and most vulnerable people more accurately. Using the data collected during the 2019 Household Income and Expenditure Survey, conducted by the Department of Census and Statistics, the MDST has improved targeting accuracy and thus the impact of social protection programs. It is therefore crucial to increase the efficiency of data collection and to compile the weighted deprivation score. Moreover, incorporating a community-level evaluation and regular monitoring is essential for maximizing the accuracy and effectiveness of targeting poverty.

Key words: poverty, social safety net, selection criteria.

1. Introduction

Under successive governments, Sri Lanka has initiated much effort to ensure sustainable and viable economic development since independence. Consequently, Sri Lanka had witnessed mixed results prior to the COVID-19 pandemic and the subsequent economic crisis. Sri Lanka's economy recorded an 8.7 per cent GDP growth

[1] Department of Census and Statistics, Sri Lanka. E-mail: dilhaniebg@gmail.com. ORCID: https://orcid.org/0000-0002-2470-4390.

[2] Department of Economics, University of Colombo, Sir Lanka. E-mail: dunusinghe@econ.cmb.ac.lk. ORCID: https://orcid.org/0000-0002-2225-711X.

rate in 2011 and the country's per capita income reached US$ 4,293 in 2017 though witnessed some setbacks in subsequent years (DCS,2023a). With the expansion of economic activities, the unemployment rate hovered around 5 per cent during the last decade. In addition, despite several global and domestic challenges, inflation had been retained at a single digit for four consecutive years as measured by year to year at Colombo Consumer Price Index (CCPI) (DCS, 2023b). Moreover, the poverty headcount index decreased dramatically from 46.8 per cent to 14.3 per cent from 2002 to 2019 (DCS, 2021a).

In the aftermath of the COVID-19 and economic crisis in 2022, Sri Lanka's economic outlook turned uncertain due to unsustainable public debt and a severe balance of payment crisis. Hence, the economy contracted by 11.7 percent year-on-year in the third quarter of 2022 (DCS, 2023a) and CCPI year-to-year. Inflation reached two digits from December 2021 and an unprecedented 69.8 per cent in September 2022 due to high food inflation of 94.9 per cent. Subsequently, it decreased to 54.2 per cent in January 2023 (DCS, 2023b).

Due to economic imbalance described above, many people in the country face severe economic hardships exacerbating vulnerability and increasing number of people live in poverty. Understanding the past economic experiences in the country it is crucial to shape the policies to overcome the existing challenges and adapt to economic dynamics effectively. Reducing poverty and improving the living standard of the poor population in Sri Lanka has been a critical agenda of the government. Hence the incumbent government has also designed and accelerated poverty-targeting programs to reduce poverty to increase the living standard of poor people. Successive governments have implemented Social Protection policies and programs since the 1940s, such as universal free education and health and food subsidy programs (Ganga & Sahan, 2015). Currently, there are many fragmented social protection schemes in the country. Ministry of Social Empowerment and Welfare (MoSEW) plays a significant role in identifying low-income families and supporting them in numerous ways to lift their living standards and achieve sustainable development by providing them cash transfers, microfinance, and various community-based and livelihood development activities. The primary safety net program currently targeting the poor in terms of Sri Lanka is the "Samurdhi/Aswasuma" program launched under the Department of Samurdhi Development/Welfare Benefit Bord (WBB). The schemes mainly cover disability, old age, and Chronic Kidney Disease of Unknown Etiology (CKDU). Besides, there are schemes covering health care, school food programs, maternal programs and other social safety net programs targeting the poor and social security schemes, old age pensions, and lump-sum payments at the retirement of government and non-government workers.

The social protection floor system is one of the main policy instruments in developing countries to target the poor to reduce chronic poverty and protect vulnerable people. One of the main targets of global and local development agendas is reduced

poverty (Goal one of MDG and SDG). The development of the human capital of the poor through social safety net programs is a long-lasting solution to poverty. Social protection covers social assistance, social security, social care, and labor market inclusion and productive employment. Developing countries have recently increased social protection coverage by expanding their social protection systems. Due to the COVID-19 impact and the economic crisis, the Sri Lanka economy has faced a sizable economic recession. Many people and households hit by the crisis face the hardships of their livelihoods. This situation further increases the focus on social protection programs to protect impoverished and vulnerable individuals and families coping with generated fiscal shock and economic crises.

During the post-independence, successive governments in Sri Lanka implemented several social protection programs such as Janasaviya, Samurdhi and the food subsidy programs, investing yet more resources. However, the outcome has not been commensurate with such investments, and none achieved its desired target (Samaraweera, 2010). The previous social protection programs have reported high inclusive and exclusive errors. Specifically, these programs have not effectively targeted their intended beneficiaries resulting in both inclusion of individuals who are not eligible with criteria (inclusive errors) and exclusion of individuals who are eligible (exclusion errors). These discrepancies challenge the effectiveness and quality of the social protection programs. Hence, it is very crucial to address these issues to improving the accuracy and impact of social protection programs. Therefore, this research investigates these errors and proposes to enhance the accuracy of social protection programs.

According to the Household Income and Expenditure Survey (HIES) 2019 conducted by the Department of Census and Statistics (DCS), out of 13 main social protection programs, currently, 33.8 per cent of poor people are not covered (Under-coverage), and 70.6 per cent of non-poor people has received transfers (leakage). Hence, the impact of social protection spending to reduce poverty has not achieved the desired results. This is due to weak targeting in which the welfare benefit has not always benefitted the needy. Thus, social protection programs have limited impact on poverty (DCS, 2021a). An early study has been carried out by the World Bank for Sri Lanka using the data from the Sri Lanka Integrated Survey (SLIS), conducted by the World Bank in collaboration with local institutions in 1999–2000, using a Proxy Means Test (PMT). However, the targeting accuracy was not as expected (Narayan & Yoshida, 2005). Kidd and Wylde (2011) studied the regression accuracy of PMT for Bangladesh, Indonesia, Rwanda, and Sri Lanka and found that high in-built inclusion and exclusion errors were high. This study has developed a criterion that enhances effectiveness of targeting which is essential in minimizing the exclusion and inclusion errors in poverty reduction programs.

This paper is structured as follows. Section 2 presents the literature review presenting different methods used as beneficiary selection criteria for targeting the

poor. Section 3 describes the methodologies employed to assess the selection criteria and new method used to compute Multidimensional Deprivation Score for identifying the new target group. Section 4 presents results and output. Finally, Section 5 concludes the paper with discussion and some recommendations.

2. Literature review

Effective targeting increases the impact of poverty reduction and raises the standard of living of the poor. Different countries have different selection criteria for identifying the poor people for targeting (Kidd &Wylde, 2011; Alatas et al., 2012; Alkire & Seth, 2013; Brown, Ravallion, & Van de Walle, 2018; Sabates-Wheeler, Hurrell, & Devereux, 2015; Diamond, et al., 2016; Bird & Hanedar, 2023). The social safety net programs have promotion and protection effects (Devereux, et al., 2017). Morestin, Grant & Ridde (2009) did a systematic review of literature on selection criteria presenting 68 experiences used by developing countries, of which 27 were in sun-Sahara Africa. This study has identified 30 incidents of the identification of the poor based on administrative, community-based, and mixed processes.

Poverty is a multidimensional phenomenon. Amartya Sen's capability concept significantly contributed to the development of multifaceted poverty measures of understanding poverty after his seminal work (Sen, 1983,1995,1997). People are poor in terms of income, and many other aspects, such as health, education, shelter, inadequate sanitation facilities, social exclusion, access to essential services and lack of assets (Sabina ,2023; Sabina, et al., 2015). Morestin, Grant & Ridde (2009) found 260 selection criteria based on 68 surveyed and categorized those into 11 dimensions. The eleven dimensions are: 1) Possession of goods and means of production; 2) Household compositions; 3) Income; 4) Condition of dwelling; 5) Occupational status; 6) Food security; 7) State of health; 8) Education; 9) Access to essential services and to credit; 10) Expenses; and 11) Physical appearance and clothing. Further, this study identified that in administrative processes, in 48 per cent of experiences, the program manager was responsible for determining the poor. In the community process, 36 per cent of studied community members have identified the poor. In the mixed method, in 20 per cent of surveys, the first selection was made by the program manager decided the final beneficiaries. Based on the study review Morestin, F., Grant & Ridde (2009) conclude that there are no perfect criteria for selecting beneficiaries and that developing countries should pay more attention to implementing an effective process for choosing beneficiaries. The effectiveness is based on inclusive and exclusive error of the selection criteria.

The Proxy Mean Test (PMT) is a widely used method to select the poor for targeting. This method is based on a score produced from a set of coefficients of variables reflecting the household living condition chosen for the best regression model (WB, 1999). This method commonly targets the poor for social safety net programs

when income or consumption expenditure data are unavailable. The early contribution of the PMT method for selection criteria was made by Grosh (1994) for Latin America. He concluded that this method produces the best targeting outcomes reducing inclusion and exclusion errors. Proxy Mean Test (PMT) model is based on a statistical method used to estimate income or expenditure based on observable characteristics correlated with income or consumption expenditure. This method is based on national household surveys. The term "Proxy Mean Test" describes estimating income or consumption when precise measures are unavailable or difficult to obtain. Brown, Ravallion and Van de Walle (2018) state that "Proxy-means testing is a popular method of poverty targeting with imperfect information". The methodology estimates household income or expenditure by associating indicators or proxies. They include demographic characteristics (such as the age of household members and size of household), human capital characteristics (such as education of household head and enrolment of children in school), physical housing characteristics (such as type of roof or floor), durable goods (such as refrigerators, televisions, or cars) and productive assets (such as land or animals), etc. It uses the weights for the variable derived through statistical analysis of household survey data like Household Income and Expenditure Survey. Using the agreed weights, a score is calculated for each household. Households that score below the cut-off point are eligible for social protection programs.

Narayan and Yoshida (2005) applied the PMT method for Sri Lanka using household data from the Sri Lanka Integrated Survey (SLIS) conducted by the World Bank in collaboration with local institutions in 1999–2000[3]. In this exercise, seven main models were developed, and different cut-offs based on per capita consumption were applied for the selection. Further, considering several modifications, four additional models, Model 8, Model 9, Model 10, and Model 1, were developed based on Model 7. The model shows that the inclusion and exclusion errors were high. For example, the under-coverage rate varies from 50 per cent to 55 per cent. The leakage rate varies from 39 per cent to 40 per cent based on a 25 per cent cut-off, and at the 40 per cent cut-off, coverage ranges from 20 per cent to 31 per cent, and leakage varies from 31 per cent to 35 per cent based on the selected models 7, 10 and 11.

Proxy Means Test has become a popular method with many advocates and detractors. The Australian Agency for International Development (AusAID) supports evidence-based debates to investigate the PMT's strengths and weaknesses further. This study assesses the regression accuracy of the PMT model in Bangladesh, Indonesia, Rwanda, and Sri Lanka, which was done in this exercise earlier and found that inclusion and exclusion vary between 44 per cent and 55 per cent with the coverage of 20 per cent of the population and 57 to 71 per cent when 10 per cent were covered (Kidd & Wylde, 2011). In addition to non-sampling errors of the dependent survey's accuracy of PMT partially depend on the interaction with error arising from the regression with the

[3] The survey data was excluded for the analysis for Northern and Eastern provinces due to conflict and concern with the quality of the data.

correlation of proxies and consumption expenditures. According to the finding of the new PMT test done by the WB based on the currently conducted survey and the assessment made by Kidd and Wylde (2011). The Australian Agency for International Development based on the PMT test done by WB for Sri Lanka in 2003 evident that PMT regression-based method is inaccurate for targeting and the majority of eligible poor households may be permanently excluded from the social grant scheme from the results from PMT scoring. Further, capturing the dynamic changes of a focus unit family/household or individual is impossible. However, it can be updated after doing a large-scale household survey frequently maintaining the integrity of the specifications.

3. Methodology

Table 1 presents the targeting accuracy of the selection method. It can be evaluated through the Type I and Type II errors, which indicate the share of under-coverage[4] and leakage[5], respectively. Type I error shows the number of individuals incorrectly excluded (exclusion error). Type II error (inclusion error) indicates the individuals incorrectly identified as eligible by the selection criteria as a share of the total population. When increased the under-coverage reduces the impact of the program and does not affect the cost of the welfare budget; however, leakage does not affect the program's impact but unnecessarily increases the cost of the welfare budget.

Table 1: Illustration of Type I and Type II errors

Type	Target group	Non-target group	Total
Eligible: predicted	Targeting Success (S1)	Type II Error (e2)	m_1
Ineligible predicted	Type I Error (e1)	Targeting Success (S2)	m_2
-	n1	n2	n

Those in the bottom quintile of per capita expenditure or poor constitute the "target group", while those predicted and grouped by eligibility threshold constitute the "eligible" group. The individual correctly classified as eligible by the formula that belongs to the target group (bottom per capita expenditure quintile or poor) is "Targeting Success". A person who is incorrectly excluded by the procedure is a case of Type I error. Conversely, a person incorrectly identified as eligible constitutes a Type II error; under-coverage is calculated by dividing the number of cases of Type I error by the total number of individuals who should get benefits [e1/n1]. Leakage is calculated by dividing the number in the Type II error category by the number of persons classified as eligible by the formula [e2/m1].

[4] Under-coverage is the percent of poor individuals that do not receive the social transfer.

[5] Leakage is the percent of individuals that receive social transfer and are not poor.

Effectiveness is the capacity to identify the actual beneficiaries or the "real" poor. Conversely, two types of errors are possible: excluding poor individuals and including persons who are not poor as beneficiaries. Therefore, it is more desirable to reduce both under-coverage and leakage for effective targeting. The efficiency of the selection criteria can be evaluated through the magnitude of Type I and Type II error. Arguably in a climate of "no method is perfect", it is essential to minimize these two errors as much as possible.

Existing beneficiary selection procedure in Sri Lanka

This will review the present main social safety net program in Sri Lanka, "Samurdhi[6]. The beneficiaries of the Samurdhi program are currently selected based on self-reported income level. However, that method generates high inclusion and exclusion errors. The 2019 Household Income and Expenditure Survey data shows that Samurdhi covered only 42 per cent (direct and indirect beneficiaries) of the total poor population, which under-coverage is 58 per cent, and leakage is 62 per cent. Among the leakage, 29 per cent are in the second quintile[7], 18.7 per cent are in the third quintile, 9.7 per cent are in the fourth quintile, and 4.5 are in the richest fifth quintile (top 20 per cent). In other words, of the non-poor population, 15.7 per cent are receiving Samurdhi benefits. To mitigate this issue, this study introduced a new criterion for identifying beneficiary's potential beneficiaries more efficiently through a criterion for effective target beneficiaries and assessing the deprivations at the family level in multidimensional aspects called "Multidimensional Deprivation Score Test (MDST)".[8] The following section presents the method of MDST.

Multidimensional Deprivation Score Test (MDST)

Multidimensional Deprivation Score Test (MDST) assesses the living standard of the poor in terms of multiple aspects, reflecting the deprivation at the family level. However, this research used the data from Household Income and expenditure survey conducted by DCS in 2019 and has collected information at the household level. Hence, this analysis considered the dimensions: Education, Health, Economic Level, Assets and Housing characteristics and Family Demography by household level.[9] Under these dimensions 22 indicators were considered (Appendix). These dimensions and indicators were selected normatively. This method proposed a data-driven weight function in which the frequency of the 'definitely poor' phenomenon weights each dimension. This weight function is built to assign lower weight to the extent in which

[6] In Household Income and Expenditure Survey in 2019 capture the Samurdhi.

[7] Based on per capita consumption expenditure.

[8] This method was introduced by DCS to the WBB (see gazette in No. 2302/23 - Thursday, October 20, 2022).

[9] This approach recommended to apply to the survey data collected from vulnerable and poor people to develop an index in computing a deprivation score for each family. Usually, social protection benefits are given to the poor and vulnerable families or individuals rather than households.

lower frequency of families is 'definitely poor', and higher significance to families with higher frequency of 'definitely poor' in a dimension. This weight can be introduced as an attempt to achieve Sustainable Development Goals (SDGs) with current information in the concept of no one behind all its form everywhere.

For example, an indicator of having safe drinking water, in an area called A, most households need access to a safe source of drinking water. Thus, definitely, the poor frequency for that indicator would be very high. Therefore, assigning a very high weight to that indicator is reasonable. In area B, the frequency of access to safe drinking water could be higher. Then a low weight was given to that indicator for that area. Each household's deprivation score is constructed based on a weighted average of the deprivations, and each household is identified as deprived or non-deprived based on a deprivation cut-off. If a household's weighted deprivation score is above the cut-off that household should be considered eligible for the social protection program.

Computation of Multidimensional Deprivation Score

MDST develops an index called the Multidimensional Deprivation Score (MDS) at the unit of the analysis. In this research, the unit of analysis is a household. This score is between 0 to 100, 0 indicates completely not deprived, and 100 means completely deprived.

Calculation of the deprivation score for a household is done in three steps:

a. Set of indicator deprivation
b. Computation of weight for indicators
c. Calculation of weighted deprivation score for each household

Every indicator is assigned a deprivation cut-off, and if a household is deprived in the relevant indicator, then it is considered completely deprived and assigned 1 for that indicator and otherwise 0. Accordingly, every indicator is assigned one and zero.

Indicator deprivation

Deprivation cut-off for each and every indicator was assigned as given in Appendix. If the deprivation cut-off is denoted as z_j then the household is considered deprived if the ith family/household achievement of indicator x_j is below the cut-off $(x_j < z_j)$.

If ith household owns indicator j, then its indicator deprivation can be calculated using the following equation:

$x_j(i)$ is the household value on indicator *j*.

Then

$\mu_j(i) = 1$; if household deprived in indicator *j*,

$\mu_j(i) = 0$; if household is not deprived in indicator *j*.

The formula for the weight function

This method uses the frequency-based data-driven weight function to weight the indicators considering the number of completely deprived household for each indicator in the area of interest (e.g. district or any administrative or geographical level). The steps for calculating the indicator weight are given below:

- Count the sum of the number of deprived households in every indicator in the area of interest.
- Get the natural log value of the inverse of the sum of the number of deprived households in every indicator in the area of interest.
- Get the total sum of natural log values obtained for every indicator for the area of interest.
- Finally, get the ratio of the natural log values to the total sum of natural log values (normalize the weight).

Getting this natural log of the inverse of deprived frequency is smoothing out the weight and reducing the over-dispersion of values. This weight function is built to assign lower weight to the indicator in which many households turn out to be 'definitely poor', and higher weight to households with a high frequency of 'definitely poor' in an indicator. The mathematical formula is given below:

$$\omega_j = \frac{ln\frac{1}{f_j}}{\sum_{j=1}^{k} ln\frac{1}{f_j}} \times 100\ ;\ j = 1,2, \dots \dots k \tag{1}$$

where f_j denotes the frequency of households completely deprived in the j[th] indicator and ω_j is the weight for the j^{th} indicator. Lower weights mean the criterion many households are less deprived of; lower weights indicate lower importance. Higher weights mean a high frequency of deprived households' in a indicator that households highly belong to deprivation of that indicator. Higher weights indicate greater importance.

Calculation of weighted deprivation score for individual

$$\mu_{wi} = \sum_{j=1}^{k} \omega_j \times \mu_j \tag{2}$$

Where μ_{wi} is the weighted deprivation score for ith households. The weighted deprivation score gets values between 0 and 100, in which zero (0) is not deprived, and one (100) is completely deprived.

4. Results

The data used for this study is the Household Income and Expenditure Survey (HIES) conducted in 2019 by the Department of Census and Statistics. The sample of this survey was drawn scientifically to represent the entire country's population. It was conducted throughout the year to capture the seasonal variation of the living standard of the household population in Sri Lanka. Two-stage stratified sampling method was used to draw the survey sample, and the sample size was 25,000 housing units in Sri Lanka. This survey collects information on household income and consumption expenditure and details on living standards and selected main social welfare programs. The Official Poverty Line (OPL) of Sri Lanka is computed based on consumption expenditure collected from this survey (DCS, 2021a).

The HIES, which was conducted in 2019, revealed that of the total population in Sri Lanka, 14.3 per cent (3.04 million individuals) live in poverty based on Official Poverty Line while from the total households, 11.9 per cent (681,800 households) live in poverty (DCS, 2021a). The Survey found that approximately out of every six (16 per cent) people are multidimensionally poor (DCS, 2021b). Further, it shows that 6.2 per cent of people has been lifted out of monetary poverty due to the thirteen social protection programs including the Samurdhi program considered in this survey. Table 1 shows the coverage of the population by the Samurdhi program by per capita expenditure decile, and Table 2 shows the distribution of direct and indirect Samurdhi beneficiaries by real per capita expenditure decile. That is the proportion of direct and indirect Samurdhi beneficiaries in each decile group. The total coverage of Samurdhi is 20.6 per cent of the total population. Both Tables 2 & 3 indicated the inefficient targeting of the Samurdhi program shows that the beneficiaries are also in the richest top two deciles. It is evident that among all beneficiaries, 4.6 per cent are in the top 20 per cent.

Table 2: Coverage of the Samurdhi program by real per capita expenditure decile

Per capita expenditure decile	Coverage of Samurdhi (Per cent)
Sri Lanka	20.6
1	40.5
2	34.6
3	32.7
4	28.0
5	21.1
6	18.2
7	12.1
8	9.6
9	7.0
10	2.4

Table 3: Distribution of beneficiaries by real per capita expenditure decile

Per capita expenditure decile	Proportion of beneficiaries (%)
Sri Lanka	100.0
1	19.6
2	16.8
3	15.8
4	13.6
5	10.2
6	8.8
7	5.8
8	4.7
9	3.4
10	1.2

The Table 4 presents the estimated number of people who are correctly and incorrectly classified as direct and indirect Samurdhi beneficiaries within the poorest 20 percent of the population based on real per capita expenditure quintile. The finding reveals that the exclusion error (under-coverage) is 62.5 and inclusion error (leakage) is 63.6 percent.

Table 4: Distribution of eligible and ineligible Samurdhi beneficiaries by target and non-target group

Specification	Target group (Q1)	Non-target group	Total
Eligible:	1,595,043 (S1)	2,786,943 (e2)	4,381,986 (m1)
Ineligible	2,654,626 (e1)	14,211,736 (S2)	16,866,362 (m2)
Total	4,249,669 (n1)	16,998,679 (n2)	21,248,348

The Target group is the individual who is in the bottom real per capita expenditure quintile (Q1)

Direct and indirect Samurdhi beneficiaries

Under-coverage[10] = [e1/n1] = 62.5 per cent

Leakage[11] = [e2/m1] = 63.6 per cent

[10] Under-coverage is the percent of poor individuals that do not receive transfer.

[11] Leakage is the percent of individuals that receive transfer and are not poor.

Comparing the coverage and distribution of beneficiaries by different approaches

Figure 1. presents the coverage of three types of targeting approaches for poor individuals and actual Samurdhi beneficiaries by per capita income deciles to examine their effectiveness for better targeting. MDSQ5 represents the individuals in the poorest 20 per cent based on Multidimensional Deprivation Score Test. OPL_new means the individuals who live in poverty based on the official monetary poverty line. MPI_poor means the individuals who are multidimensionally poor on official multidimensional poverty index based on the Alkire and Foster method. Finally, Samurdhi represents the actual direct and indirect beneficiaries currently receiving benefits, while looking at the output reveals that among the poorest 40 per cent (bottom four deciles), the highest number of poor individuals are covered by the individuals identified by the MDST method.

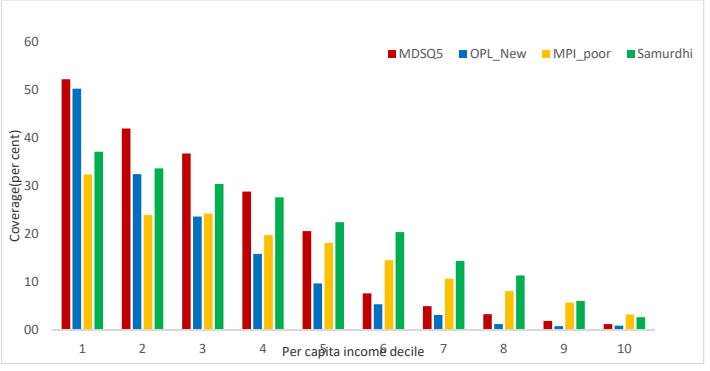


Figure 1. Distribution of predicted, targeted and Samurdhi beneficiaries by per capita income decile

There are vast discrepancies in coverage of actual direct and indirect Samurdhi beneficiaries and targeted individuals across districts (Figure 2). According to the official multidimensional poverty index the Colombo district (3.5 per cent) has the lowest incidence of poverty. In comparison, Nuwara Eliya (44.2 per cent) shows the highest poverty (DCS, 2021b). However, based on official monetary poverty based on consumption expenditure, the lowest poverty incidence was reported from Colombo (2.3 per cent), while the highest was from Mullaitivu (44.5 per cent) (DCS, 2021a). When examining the Nuwara Eliya district, more than half of the individuals are poor on MDS, more than two-fifths are poor on MPI, and more than one-fourth are poor on OPL, but coverage of Samurdhi is 10 per cent. The situation is different in Mullaitivu; two-fifths are poor in terms of OPL, three-tenth and more than one-tenth are poor in terms of MDST and MPI, and the Samurdhi coverage is almost 50 per cent, while in the Mannar district, the coverage of the Samurdhi is much higher than the share of the targeted beneficiaries. These findings demonstrate that the existing beneficiary selection method for the leading social net program in Sri Lanka should be revised for effective targeting.

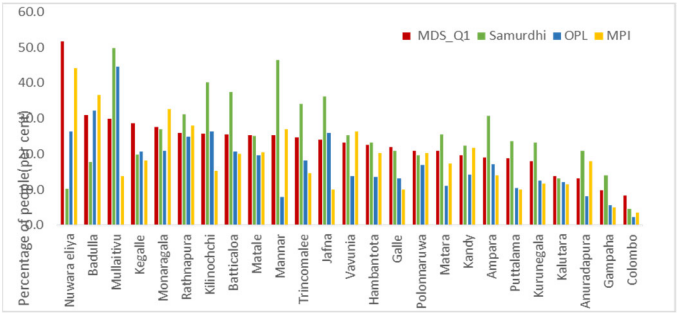


Figure 2: Distribution of predicted, targeted and Samurdhi beneficiaries by district

Figure 3 presents the graphical presentation of the distribution of the Multidimensional Deprivation Score. It appears as the normal distribution and has no skewness.

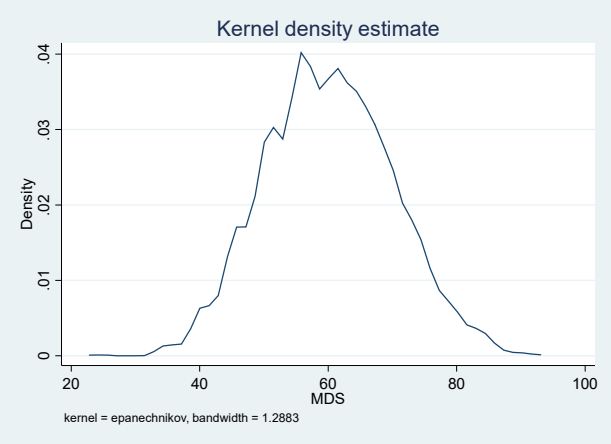


Figure 3: Distribution of Multidimensional Deprivation Score (MDS)

Selection cut-off

It is essential to identify the most appropriate cut-off for selection of beneficiaries for welfare programs. For this purpose, it is necessary to decide the targeting group either in monetary, non-monitory or mixed approach or to decide normatively on policy decisions. For instance, it can be per capita income or consumption expenditure decile or quintile or multidimensional deprivation quintile or decile. The coverage of the target population is very high; then the selection cut-off is more accurate with less under-coverage. Figure 4 plots the percentage of deprived people based on multidimensional deprivation scores by different cut-offs concerning the per capita expenditure quintiles. The graph shows that the MDST cut-off concerning the AA' line covers 100 per cent of the bottom 20 per cent of the poor individual (first per capita expenditure quintile). The exclusion error is very low, and the cut-off on the BB' line shows that the richest top 20 per cent is excluded 100 per cent, and the inclusion error is significantly less. Further, it reveals that when increase the cut-off exclusion error is

reduced. Accordingly, the plot provides valuable information to decide the cut-off with minimum inclusion and exclusion errors.

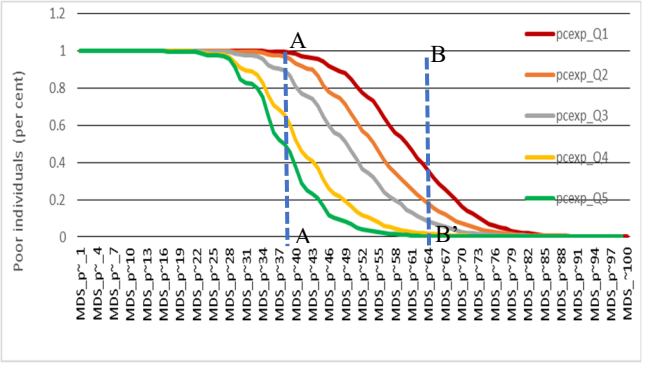


Figure 4: Distribution of multidimensional poor of per capita expenditure quintile by different deprivation cut-off k

Target Performance

Table 5 presents the under-coverage and leakage of currently existing Samurdhi beneficiaries and predicted Samurdhi beneficiaries on MDS test considering different target groups. To assess the existing Samurdhi beneficiaries three target groups were considered[12]. For predicted Samurdhi beneficiaries instead of target group 3 a new target group 4 was considered[13].

Table 5 shows that the existing selection method report high exclusion errors on all three targeting groups. Further, it reveals that the predicted Samurdhi beneficiaries-based om MDST is much more accurate than the currently available method, (under coverage and leakage is less for three types of targeting groups on MDST in compared with the currently available selection method). Nevertheless, these findings strongly suggest that the current selection beneficiary method should be reevaluated. Table 6 shows the exclusion errors with three different MDS selection cut-offs. It reveals that when increase the cut-off exclusion error is reduced due to increase of the coverage.

[12] 1). Target group 1 - Target group is poor with respect to OPL- 2019 (Updated 2012/13_NCPI)
2). Target group 2 - First real per capita expenditure quintile Q1
3). Target group 3 - Multidimensional deprivation score 5th quintile-Q5

[13] Target group 4- MPI poor

Table 5: Targeting errors on existing and predicted Samurdhi beneficiaries on different target groups

	Existing Samurdhi beneficiaries		
	Target group1(OPL)	Target group2 (rlpcexpQ1)	Target group3 (MDSQ5)
Under-coverage	60.4	61.9	62.1
Leakage	72.5	63.1	63.4
	Predicted Samurdhi beneficiaries (MDS 5th quintile)		
	Target group1(OPL)	Target group2 (rlpcexpQ1)	Target group4 (MPI poor)
Under-coverage	47.3	50.5	55.1
Leakage	62.2	50.3	63.8

Note: Under-coverage – exclusion errors.
Leakage - inclusion errors.

Key findings of multidimensional poverty on MDS approach

Multidimensional Deprivation Score can be used as a tool to measure poverty through multidimensional lens. To identify the poor individuals, it necessitates to identify the poverty cut-off. With the evidence of Figure 4, the poverty cut-off was set as k=0.5. It says that if an individual is deprived at least 11 indicators out of 22, that person is considered multidimensionally poor. Accordingly, Table 6 presents the key significant finding of MDS poor.

Table 6 reveals the incidence of poverty on MDS multidimensional score test, i.e. that the percentage of poor people is 47 per cent. The average of deprivation experience by multidimensional poor individual is 60%. That is the average proportion of weighted indicators experience by a poor person. MPI means that the poor people experience 28.2 percent of total deprivation if all people were deprived in all indicators.

Table 6: Incidence, Intensity and Multidimensional Poverty Index (MPI) for MDS, 2019

Specification	Index	Value	Confidence interval (95%)	
Poverty cut-off K=50%	MDS_MPI	0.282	0.276	0.288
	Incidence, H (%)	47.0%	46.1%	48.0%
	Intensity, A (%)	60.1%	0.599	0.603

Comparison of poverty measures by different approach

The approaches use to measure poverty depends on different objectives. The main objective of monetary poverty is to identify the individual or household experiencing economic hardship and lack of resources necessary for minimum standard of living to inform policy makers to design targeted interventions allocating necessary resources for reducing poverty effectively. The multidimensional poverty measure aims to identify the individuals who are experiencing deprivation in non-monetary aspect from different factors at the same time to targeting poor by identifying specific indicators

cause poor to formulate policies to reduce poverty at whole more effectively. Hence, multidimensional poverty measures are complimentary to the monetary poverty. Poverty headcount depends on the conditions and the techniques used in each method.

Figure 5 shows the poverty headcount given by three main approaches in Sri Lanka. Multidimensional deprivation Test Score mainly focuses on identifying the targeting beneficiaries among poor reducing exclusion and inclusion errors to support them by providing social protection assistance to uplift their living standard. Hence, in the selection process it is important to identify the individuals who rely on assistance from others to meet their daily living needs.

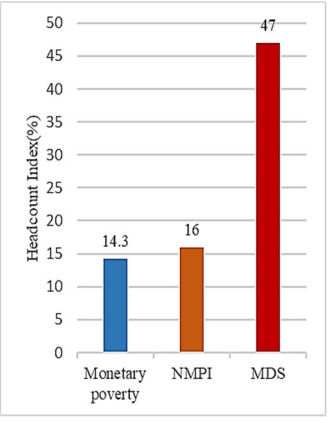


Figure 5: Poverty headcount ratio by different approaches

5. Conclusion and discussion

Developing countries face a massive challenge in implementing effective poverty reduction programs due to less effective criteria for identifying eligible welfare recipients and political interferences. The people are poor not only lack money but also the experience of deprivation in other dimensions such as health, education, shelter, nutrition, and assets at the same time. Therefore, for effective targeting, it is essential to correctly identify the needy through a selection criterion on a multidimensional approach to provide social welfare benefits.

Poverty reduction is the main objective lined with social safety net programs. Subsequently, policymakers are more concerned about exclusion errors than inclusion errors with the allocated budget. To achieve this, a proper method should be applied to cover the needy people broadly. The countries use different methods for selecting beneficiaries. Proxy Mean Test (PMT) is widely used by developing countries. However, many countries have reported a significant exclusion error based on some conceptual and methodological limitations. Hence, the countries are rethinking new selection criteria.

This paper discussed a multidimensional selection criterion for the leading social safety net for Sri Lanka, Multidimensional Deprivation Score Test (MDST). In this paper, this method has been applied to the HIES-2019 data and reveals that the exclusion error is less than the existing selection criteria when compared with different targeted groups. The MDS method computes a multidimensional deprivation score for every household. Thus, according to the selection cut-off, Samurdhi/welfare beneficiaries can be identified. The cut-off is the more critical policy decision and should be determined in terms of the impact of poverty and for an affordability within a budget. In addition, to impact of poverty, the transfer schemes should be varied concerning the severity of poverty. Otherwise, if all the beneficiaries get the same amount of money, the impact on poverty is unlikely to change significantly. In addition, to identify the suitable beneficiaries, MDST help to compute the contribution of deprivation in every dimension, which is taken into consideration by household or family, community, or geographical levels.

The results of the MDS method show that the individuals who are not identified as poor based on official poverty measures are poor in terms of the MDS method, and there are considerable gaps of the incidence of poverty across districts. Further, when compared with current Samurdhi targeting, the performance varies across district and evidence that the current selection method is associated with high exclusion errors.

Sri Lanka is currently selecting the beneficiaries considering the family aspect based on monetary measures. This paper utilizes HIES 2019 data and assesses the selection performance at the household level. Consequently, the outcome performance might not match accurately. The Samurdhi beneficiary family background might be different from the household background.

Poverty-targeting measures are more productive when the analysis is focused on poor people. The MDST method for selection criteria is more productive to apply to get information from existing and potential beneficiaries first and then apply the MDST criteria. The MDST method is a data-driven approach focusing on the target population to make an evidence-based policy decision to reduce poverty based on current information. MDST depends on the dimensions and indicators decided use for the criteria and the selection cut-off. This test provides important policy decisions for the government for effective targeting to reduce poverty.

This analysis has been carried out considering the entire population based on a representative sample used for the Household Income and Expenditure survey conducted in 2019. The proposed MDST for the selection performance can be properly assessed when applied to the targeting group based on the multidimensional poverty approach and considering the selected beneficiaries from the MDS method. To improve the effectiveness of this method it would be more accurate to collect the information from existing and potential beneficiaries and assess the targeting performance through a subjective evaluation at the community level.

References

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., and Tobias, J., (2012). Targeting the poor: evidence from a field experiment in Indonesia. *American Economic Review, 102*(4), pp. 1206–1240.

Alkire, S. K., (2023). Moderate multidimensional poverty index: paving the way out of poverty. *Social Indicators Research, 168*(1), pp. 409–445.

Alkire, S., Seth, S., (2013). Selecting a targeting method to identify BPL households in India. *Social indicators research, 112*, pp. 417–446.

Alkire, S., Roche, J. M., Ballon, P., Foster, J., Santos, M. E. and Seth, S., (2015). *Multidimensional poverty measurement and analysis*. Oxford University Press, USA.

Beuermann, D. W., Hoffmann, B., Stampini, M., Vargas, D. and Cossío, D. A., (2024). Shooting a moving target: Choosing targeting tools for social programs. *IDB Working Paper Series, No. IDB-WP-1559*.

Bird, N., Hanedar, E., (2023). Expanding and Improving Social Safety Nets Through Digitalization. *International Monetary Fund*.

Brown, C., Ravallion, M. and Van de Walle, D., (2018). A poor means test? Econometric targeting in Africa. *Journal of Development Economics, 134*, pp. 109–124.

Department of Census and Statistics, (2021b). *Multidimensional Poverty in Sri Lanka-2019*. Department of Census and Statistics, Sri Lanka.

Department of Census and Statistics, (2023a, January). Retrieved from Department of Census and Statistics: www.statistics.gov.lk/NationalAccounts/StaticalInformation/Reports/press_note_2022q3_en.

Department of Census and Statistics, (2021a). Poverty Indicators-2019.

Department of Census and Statistics, (2023b, February). Retrieved from www.statistics.gov.lk/InflationAndPrices/StaticalInformation/MonthlyCCPI/Inflation-FoodAndNonFoodGroups.

Devereux, S., Masset, E., Sabates-Wheeler, R., Samson, M., Rivas, A. M. and Te Lintelo, D., (2017). The targeting effectiveness of social transfers. *Journal of Development Effectiveness, 9*(2), pp. 162–211.

Diamond, A., Gill, M., Rebolledo Dellepiane, M. A., Skoufias, E., Vinha, K. and Xu, Y., (2016). Estimating poverty rates in target populations: An assessment of the simple poverty scorecard and alternative approaches. *World Bank Policy Research Working Pape, 7393*.

Grosh, M. E., (1994). *Administering targeted social programs in Latin America: From platitudes to practice* (Vol. 94). World Bank Publications..

Kidd, S., Wylde, E., (2011). Targeting the Poorest: An assessment of the proxy means test methodology. *AusAID Research Paper*.

Morestin, F., Grant, P. and Ridde, V., (2009). Criteria and processes for identifying the poor as beneficiaries of programs in developing countries. *Policy brief. Montreal: University of Montreal.*

Narayan, A., Yoshida, N., (2005). Proxy Means Tests for Targeting Welfare Benefits in Sri Lanka. *South Asia Poverty Reduction and Economic Management*.

Sabates-Wheeler, R., Hurrell, A. and Devereux, S., (2015). Targeting social transfer programmes: Comparing design and implementation errors across alternative mechanisms. *Journal of International Development, 27*(8), pp. 1521–1545.

Samaraweera, G. C., (2010). Economic and social assessment of poverty alleviation programs in Sri Lanka-special reference to the Gemidiriya community development and livelihood improvement project. *Journal of Emerging Trends in Economics and Management Sciences, 1*(1), pp. 60–65.

Sen, A., (1983). Poor, relatively speaking. *Oxford economic papers, 35*(2), pp. 153–169.

Sen, A., (1995). *Inequality reexamined.* Harvard university press.

Sen, A. K., (1997). From income inequality to economic inequality. *Southern Economic Journal, 64*(2), pp. 384-401.

Tilakaratna, G., Jayawardana, S., (2015). *Social protection in Sri Lanka: current status and effect on labor market outcomes.*

World Bank, (1999). *Improving social assistance in Armenia.* Human Development Unit, Country Department III, Europe and Central Asia Region, World Bank, Washington, DC.

Appendix

The list of Dimensions, Indicators, and definition

Dimension	Indicator	Definition
1. Education	1. Education Level of family members	A household is considered poor based on this indicator when all household members have less than O/L (or poor) education
	2. Number of non-school going children between the age of 5-16 years	A household is considered poor based on this indicator if at least one school aged (5-16) child is not enrolled in school
2. Health	1. Family members suffering from long term chronic diseases	A household is considered poor based on this indicator if at least one family member has suffered from a chronic disease
	2. Family members with disabilities	A household is considered poor based on this indicator if at least one family member is disabled
3. Economic Level	1. Monthly per capita expenditure	A household is considered poor based on this indicator when monthly per capita expenditure is less than Rs. 13,500
	2. Monthly per capita income	A household is considered poor based on this indicator when monthly per capita income is less than Rs. 14,000
	3. Electricity consumption less than 60 units per month	A household is considered poor based on this indicator when electricity consumption is less than 60 units (Rs.472) per month
4. Assets	1. Not having ownership of the occupied house and land to a family member	A household is considered poor based on this indicator if it does not have ownership of the occupied house and land to a family member
	2. Not having ownership of other house or a building to a family member	A household is considered poor based on this indicator if it does not have ownership of other houses and buildings
	3. Not having at least 0.5 acre of cultivable highland to a family	A household is considered poor based on this indicator if it does not have at least 0.5 acre of highland to a family
	4. Not having at least one acre of cultivable paddy land to a family	A household is considered poor based on this indicator if it does not have at least one acre of paddy land to a family
	5. Not having at least one asset related to mobility (Motor bike CC 125>, Three-wheeler, Car, Van, Jeep, Bus, Lorry, Tipper, Hand tractor (2 wheels), Tractor (4 wheels)	A household is considered poor based on this indicator if it does not have at least one asset related to mobility

The list of Dimensions, Indicators, and definition (cont.)

Dimension	Indicator	Definition
	6. Not having at least one asset related to economic activity (Fishing boat, Combined harvest machines, Threshers)	A household is considered poor based on this indicator if it does not have at least one asset related to mobility
	7. Not having at least one asset related to livelihood (5 cattle for milk, 20 goats, 50 chickens, 50 ducks, 10 swine)	A household is considered poor based on this indicator if it does not have at least one asset related to livelihood
5. Housing condition	1. Living in line room/row house/slum/shanty or other.	A household is considered poor based on this indicator when living in line room/row house/slum/shanty or other
	2. Not having a living home with a permanent wall and permanent floor and permanent roof	A household is considered poor based on this indicator if it does not have a living home with a permanent wall, floor, and roof
	3. Total floor area is less than 500 square feet	A household is considered poor based on this indicator if it lives in a house with floor area less than 500 square feet
	4. No access to clean drinking water	A household is considered poor based on this indicator if it does not have access to clean drinking water
	5. No access to adequate sanitation	A household is considered poor based on this indicator if it does not have access to adequate sanitation
	6. Not access to electricity	A household is considered poor based on this indicator if it does not have access to electricity
6. Family Demography	1. Dependency ratio (number of people aged 0-14 and those aged 65 and over/number of people aged 15 – 64) greater than 0.65	A household is considered poor based on this indicator if dependency ratio is greater than 0.65
	2. Single parent family	A household is considered poor based on this indicator when the family is a single parent family. ** In HIES data file households are nuclear families or extended families or one person **Here we can only identify single parents with children age<18, when a single parent is the head of the household only

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 51–67, https://doi.org/10.59139/stattrans-2025-004
Received – 18.04.2023; accepted – 11.07.2024

On survival estimation of Lomax distribution under adaptive progressive type-II censoring

Hemani Sharma[1], **Parmil Kumar**[2]

Abstract

The main objective of the research described in the article is to study the maximum likelihood (ML) estimation and the Bayesian approach for parameter estimation of the Lomax distribution. Additionally, the study aims to determine the approximate intervals for the parameters and the survival function based on adaptive progressive type-II censored data. The ML estimators of the probability distribution's parameters were calculated using the Newton-Raphson method, while the delta method was utilised to compute the approximate confidence intervals for the survival function. The Bayesian approach was also used to estimate the unknown parameters and survival function. This was achieved through the construction of Bayesian estimators under an informative and non-informative prior based on the squared error loss function (SELF) and approximate credible intervals. The Markov Chain Monte Carlo (MCMC) method was employed for this purpose. A Monte Carlo analysis was conducted to test the efficiency of the proposed method in various situations based on different criteria such as mean-squared error, bias, coverage probability, and expected length-estimated criteria. The results indicate that the Bayesian approach out-performs the likelihood method in estimating the Lomax model parameters. Finally, the study includes an application of these methods to real data.

Key words: Lomax distribution, maximum likelihood (ML); bayesian estimation; adaptive progressive type-II censoring scheme; squared error loss function (SELF).

1. Introduction

The Lomax distribution is a probability distribution that is widely used in reliability and survival analysis. The distribution is named after K. S. Lomax (1954), who first introduced it in 1954. It is a parametric distribution that is used to model the lifetime of products or systems, and it has several applications in engineering, medical sciences, and social sciences. The Lomax distribution is also known as the Pareto Type II distribution. The PDF of the Lomax distribution with shape parameter β and scale parameter θ is given by

$$f(x;\beta,\theta) = \theta\beta(1+\theta x)^{-(\beta+1)}; x,\beta > 0, \theta > 0 \tag{1}$$

and the corresponding Cumulative Distribution Function (CDF) and Survival Function is given as

$$F(x;\beta,\theta) = 1-(1+\theta x)^{-\beta}; x,\beta,\theta > 0 \tag{2}$$

[1]University of Jammu, J&K, Jammu, India. E-mail: hemanisharma124@gmail.com. ORCID: https://orcid.org/0009-0006-1229-8360.

[2]University of Jammu, J&K, Jammu, India. E-mail: parmil@yahoo.com. ORCID: https://orcid.org/0000-0003-1299-8587.

$$S(t;\beta,\theta) = (1+\theta t)^{-\beta}; t,\beta,\theta > 0 \tag{3}$$

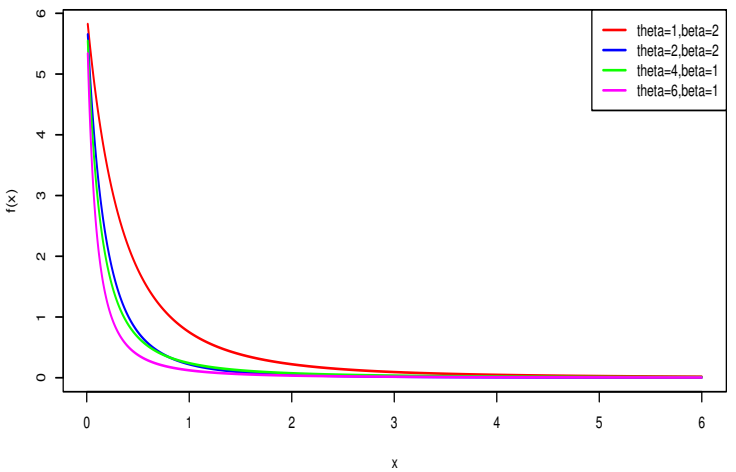


Figure 1: PDF of Lomax distribution for different values of parameters

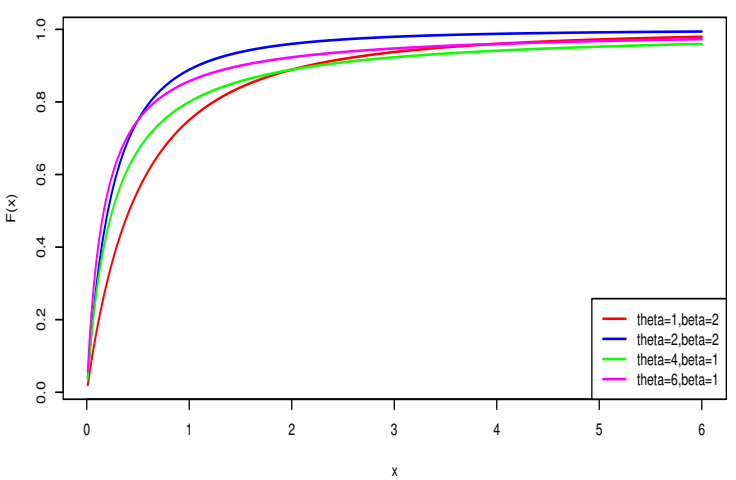


Figure 2: CDF of Lomax distribution for different values of parameters

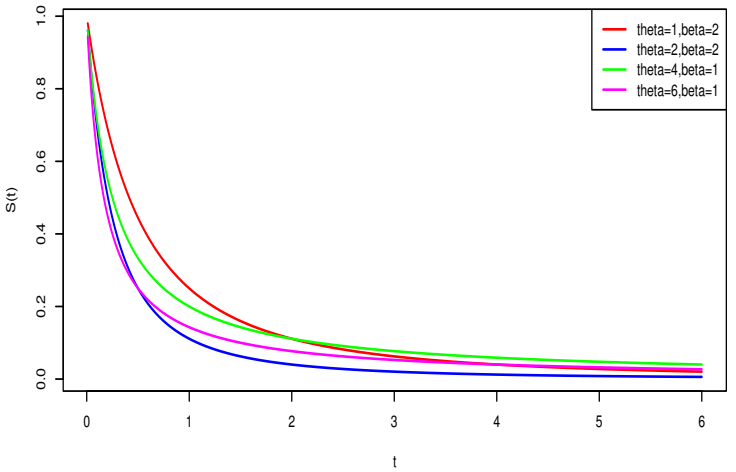


Figure 3: Survival Function of Lomax distribution for different values of parameters

The two-parameter Lomax distribution model, traditionally characterized by its shape parameter β and scale parameter θ, can be generalized to include the effect of explanatory variables. In this generalized model, the scale parameter θ is expressed as a function of the covariates through a log-linear relationship. This allows the model to account for the influence of various factors (denoted by Z) on the scale parameter, thereby providing a more flexible and comprehensive framework for modeling data that may be influenced by multiple explanatory variables. When incorporating explanatory variables Z, we often model β or θ (or both) as functions of Z, similar to the approaches used by Altun (2021) and Khan and Khan (2020). These authors demonstrated that using such link functions provides a flexible alternative to models like gamma regression, allowing for more nuanced analysis by accounting for the effects of various control variables on the distribution's parameters. Both MLE and Bayesian Estimation allows for the inclusion of explanatory variables by modeling the parameters of the Lomax distribution as functions of these variables. This generalization enhances the model's applicability in fields like survival analysis and reliability engineering, where understanding the impact of multiple covariates is crucial.

Balkema and de Haan (1974) has used this distribution for reliability and life testing experiment. Hassan and Al-Ghamdi (2009) studied the optimum step stress accelerated life testing for the Lomax distribution using maximum likelihood procedure. In many real-life situations, the lifetime of a product or system is subject to progressive type-II censoring. In such cases, the lifetime of a unit is only observed up to a certain point, and then it is censored. Adaptive progressive type-II censoring is a type of censoring where the sample size changes based on the current state of the experiment. This type of censoring is commonly used in reliability testing and is considered more efficient than traditional censoring methods. In some experiments, it may not be possible to observe the lifetime of all experimental

units within the available time. In such cases, censoring is used to reduce the duration and costs associated with the experiment. The two most common censoring schemes are type-I and type-II censoring, which end the experiment at a predetermined time or after a specified number of failures, respectively. However, these schemes lack the flexibility to remove units at points other than the end of the experiment. To address this, a progressive type-II censoring scheme was introduced in real-life tests, and a more flexible scheme called the type-II hybrid progressive censoring was proposed. An adaptive type-II progressive censoring scheme that combines type-I and type-II progressive censoring was also proposed for real-life studies. In a reliability experiment with n identical, independent units, the values of m and n are predetermined and before the experiment begins, a progressive censoring scheme $R = (R_1, ..., Rm)$ is given. It is possible that the experimental total time may exceed the pre-fixed time T. J denotes the observed failure times before the predetermined time T, i.e. $X_{J:m:n} < T < X_{J+1:m:n}, J = 0, 1, ..., m$ where$X_{J:m:n}$, $T < X_{J+1:m:n}, J = 0, 1, ..., m$ where $X_{0:m:n} = 0$ and $X_{m+1:m:n} = \infty$. When the experiment's total time exceeds the ideal test time T, the scheme sets $R_{J+1} = ... = R_{m-1} = 0$ and $R_m = n - m - \sum_{i=J}^{m} R_i$. This allows the experiment to end as soon as possible, with no survival units removed except at the time of the m^{th} failure. There have been several studies on the Lomax distribution under different types of censoring. Cramer and Schmiedt (2011) has considered progressively type-II censored competing risks data from the Lomax distribution and discuss the applicability of the model in the presence of censoring schemes. In recent years, the Adaptive IIPH censoring scheme has been studied by a vast number of authors, including Cui et al. (2019), who discussed the problem of estimating the Weibull distribution parameters in a constant-stress accelerated life test. Sewailem and Baklizi (2019) provided inference for the log-logistic distribution based on an adaptive progressive type-II censoring scheme. Ye et al. (2014) estimated the parameters of the extreme value distribution using the maximum likelihood technique (MLE). Helu and Samawi (2021) studied Statistical analysis based on adaptive progressive hybrid censored data from the Lomax distribution. Helu (2022) discussed Adaptive Type-II Hybrid Progressive Schemes Based on Maximum Product of Spacings for Parameter Estimation of Kumaraswamy Distribution. Nassr et al. (2021) studied statistical inference for the extended Weibull distribution based on adaptive type-II progressive hybrid censored competing risks data. Chen and Gui (2020) discussed the problem of estimating the parameters of the bathtub-shaped failure rate function. Panahi et al. (2021) derived the maximum likelihood and Bayes estimates for the Burr Type-III distribution. Kohansal and Shoaee (2021) studied the statistical inferences for a multicomponent stress-strength reliability model. Okasha et al. (2021) discussed Reliability Estimation of the Lomax Distribution under Adaptive Type-I Progressive Hybrid Censoring Scheme. The purpose of this study is to explore and investigate the Lomax distribution under adaptive progressive type-II censoring. Specifically, this study aims to estimate the parameters and survival function of the Lomax distribution based on the adaptive progressive type-II censored data.

The structure of the article is as follows: Section 1 provides an introduction, outlining the research problem and objectives. Section 2 focuses on estimating the parameters and survival function using Maximum Likelihood Estimation (MLE). Section 3 presents the confidence intervals for the parameters and survival function. Section 4 presents the Bayesian estimators for the parameters and survival function based on SELF. To assess the performance of

the estimators, a simulation study is conducted in Section 5 and the estimators are compared using the R software. Section 6 presents the analysis of a real-life dataset to demonstrate the practical application of the proposed estimators. Finally, in Section 7, the article concludes by summarizing the key findings and implications of the study.

2. Maximum Likelihood Estimation

Suppose that $X^R_{1:m:n}, X^R_{2:m:n}, ...X^R_{m:m:n}$ is an adaptive progressive type-II censored sample of size m from a sample of size n with censoring scheme $R = (R_1, R_2, ..., R_m)$ taken from distribution having f(x) as the PDF and F(x) as the CDF, and $X_{J:m:n}$ is the last observed failure before T which is prefixed best testing time. The observed values of an adaptive type-II progressively censored sample are represented by $x = x^R_{1:m:n}, x^R_{2:m:n}, ...x^R_{m:m:n}$ (simplified as $x = x_1, x_2, ...x_m$ in later equations). On this basis, the corresponding likelihood function is given by

$$L(x^R_{1:m:n}, x^R_{2:m:n}, ..., x^R_{m:m:n}) = D_J \prod_{i=1}^{m} f(x_{i:m:n}) \left[\prod_{i=1}^{J}(1 - F(x_{i:m:n}))\right]^{R_i} \left[(1 - F(x_{m:m:n}))\right]^{R_J} \quad (4)$$

$D_J = \prod_{i=1}^{m}[n - i + 1 - \sum_{k=1}^{max(i-1,J)} R_k]$ and $R_J = n - m - \sum_{i=i}^{J} R_i$.

The Likelihood function for $x^R_{1:m:n}, x^R_{2:m:n}, ..., x^R_{m:m:n}$based on the Lomax distribution is written as

$$L(\beta, \theta; x) = D_J \prod_{i=1}^{m} \left[\theta\beta(1 + \theta x_i)^{-(\beta+1)}\right] \left[\prod_{i=1}^{J}(1 + \theta x_i)^{-\beta}\right]^{R_i} \left[(1 + \theta x_m)^{-\beta}\right]^{R_J} \quad (5)$$

Further, the log-likelihood function can be written as

$$\ln L(\beta, \theta; x) = m\ln(\theta) + m\ln(\beta) - (\beta + 1)\sum_{i=1}^{m} \ln(1 + \theta x_i) - \beta \sum_{i=1}^{J} R_i \ln(1 + \theta x_i) - \beta R_J \ln(1 + \theta x_m) \quad (6)$$

Then, take the partial derivative of the log-likelihood function, and obtain the likelihood equations as:

$$\frac{\partial \ln L(\beta, \theta; x)}{\partial \theta} = \frac{m}{\theta} - (\beta + 1)\sum_{i=1}^{m} \frac{x_i}{1 + \theta x_i} - \beta \sum_{i=1}^{J} \frac{R_i x_i}{1 + \theta x_i} - \beta R_J \frac{x_m}{1 + \theta x_m} = 0 \quad (7)$$

$$\frac{\partial \ln L(\beta, \theta; x)}{\partial \beta} = \frac{m}{\beta} - \sum_{i=1}^{m} \ln(1 + \theta x_i) - \sum_{i=1}^{J} R_i \ln(1 + \theta x_i) - R_J(1 + \theta x_m) = 0 \quad (8)$$

Equations (7) and (8) cannot be solved for β and θ explicitly. So, these equations required numerical solving.

The ML estimator for the survival function by using the invariance property of ML

estimator is as follows:

$$\hat{S(t)} = (1+\hat{\theta}t)^{-\hat{\beta}} \tag{9}$$

3. Asymptotic Confidence Intervals

The Fisher information matrix was discussed by Aldrich (1997) and the consequently observed Fisher information matrix of the parameters β and θ for large n, is given as follows:

$$I(\hat{\beta},\hat{\theta}) = \begin{bmatrix} -\frac{\partial^2 \ln L(\beta,\theta;x)}{\partial \beta^2} & -\frac{\partial^2 \ln L(\beta,\theta;x)}{\partial \beta \partial \theta} \\ -\frac{\partial^2 \ln L(\beta,\theta;x)}{\partial \theta \partial \beta} & -\frac{\partial^2 \ln L(\beta,\theta;x)}{\partial \theta^2} \end{bmatrix}_{\hat{\beta},\hat{\theta}} \tag{10}$$

where

$$\frac{\partial^2 \ln L(\beta,\theta;x)}{\partial \beta^2} = -\frac{m}{\beta^2}$$

$$\frac{\partial^2 \ln L(\beta,\theta;x)}{\partial \theta^2} = -\frac{m}{\theta^2} + (\beta+1)\sum_{i=1}^{m}\frac{x_i^2}{(1+\theta x_i)^2} + \beta\sum_{i=1}^{J}\frac{R_i x_i^2}{(1+\theta x_i^2)} + \beta R_J \frac{x_m^2}{(1+\theta x_m)^2}$$

$$\frac{\partial^2 \ln L(\beta,\theta;x)}{\partial \theta \partial \beta} = -\sum_{i=1}^{m}\frac{x_i}{(1+\theta x_i)} - \sum_{i=1}^{J}\frac{R_i x_i}{(1+\theta x_i)} - R_J\frac{x_m}{(1+\theta x_m)}$$

$$\frac{\partial^2 \ln L(\beta,\theta;x)}{\partial \beta \partial \theta} = -\sum_{i=1}^{m}\frac{x_i}{(1+\theta x_i)} - \sum_{i=1}^{J}\frac{R_i x_i}{(1+\theta x_i)} - R_J\frac{x_m}{(1+\theta x_m)}$$

It is difficult to find the expected Fisher information analytically. Therefore, by using the concept of large sample theory and the variance covariance matrix, which is the inverse of the observed Fisher information matrix $I^{-1}(\hat{\beta},\hat{\theta})$, the approximate $100(1-\alpha)$ normal confidence intervals for the parameters β and θ are given respectively as

$$\left(\hat{\beta} - z_{\frac{\alpha}{2}}\sqrt{var(\hat{\beta})}, \hat{\beta} + z_{\frac{\alpha}{2}}\sqrt{var(\hat{\beta})}\right) \tag{11}$$

$$\left(\hat{\theta} - z_{\frac{\alpha}{2}}\sqrt{var(\hat{\theta})}, \hat{\theta} + z_{\frac{\alpha}{2}}\sqrt{var(\hat{\theta})}\right) \tag{12}$$

where $z_{\frac{\alpha}{2}}$ is the percentile of the standard normal distribution N(0,1) with right-tail probability $\frac{\alpha}{2}$. In addition, the Delta method (Greene, 2010), is applied to evaluate the approximate confidence intervals for the survival functions S(t). This is a natural way for calculating the confidence interval for the functions of the ML estimators, in which these functions are intractable to calculating the variance analytically. Then, we create linear approximations of this survival function and then calculate the variance of linear approximation as follows:

$$C = \begin{pmatrix} \frac{\partial S(t)}{\partial \beta} & \frac{\partial S(t)}{\partial \theta} \end{pmatrix} \tag{13}$$

where

$$\frac{\partial S(t)}{\partial \beta} = -(1+\theta t)^{-\beta}.\ln(1+\theta t) \tag{14}$$

$$\frac{\partial S(t)}{\partial \theta} = -\beta t(1+\theta t)^{(-\beta-1)} \tag{15}$$

The approximate estimate of the variance of S(t) is given by the following:

$$var(\hat{S(t)}) = \left[C^t I^{-1}(\beta,\theta)C\right]_{\hat{\beta},\hat{\theta}}$$

Then, the approximate confidence interval for S(t) is as follows:

$$\left(\hat{S(t)} - z_{\frac{\alpha}{2}}\sqrt{var(\hat{S(t)})}, \hat{S(t)} + z_{\frac{\alpha}{2}}\sqrt{var(\hat{S(t)})}\right) \tag{16}$$

where $z_{\frac{\alpha}{2}}$ is the upper $(\frac{\alpha}{2})^{th}$ quantile of the standardized normal distribution.

4. Bayesian Estimation

Bayesian estimation is a statistical method for estimating the parameters of a probability distribution based on prior knowledge and observed data. In this approach, the unknown parameters are treated as random variables with their own prior probability distributions, and the observed data are used to update these prior distributions to obtain a posterior distribution that reflects both the prior information and the new evidence provided by the data. It includes the ability to incorporate prior knowledge into the analysis, the flexibility to handle complex models and data structures, and the ability to quantify uncertainty in a more intuitive way than traditional frequentist methods. In this paper, the Bayes estimates under the Squared Error Loss Function (SELF) are constructed for the unknown parameters (θ,β) and for the survival function. The corresponding credible intervals for these quantities are calculated. It is supposed that the unknown parameters β and θ are independent and follow the gamma distributions as

$$\pi(\beta) \propto \beta^{a_1-1}e^{-b_1\beta}; a_1, b_1 > 0$$

$$\pi(\theta) \propto \theta^{a_2-1}e^{-b_2\theta}; a_2, b_2 > 0$$

Thus, the joint prior distribution becomes

$$\pi(\beta,\theta) \propto \beta^{a_1-1}\theta^{a_2-1}e^{-(b_1\beta+b_2\theta)} \tag{17}$$

The non-informative priors for both parameters β and θ are considered to be $\pi_1(\theta) \propto 1$ and $\pi_2(\beta|\theta) \propto \frac{1}{\beta}$. When $\pi_1(\theta)$ is multiplied by the $\pi_2(\beta|\theta)$, corresponding prior density of β and θ is given by $\pi(\beta,\theta) = \pi_1(\theta) * \pi_2(\beta|\theta)$; Clearly, $\pi(\beta,\theta) \propto \frac{1}{\beta}$. Subsequently, the general form of the posterior density is proportional to the likelihood function time of the prior density function, as follows:

$$p(\beta,\theta|x) \propto (likelihood \times prior)$$

And the corresponding joint posterior conditional density function with informative priors is

$$p(\beta,\theta|x) \propto \left[\prod_{i=1}^{m}\theta\beta(1+\theta x_i)^{-(\beta+1)}\right]\left[\prod_{i=1}^{J}(1+\theta x_i)^{-\beta}\right]^{R_i}\left[(1+\theta x_m)^{-\beta}\right]^{n-m-\sum_{i=i}^{J}R_i} \times\beta^{a_1-1}\theta^{a_2-1}e^{-(b_1\beta+b_2\theta)} \quad (18)$$

The corresponding joint posterior conditional density function with non-informative priors is

$$p(\beta,\theta|x) \propto \left[\prod_{i=1}^{m}\theta\beta(1+\theta x_i)^{-(\beta+1)}\right]\left[\prod_{i=1}^{J}(1+\theta x_i)^{-\beta}\right]^{R_i}\left[(1+\theta x_m)^{-\beta}\right]^{n-m-\sum_{i=i}^{J}R_i} \times\frac{1}{\beta} \quad (19)$$

Hence, the Bayes estimates of any function of θ and β such as $g(\beta,\theta)$, based on SELF is obtained as

$$\widehat{g(\beta,\theta)} = E_{\beta,\theta|x}\widehat{g(\beta,\theta)} = \frac{\int_0^\infty\int_0^\infty \widehat{g(\beta,\theta)}L(\beta,\theta|x)\times\pi(\beta,\theta)d\beta d\theta}{\int_0^\infty\int_0^\infty L(\beta,\theta|x)\times\pi(\beta,\theta)d\beta d\theta} \quad (20)$$

Clearly, calculating the Bayes estimators using (18), (19) and (20) analytically is unattainable. As a result, we advocate employing the MCMC technique to obtain the Bayes estimates of θ and β and the associated credible intervals. The Metropolis-Hastings algorithm is a Markov chain Monte Carlo (MCMC) method for sampling from a probability distribution that is difficult to sample directly. It is a general algorithm that can be used to sample from any distribution, as long as the distribution can be evaluated up to a constant proportionality factor. The algorithm works by defining a proposal distribution, which is used to generate a candidate sample from the current state of the chain. The candidate sample is then accepted or rejected based on the probability of moving from the current state to the candidate state, as determined by a Metropolis-Hastings acceptance probability. To apply the MCMC technique, we should first derive the full conditional distributions of β and θ as follows:

$$h(\beta|\theta,x) \propto \beta^{m+a_1-1}e^{b_1\beta}\prod_{i=1}^{m}\left[(1+\theta x_i)^{-(\beta+1)}\right]\left[\prod_{i=1}^{J}(1+\theta x_i)^{-\beta}\right]^{R_i}\left[(1+\theta x_m)^{-\beta}\right]^{R_J} \quad (21)$$

$$h(\theta|\beta,x) \propto \theta^{m+a_2-1}e^{b_2\theta}\prod_{i=1}^{m}\left[(1+\theta x_i)^{-(\beta+1)}\right]\left[\prod_{i=1}^{J}(1+\theta x_i)^{-\beta}\right]^{R_i}\left[(1+\theta x_m)^{-\beta}\right]^{R_J} \quad (22)$$

To involve the MH sampling, we assume the normal distribution as the proposal distribution to acquire the Bayesian estimates and to obtain the credible intervals. Here, we simulate samples from the full conditional posterior distribution and the proposal proceeds by proposing a joint move on (θ, β). The Metropolis-Hasting algorithm is illustrated below.

1) Initialize j=0, $\theta^{(j)} = 1.5$, $\beta^{(j)} = 1$
2) j=1

3) Generate θ and β using normal candidate distribution.
4) Compute the acceptance probability s= $\min\left(1, \frac{p(\theta^*|data)}{p(\theta^{j-1}|data)} \frac{f(\theta^{j-1}|\theta^*)}{f(\theta^*|\theta^{j-1})}\right)$
5) Draw u from a uniform (0,1) density.
6) If $u \leq r$; set $\theta^j = \theta^*$ and otherwise $\theta^j = \theta^{j-1}$
8) Increment j and repeat steps 3 to 6 for N = 11,000 times.
9) Approximate Bayes estimates of θ and β using MCMC samples based on the SELF as $\hat{\theta}_B = \frac{1}{N-M}\sum_{i=M+1}^{N}\theta^{(i)}$ and $\hat{\beta}_B = \frac{1}{N-M}\sum_{i=M+1}^{N}\beta^{(i)}$ where M is burn-in.
10) An approximate Bayesian estimates of the S(t), based on the SELF, can be found as $\widehat{S(t)_B} = \frac{1}{N-M}\sum_{i=M+1}^{N} S^{(i)}(t)$
11) Compute the credible intervals of θ and β, order $\theta_{M+1}, \theta_{M+2}..., \theta_N$ and $\beta_{M+1}, \beta_{M+2}..., \beta_N$ as $\theta_1, \theta_2..., \theta_{N-M}$ and $\beta_M, \beta_M..., \beta_{N-M}$ Then, the $100(1-\alpha)\%$ symmetric credible intervals of θ and β constructed as $\left(\theta_{((N-M)(\frac{\alpha}{2}))}, \theta_{((N-M)(1-\frac{\alpha}{2}))}\right)$ and $\left(\beta_{((N-M)(\frac{\alpha}{2}))}, \beta_{((N-M)(1-\frac{\alpha}{2}))}\right)$.
12) Compute the credible intervals of $S(t)$ order $S_{M+1}(t), S_{M+2}(t)..., S_N(t)$ as $S_1(t) < S_2(t) < ... < S_{N-M}(t)$Then, the $100(1-\alpha)\%$ symmetric credible intervals of θ and β constructed as $\left(S_{((N-M)(\frac{\alpha}{2}))}(t), S_{((N-M)(1-\frac{\alpha}{2}))}(t)\right)$.

5. Simulation Study

In this section, Monte Carlo simulations are performed to know the performance of the proposed estimators developed in the previous sections of the parameters, the survival function based on an adaptive progressive type-II censoring scheme. The process of generating an adaptive progressive type-II censored sample with a pre-determined number of n and m and the progressive censoring schemes with given values of the ideal censoring time T from the Lomax distribution is described below using the procedure described by Balakrishnan and Sandhu (1995) and by Ng et al. (2009). The steps are as follows:
1) Define the values of n, m, θ, β, T and $R = (R_1, R_2...R_m)$.
2) Simulate m random variables from uniform (0,1) as $W_1, W_2...W_m$.
3) Set $V_i = W_i^{\frac{1}{(i+R_m+R_{m-1}+...+R_{m-i+1})}}$ for i=1,2,...m.
4) Set $U_i = V_m V_{m-1}...V_{m-i+1}$, for i=1,2,...m. Then, $U_1, U_2, ...U_m$, is the m progressive type-II observed sample from the Uniform (0,1) distribution.
5) Set $x_i = F^{-1}(Ui)$ for i=1,2,... m, where $F^{-1}(Ui)$represent the quantile function of the Lomax distribution. Thus, $x_1, x_2, ..., x_m$, is the needed progressive type-II observed sample from the specified distribution F(.) by using the inverse transformation method.
6) Identify the value of J, where $x_{J:m:n} < T < x_{J+1:m:n}$, discard the sample $x_{j+2:m:n}, ..., x_{m:m:n}$.
7) Simulate the first m -J -1 order statistics from a truncated distribution considered as $\frac{f(x)}{[1-f(x_{J+1:m:n})]}$ with sample size $\left(n - \sum_{i=1}^{j} R_i - J - 1\right)$ as $x_{j+2:m:n}, x_{j+3:m:n}, ..., x_{m:m:n}$.

Hence, a simulation study was executed using the ideal total test time T=1. To generate the data, we supposed that the initial true values of the parameters θ and β were (1.5, 1), we used the values of t= 0.5, 1, the corresponding values of the survival function are S(t)

are 0.5714 and 0.4 respectively. For prior information, the hyperparameters ($a_1 = 1, b_1 = 0, a_2 = 0, b_2 = 1$) were considered. To find the Bayesian estimates and the 95% Bayes intervals for the unknown parameters, we simulate 10,000 MCMC values from the target distribution using the Metropolis–Hastings algorithm.

Table 1: Average Estimate(AE), Bias, MSE, AL and CP of scale(θ) and shape(β) Parameters Based on T=1

(n,m)	CS	MLE		Bayes Informative		Bayes Non-Informative	
(50,20)	$(20, 0^{19})$	θ	β	θ	β	θ	β
AE		1.7038	1.1901	1.3942	0.9590	1.3112	1.0904
Bias		0.2038	0.1901	0.1058	0.0410	0.1888	0.1090
MSE		1.5171	1.5063	0.8998	0.9327	1.527	0.7821
CI		(-0.7851,0.9574)	(0.12011,1.2738)	(0.6177,1.7736)	(0.5848,1.6417)	(0.5283,1.6523)	(0.6384,1.7601)
AL		1.7426	1.15374	1.1559	1.0569	1.1239	1.1216
CP		0.912	0.905	0.925	0.965	0.901	0.945
AE	$(2^5, 1^{10}, 0^5)$	1.5862	0.9441	1.4156	1.1901	1.2270	1.1964
Bias		0.0862	-0.0558	-0.0843	0.1905	-0.2729	-0.1964
MSE		1.4522	1.1055	0.9428	0.7291	1.4048	0.9706
CI		(-0.9843,0.7409)	(-0.0415,1.2507)	(0.7000,1.7187)	(0.6674,1.6002)	(0.5484,1.6839)	(0.5979,1.7106)
AL		1.7252	1.2922	1.0186	1.0328	1.1355	1.1127
CP		0.930	0.935	0.920	0.95	0.93	0.985
AE	(1^{20})	1.7997	0.9693	1.3257	0.8897	1.4754	1.1809
Bias		0.2997	0.0306	-0.1742	-0.1102	-0.0245	0.1809
MSE		1.6223	1.3605	1.0921	0.7793	0.5040	0.3043
CI		(-0.8943,0.9702)	(0.1150,1.2709)	(0.6177,1.7736)	(0.5848,1.6417)	(0.6042,1.7491)	(0.5964,1.7373)
AL		1.8646	1.1558	1.1216	1.0895	1.4493	1.1409
CP		0.919	0.925	0.905	0.975	0.91	0.97
(70,30)	$(30,0^{29})$						
AE		1.3252	1.0809	1.7165	1.0109	1.7409	0.8815
Bias		0.1748	-0.0809	0.2165	0.0109	0.2409	-0.1184
MSE		1.1773	1.1999	0.7331	0.5108	0.9231	0.4391
CI		(-0.7000,1.0940)	(-0.0109,1.4870)	(0.5838,1.7535)	(0.6278,1.7433)	(0.5589,1.6820)	(0.5800,1.6676)
AL		1.7241	1.1980	1.0696	1.0154	1.1031	1.0875
CP		0.92	0.95	0.915	0.965	0.915	0.975
AE	$(2^{20}, 1^{10}, 0^{10})$	1.4701	1.0816	1.4667	1.0401	1.5763	0.9887
Bias		-0.0298	0.0816	-0.0332	0.0401	0.0763	-0.0112
MSE		1.1836	1.0439	0.0520	0.0190	0.9508	0.3430
CI		(-0.8295,0.9942)	(-0.1112,1.3783)	(0.6065,1.7594)	(0.5945,1.6390)	(0.5847,1.7599)	(0.5972,1.6978)
AL		1.5238	1.1896	1.0052	1.0044	1.0752	1.1006
CP		0.915	0.945	0.910	0.945	0.91	0.97
AE	(1^{30})	1.7147	0.9441	1.3224	1.0036	1.4864	1.0687
Bias		0.2147	-0.0558	-0.1775	0.0036	-0.0135	0.0687
MSE		1.5038	0.9678	0.8607	0.5595	0.2036	0.0937
CI		(-0.9605,0.7143)	(-0.0223,1.3828)	(0.6742,1.7820)	(0.6395,1.6097)	(0.6317,1.7392)	(0.6522,1.7104)
AL		1.6749	1.2052	1.1078	0.9002	1.1074	1.0582
CP		0.922	0.910	0.91	0.945	0.905	0.95
(90,40)	$(40,0^{39})$						
AE		1.3552	1.0911	1.4690	1.0950	1.2791	0.8908
Bias		-0.1448	0.0911	-0.0309	0.0950	-0.2208	0.1092
MSE		0.8560	0.8938	0.6611	0.5093	0.7338	0.3827
CI		(-0.4553,1.3351)	(0.1222,1.8344)	(0.5587,1.6813)	(0.6407,1.7708)	(0.6552,1.7481)	(0.6160,1.5899)
AL		1.6905	1.0122	1.0226	1.0250	1.0929	1.1138
CP		0.91	0.93	0.90	0.95	0.925	0.975
AE	$(2^{20}, 1^{10}, 0^{10})$	1.4701	1.2008	1.4767	1.0201	1.3709	1.0938
Bias		-0.0298	0.2008	-0.0233	0.0201	-0.1290	-0.0938
MSE		0.7395	0.8214	0.0420	0.0160	0.4358	0.2306
CI		(-0.1921,1.3903)	(0.7361,1.8359)	(0.5741,1.7177)	(0.5033,1.5436)	(0.5534,1.7066)	(0.5505,1.5849)
AL		1.4824	0.9370	0.9836	0.9943	1.0031	1.0344
CP		0.905	0.925	0.90	0.94	0.93	0.955
AE	(1^{40})	1.6753	1.0246	1.2259	1.2099	1.6934	1.0161
Bias		0.1752	0.0246	-0.2640	0.2099	0.1934	0.0161
MSE		0.8927	0.7349	0.8419	0.5322	0.1945	0.0083
CI		(-0.7775,0.8605)	(-0.0972,1.4627)	(0.6231,1.7030)	(0.6549,1.7312)	(0.6401,1.7281)	(0.6157,1.6686)
AL		1.6381	1.0599	1.0798	1.0062	1.0880	1.0528
CP		0.910	0.930	0.905	0.955	0.90	0.97

Table 2: Average Estimate(AE), Bias, MSE, AL and CP of S(t), t=0.5, 1 Parameters Based on T=1

(n,m)	CS	MLE		Bayes Informative		Bayes Non-Informative	
(50,20)	$(20,0^{19})$	**S(0.5)**	**S(1)**	**S(0.5)**	**S(1)**	**S(0.5)**	**S(1)**
AE		0.7856	0.5863	0.5422	0.3675	0.6225	0.4614
Bias		0.2141	0.1863	-0.0292	-0.0325	0.0510	0.0614
MSE		0.1111	0.1259	0.0041	0.0.0046	0.0096	0.0141
CI		(0.6302,0.9456)	(0.4404,0.9110)	(0.5164,0.5685)	(0.3383,0.3977)	(0.5277,0.7128)	(0.3488,0.5709)
AL		0.3153	0.4705	0.0578	0.0671	0.1864	0.2221
CP		0.925	0.915	0.943	0.952	0.925	0.935
AE	$(2^5,1^{10},0^5)$	0.7694	0.5671	0.6107	0.4471	0.5889	0.4248
Bias		0.1980	0.1671	0.0392	0.0471	0.0175	0.0248
MSE		0.0953	0.0959	0.0711	0.0191	0.0252	0.0388
CI		(0.6041,0.9072)	(0.4250,0.8495)	(0.5245,0.7034)	(0.3447,0.5597)	(0.5362,0.6298)	(0.3587,0.4770)
AL		0.3030	0.4245	0.1788	0.2150	0.1936	0.2402
CP		0.915	0.92	0.935	0.94	0.93	0.905
AE	(1^{20})	0.7398	0.5418	0.6063	0.4414	0.6169	0.4584
Bias		0.1684	0.1418	0.0348	0.0413	0.0455	0.0584
MSE		0.1134	0.1022	0.0091	0.0121	0.0158	0.0238
CI		(0.5214,0.7963)	(0.4072,0.7662)	(0.5217,0.7053)	(0.3442,0.5647)	(0.5699,0.8347)	(0.3938,0.7460)
AL		0.2748	0.3589	0.1836	0.2205	0.2648	0.3481
CP		0.92	0.93	0.945	0.95	0.925	0.93
(70,30)	$(30,0^{29})$						
AE		0.7514	0.5366	0.5422	0.3675	0.6231	0.4602
Bias		0.1799	0.1366	-0.0292	-0.0325	0.0517	0.0602
MSE		0.0820	0.0696	0.0041	0.0046	0.0094	0.0131
CI		(0.5632,0.8759)	(0.3614,0.7911)	(0.5164,0.5685)	(0.3383,0.3977)	(0.5306,0.7171)	(0.3519,0.5753)
AL		0.3127	0.4297	0.1751	0.2203	0.0521	0.0594
CP		0.925	0.93	0.955	0.95	0.91	0.935
AE	$(2^{10},1^{10},0^{10})$	0.7391	0.5236	0.6165	0.4537	0.6605	0.5134
Bias		0.1676	0.1238	0.0451	0.0537	0.0891	0.0113
MSE		0.0719	0.0537	0.0210	0.0172	0.0119	0.0166
CI		(0.5537,0.8362)	(0.3654,0.7352)	(0.5211,0.7150)	(0.3423,0.5731)	(0.5533,0.7352)	(0.3770,0.6174)
AL		0.2825	0.3697	0.1639	0.2008	0.1818	0.2103
CP		0.905	0.92	0.935	0.935	0.92	0.93
AE	(1^{30})	0.7391	0.5284	0.6136	0.4487	0.6824	0.5395
Bias		0.1667	0.1284	0.0422	0.0487	0.1110	0.1395
MSE		0.0796	0.0632	0.00811	0.0113	0.0126	0.0125
CI		(0.5561,0.8299)	(0.3806,0.7390)	(0.5256,0.7040)	(0.3466,0.5611)	(0.5539,0.7685)	(0.3776,0.6608)
AL		0.2738	0.3583	0.1784	0.2144	0.2145	0.2832
CP		0.93	0.89	0.94	0.945	0.93	0.915
(90,40)	$(40,0^{39})$						
AE		0.7309	0.5086	0.5571	0.3832	0.6208	0.4587
Bias		0.1595	0.1086	-0.0144	-0.0167	0.0493	0.0587
MSE		0.0668	0.0424	0.0022	0.0027	0.0089	0.0126
CI		(0.5400,0.8170)	(0.3414,0.6989)	(0.5279,0.5783)	(0.3493,0.4082)	(0.5240,0.7145)	(0.3450,0.5729)
AL		0.2770	0.3574	0.1505	0.2079	0.0503	0.0589
CP		0.925	0.91	0.955	0.955	0.93	0.925
AE	$(2^{20},1^{10},0^{10})$	0.6809	0.4522	0.6362	0.4598	0.6183	0.4789
Bias		0.1094	0.0522	0.0648	0.0598	0.0468	0.0789
MSE		0.0539	0.0234	0.0115	0.0158	0.0116	0.0161
CI		(0.4578,0.6991)	(0.2710,0.5441)	(0.2285,0.4721)	(0.5274,0.7409)	(0.3503,0.6116)	(0.3354,0.6028)
AL		0.2413	0.2731	0.1234	0.1612	0.1589	0.1973
CP		0.91	0.92	0.95	0.94	0.93	0.935
AE	(1^{40})	0.7210	0.5043	0.6156	0.4543	0.7160	0.5792
Bias		0.1496	0.1043	0.0442	0.0543	0.1446	0.1792
MSE		0.0654	0.0407	0.0077	0.0101	0.0111	0.0118
CI		(0.5654,0.8275)	(0.3393,0.6841)	(0.5191,0.6973)	(0.3416,0.5519)	(0.5463,0.6668)	(0.3701,0.5263)
AL		0.2621	0.3448	0.1782	0.2103	0.1205	0.1562
CP		0.925	0.91	0.94	0.935	0.925	0.905

We generated 10,000 MCMC samples and then discard the first 1000 random values. Table 1 and 2 summarizes the ML estimators and the Bayes estimators for the parameters θ,β and S(t)) via the censored sample. Furthermore, from this table, it seems that the Bayes estimates under the non-informative prior and the ML Estimator were close to each other. The approximate 95% confidence intervals were computed together with the corresponding length for each interval, as reported below in Table 1 and 2. From these tables, it was discovered that the average length of the confidence interval and the credible interval decreased as n and m increased. The coverage probabilities of the confidence intervals based on the likelihood are close to the nominal level of 0.95 for θ and β, and S (t =0.5, 1) as n grew larger, but failed to reach the desired level for small values of n. On the other hand, the coverage probabilities of the credible intervals approached the nominal level of 0.95 for θ and β and S (t =0.5, 1) in most cases.

6. Real Data Analysis

In this section, we consider a real life data to demonstrate the proposed method and verify how our estimates work in practice. The dataset was initially considered by Chhikara and Folks (1977). It represents the 46 repair times (in hours) for an airborne communication transceiver. The ordered dataset is presented below:

0.2, 0.3, 0.5, 0.5, 0.5, 0.5, 0.6, 0.6, 0.7, 0.7, 0.7, 0.8, 0.8, 1.0, 1.0, 1.0, 1.0, 1.1, 1.3, 1.5, 1.5, 1.5, 1.5, 2.0, 2.0, 2.2, 2.5, 2.7, 3.0, 3.0, 3.3, 3.3, 4.0, 4.0, 4.5, 4.7, 5.0, 5.4, 5.4, 7.0, 7.5, 8.8, 9.0, 10.3, 22.0, 24.5.

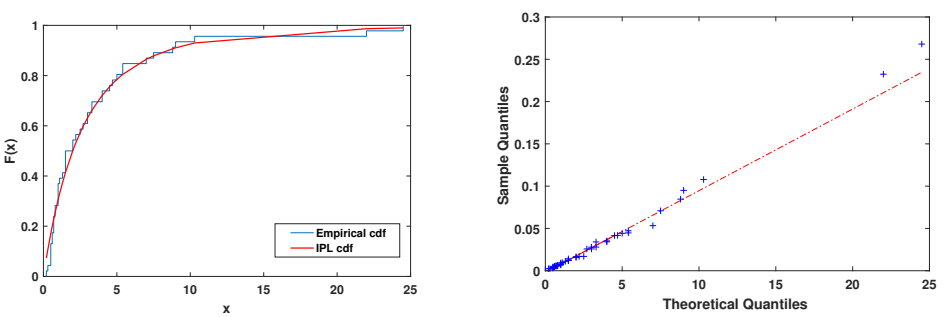


Figure 4: (a) ECDF plot for the dataset I (b) Q-Q plot for the dataset I

Table 3: Adaptive Progressive Type-II censored sample for n=46, m=20

T=7.5, j=20	0.3, 0.5, 0.5, 0.5, 0.5, 0.7, 0.7, 1.0, 1.0, 1.5, 1.5, 2.0, 2.2, 2.5, 2.7, 3.3, 4.5, 4.7, 5.4, 7.0
T=2.0, j=11	0.2, 0.3, 0.5, 0.5, 0.6, 0.7, 1.0, 1.0, 1.0, 1.3, 1.5, 2.0, 2.2, 2.7, 4.0, 4.0, 4.75.4, 7.5, 22.0.

In this illustration, the value of the Kolmogorov-Smirnov (K-S) distance and its corresponding p-value for the dataset are 0.1272 and 0.4462 respectively. It indicates that the dataset fits well through this distribution. This can further be seen through the visualization of the empirical Cumulative Distribution Function (ECDF) plot, the quantile-quantile (Q-Q) plot, as shown in Figure 4. The ML estimators for the unknown quantities are computed for the complete sample (uncensored), i.e. n=m, (θ=0.1082 and β=3.5494) the dataset was used to simulate an adaptive progressive type-II censored sample with m = 20 and with two distinct values of ideal total test time T (2.0,7.5), as displayed in Table 3. For clarity $R = (5, 0^5)$ is given as a short form of $R = (5, 0, 0, 0, 0, 0)$. Thus, the observed adaptive progressive type-II censored samples are shown below in Table 3, for two different values of T and two distinct values of J. If J = 11 means that only 11 observed failures were observed before time T = 2.0 and J = 20 means that all the observed failure times were observed before time T = 7.5, then this implies that the experiment ended before time T. Table 4 and 5 represents the average estimates, CI and AL based on dataset I for the different values of T and R.

Table 4: AE, CI, and AL of θ, β and S(t), t=0.5,1 Parameters Based on Real dataset I for n=46, m=20, T=2.0, R=(20,0^{19})

MLE				**Bayesian**			
θ	β	**S(0.5)**	**S(1)**	θ	β	**S(0.5)**	**S(1)**
0.5277	0.3698	0.9170	0.8549	0.9188	0.9241	0.4620	0.3765
(-0.1641,1.2197)	(0.0620,0.6776)	(0.7639,1.0700)	(0.6097,1.1000)	(0.0921,1.5362)	(0.1335,1.7083)	(0.1652,0.7379)	(0.0970,0.6672)
1.3838	0.6156	0.3061	0.4903	1.4441	1.5747	0.5726	0.5701

Table 5: AE, CI, and AL of θ, β and S(t), t=0.5, 1 Parameters Based on Real dataset I for n=46, m=20, T=7.5, R=(10,0^{18},10)

MLE				**Bayesian**			
θ	β	**S(0.5)**	**S(1)**	θ	β	**S(0.5)**	**S(1)**
0.2651	1.0303	0.8796	0.7847	1.2675	0.9710	0.3795	0.2685
(-0.2903,0.8206)	(-0.5284,1.2289)	(0.4890,1.2701)	(0.1478,1.4216)	(0.1240.1.7566)	(0.2088,1.5464)	(0.2746,0.6518)	(0.1441,0.5712)
1.1103	1.7573	0.0.7811	1.2738	1.6325	1.3375	0.3772	0.4271

Dataset 2: The data represents the breakdown time of an insulating fluid between electrodes at a voltage of 34 kV studied by Nelson (1982). The data are recorded as follows:

0.96, 4.15, 0.19, 0.78, 8.01, 31.75, 7.35, 6.50, 8.27, 33.91, 32.52, 3.16,4.85, 2.78, 4.67, 1.31, 12.06, 36.71, 72.89.

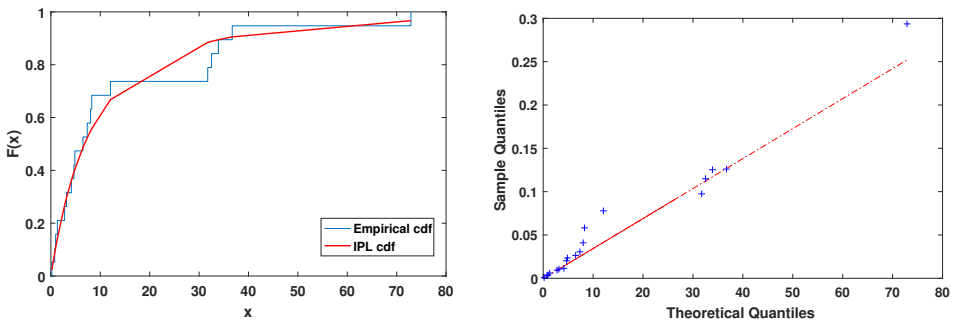


Figure 5: (a) ECDF plot for the dataset II (b) Q-Q plot for the dataset II

Table 6: Adaptive Progressive Type-II censored Sample for n=19 and m=10

T=7, j=5	0.19, 0.96, 4.15, 4.85, 6.50, 8.01, 31.75, 32.52, 33.91, 36.71
T=37, j=10	0.19, 1.31, 2.78, 3.16, 4.15, 4.85, 6.50, 32.52, 33.91, 36.71.

In this illustration, the Kolmogorov-Smirnov (K-S) distance and its corresponding p-value for the dataset are 0.1479 and 0.7467 respectively. It indicates that the dataset fits well through this distribution. This can further be seen through the visualization of the empirical Cumulative Distribution Function (ECDF) plot, the quantile-quantile (Q-Q) plot as shown in Figure 5. The ML estimators for the complete sample (uncensored), i.e. n=m, (θ=0.0597 and β=2.0323) the dataset was used to simulate an adaptive progressive type-II censored sample, as displayed in Table 6 with m = 10 and with two distinct values of ideal total test time T (7,37). Thus, the observed adaptive progressive type-II censored samples are shown below in Table 6, for two different numbers of T and two distinct numbers of J. If J =5 means that only 5 observed failures were observed before time T = 7 and J = 10 means that all the observed failure times were observed before time T = 37, then this implies that the experiment ended before time T. Table 7 and 8 presents the values of AEs, CI and AL based on dataset II for different values of T and R.

Table 7: AE, CI, and AL of θ, β and S(t), t=0.5, 1 Parameters Based on Real dataset II for n=19, m=10, T=37, R=(5,0^8,5)

MLE				**Bayesian**			
θ	β	**S(0.5)**	**S(1)**	θ	β	**S(0.5)**	**S(1)**
0.1331	0.5242	0.9667	0.9365	1.1017	1.0280	0.3675	0.2657
(-0.2363,0.5027)	(-0.3422,1.3906)	(0.8274,1.1061)	(0.6795,1.1935)	(0.1685.1.5000)	(0.3529,1.5228)	(0.2603,0.5294)	(0.1824,0.4358)
1.3314	1.1698	0.2690	0.2534	0.5333	1.5530	0.4571	0.8790

Table 8: AE, CI, and AL of θ, β and S(t), t=0.5, 1 Parameters Based on Real dataset II for n=19, m=10, T=7, R=(5,0^9)

MLE				**Bayesian**			
θ	β	**S(0.5)**	**S(1)**	θ	β	**S(0.5)**	**S(1)**
0.4370	1.0630	0.9772	0.9554	1.2704	1.0935	0.3067	0.1783
(-0.2229,0.3103)	(0.2969,1.2561)	(0.7486,1.2058)	(0.5160,1.3950)	(0.9188.1.5265)	(0.9053,1.7482)	(0.9103,0.3543)	(0.1115,0.2101)
0.5330	1.5530	0.4571	0.8790	0.6077	0.5428	0.1641	0.0985

7. Conclusion

In this study, the likelihood and Bayesian approaches were utilized to estimate the parameters of the Lomax distribution and survival function, under an adaptive progressive type-II censored data. However, closed-form solutions for the ML estimators of the parameters and survival function were unavailable, which led to the use of the Newton-Raphson numerical method for computation. Moreover, the study constructed asymptotic confidence intervals for θ and β, and an approximate confidence interval for the reliability function was obtained through the Delta method. The Bayesian approach used in the study employed both informative prior and non-informative prior. However, the Bayes estimates under the squared error loss function could not be derived analytically. As a result, the Metropolis-Hastings algorithm was utilized to generate 10,000 samples for estimation of the two unknown parameters, and credible intervals were computed for these quantities, as well as for the survival function. Furthermore, a simulation study was conducted to investigate the proposed methods for various sample sizes n, effective sample sizes m, and three different progressive censoring schemes, replicated 2000 times. The study also evaluated the proposed methods based on a real-life example. The estimators were observed to have small biases in all situations, indicating approximate unbiasedness. The average length of the estimators decreases with increase in the value of m and n. The MSEs of the estimators decreases with increase in the sample size. Overall, the study suggests that the Bayesian inference approach performs better than the classical approach. In the future endeavours, one could explore these estimation techniques in the presence of explanatory variables and develop more efficient computational algorithms to handle high-dimensional data and complex models. Further studies might also investigate the application of these generalized Lomax models in various domains, such as finance and biomedical sciences, to validate their practical utility.

Acknowledgements

The first author would like to thanks the Department of Science and Technology for providing financial assistance for conducting this work in the form of INSPIRE Fellowship. The authors would like to thank the referees for valuable comments and suggestions which greatly improved the paper.

References

Aldrich, J., (1997). RA Fisher and the making of maximum likelihood 1912-1922. *Statistical science*, 12, no. 3, pp. 162–176.

Altun, E., (2021). The Lomax regression model with residual analysis: an application to insurance data. *Journal of Applied Statistics*, 48, no. 13, pp. 2515–2524.

Balakrishnan, N., Sandhu, R. A., (1995). A simple simulational algorithm for generating progressive Type-II censored samples. *The American Statistician*, 49, no. 2, pp. 229–230.

Balkema, A. A., De Haan, L., (1974). Residual life time at great age. *The Annals of probability*, 2, no. 5, pp. 792–804.

Chen, S., Gui, W., (2020). Statistical analysis of a lifetime distribution with a bathtub-shaped failure rate function under adaptive progressive type-II censoring. *Mathematics*, 8, no. 5, p.670.

Chhikara, R. S., Folks, J. L., (1977). The inverse Gaussian distribution as a lifetime model. *Technometrics*, 19, no. 4, pp. 461–468.

Cramer, E., Schmiedt, A. B., (2011). Progressively Type-II censored competing risks data from Lomax distributions. *Computational Statistics & Data Analysis*, 55, no. 3, pp. 1285–1303.

Cui, W., Peng, X. Y. and Yan, Z. Z., (2019). Bayesian analysis of a constant-stress accelerated life testing with thermal aging life model under general progressive type-II censored data. *Thermal Science*, 23, no. 4, pp. 2509–2516.

Greene, W. H., (2010). Econometric analysis (7th ed.).New York: Pearson.

Ellah, H., (2007). Comparison of estimates using record statistics from Lomax model: Bayesian and non Bayesian approaches. *Journal of Statistical Research of Iran JSRI*, 3, no. 2, pp. 139–158.

Hassan, A., Al-Ghamdi, A., (2009). Optimum Step Stress Accelerated Life Testing for Lomax Distribution. *Journal of Applied Sciences Research*, 5, no. 12, pp. 2153–2164.

Helu, A., (2022). Adaptive type-II hybrid progressive schemes based on maximum product of spacings for parameter estimation of Kumaraswamy distribution. *Applied and Computational Mathematics*, 11, no. 4, pp. 102-–115.

Helu, A., Samawi, H., (2021). Statistical analysis based on adaptive progressive hybrid censored data from Lomax distribution. *Statistics Optimization & Information Computing*, 9, no. 4, p.789.

Khan, Y., Khan, A. A., (2020). Bayesian Analysis of Lomax Family of Distributions Using Simulation and Optimisation. *Global Journal of Pure and Applied Mathematics*, 16, no. 1, pp. 53–77.

Kohansal, A., Shoaee, S., (2021). Bayesian and classical estimation of reliability in a multicomponent stress-strength model under adaptive hybrid progressive censored data. *Statistical Papers*, 62, no. 1, pp. 309–359.

Lomax, K. S., (1954). Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association*, 49, no. 268, pp. 847–852.

Nassr, S. G., Almetwally, E. M. and El Azm, W. S. A., (2021). Statistical inference for the extended Weibull distribution based on adaptive type-II progressive hybrid censored competing risks data. *Thailand Statistician*, 19, no. 3, pp. 547–564.

Nelson, W., (1982). *Lifetime data analysis*.

Ng, H. K. T., Kundu, D. and Chan, P. S., (2009). Statistical analysis of Exponential lifetimes under an adaptive Type-II progressive censoring scheme. *Naval Research Logistics (NRL)*, 56, no. 8, pp. 687–698.

Okasha, H., Lio, Y. and Albassam, M., (2021). On reliability estimation of Lomax distribution under adaptive type-i progressive hybrid censoring scheme. *Mathematics*, 9, no. 28, p. 2903.

Panahi, H., Asadi, S., (2021). On adaptive progressive hybrid censored Burr type III distribution: Application to the nano droplet dispersion data. *Quality Technology & Quantitative Management*, 18, no. 2, pp. 179–201.

Sewailem, M. F., Baklizi, A., (2019). Inference for the log-logistic distribution based on an adaptive progressive type-II censoring scheme. *Cogent Mathematics & Statistics*, 6, no. 1, p. 1684228.

Ye, Z. S., Chan, P. S., Xie, M. and Ng, H. K. T., (2014). Statistical inference for the extreme value distribution under adaptive Type-II progressive censoring schemes. *Journal of Statistical Computation and Simulation*, 84, no. 5, pp. 1099–1114.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1 pp. 69–91, https://doi.org/10.59139/stattrans-2025-005
Received – 30.01.2024; accepted – 25.10.2024

A fuzzy hybrid MCDM approach to the evaluation of subjective household poverty

Aleksandra Łuczak [1], **Sławomir Kalinowski** [2]

Abstract

Poverty is one of the most important global socio-economic problems. Despite a strong interest in this phenomenon, there is no unified concept for measuring it. It is difficult to quantify due to the diversity of the dimensions of perceived poverty, particularly subjective ones. Thus, the aim of the research described in the article is to propose a comprehensive procedure for constructing a synthetic measure of subjective poverty in households. This involves aggregating factors describing the present, future, and past, which make it easier to grasp the feeling of deprivation. Methods such as fuzzy TOPSIS and fuzzy hierarchical analysis (FHA) based on the fuzzy sets theory were used for this purpose, which is not standardly used for this type of research. This innovative procedure was applied to assess the level of subjective household poverty in Poland. The analyses are based on data from primary research carried out in three stages in 2020 using the CAWI method. The results show that the assessment of households' current socio-economic situation is also influenced by past events as well as projections of future developments. Changes in the values of the synthetic index illustrate the trajectory of switching from panic to negation, and attempting to cope with the situation or, alternatively, switching to a state of irritation. The research results can form the basis for formulating policies and strategies to combat poverty.

Key words: fuzzy TOPSIS, fuzzy hierarchical analysis (FHA), MCDM, subjective poverty, household, CAWI

1. Introduction

Due to its interdisciplinary nature, poverty is a specific research category. Understanding its specificity requires various scientific disciplines – economics (including behavioral), sociology, social policy, or psychology. The considerations of poverty highlight that it is the result of many overlapping social and economic difficulties, including the lack of work, low income, dysfunctions, limited opportunities or low human capital. Schiller (1989) points to three causes: flawed character, restricted

[1] Faculty of Economics, Poznań University of Life Sciences, Poznań, Poland.
E-mail: aleksandra.luczak@up.poznan.pl. ORCID: https://orcid.org/0000-0002-3149-7748.
[2] Institute of Rural and Agricultural Development, Polish Academy of Sciences, Warsaw, Poland.
E-mail: skalinowski@irwirpan.waw.pl. ORCID: https://orcid.org/0000-0002-8068-4312.

opportunities and inefficient state policy, which Schiller describes as Big Brother. Bradshaw (2007) suggests that it is the "effect of individual deficiencies, cultural belief systems that support subcultures in poverty, political-economic distortions, geographic disparities, or cumulative and circumstantial origins". Given the wide range of causes of poverty, it can be assumed that it is an anomic feature of the contemporary world. Although there are many causes (Brandt, 1908; Thurow, 1967; Shaw, 1996; Jennings, 1999; Dudek, 2008; Dudek & Szczesny, 2021; Brady, 2019; Kalinowski, 2020), the problem of the COVID-19 pandemic and its negative effects on the functioning of households seems to have been the most important in recent years (Kalinowski & Wyduba 2020; Gupta et al., 2021; Asfaw, 2021).

Although 120 years have passed since Rowntree's first poverty research (1901), there is still no unified definition. The concept of poverty is unclear, which makes it difficult to define it (Blank, 1997; van Praag at al., 2008), as a result of which there is also no generally accepted method of measuring it (Kalinowski, 2015). In most research into poverty, a person is classified as poor if he or she lacks sufficient resources to achieve an acceptable standard of living. Usually the analysis is limited to economic deprivation and distress. However, as Shaw (1996) and Blank (2003) (among others) point out, poverty is a very complex social problem with many variants and roots, all of which are important depending on the situation. The very attempt to define poverty itself is a consequence of research traditions resulting from the overlapping of behavioral, social and economic factors, reinforced by political considerations.

The essence of poverty is inequality (Valentine, 1968). It can be reflected both in unequal income and consumer spending, as well as in the level of perceived needs and the way in which they are perceived. Thus it can be assumed that inequality in terms of perceived needs may favor various levels of satisfaction, regardless of the objective dimension of satisfying the needs. The amount of funds held cannot reflect satisfaction. It can be assumed after Ahuvia (2008) that the chances of determining an individual's situation are greater when knowing the evaluation of satisfaction with life as a whole rather than by knowing the level of income. Thus the objective dimension expressed in income or expenditure will not be reflected in the subjective satisfaction with the various dimensions of life (cf. Easterlin, 1974; Nettle, 2005; Rayo & Becker, 2007; Michoń, 2010).

Since the objective dimension is not sufficient to describe multidimensional poverty, we have chosen to redefine subjective poverty. "We assumed that this is an awareness of the lack of sufficient resources to meet one's needs in terms of socio-economic status (income and current financial situation, level of education and occupation, residence, lifestyle and leisure) and one's own aspirations to achieve and maintain the desired standard of living" (Łuczak & Kalinowski, 2022). We recognized that to some extent subjective poverty is a consequence of the emphasis on relative deprivation of needs discussed by Townsend (1979) and Runciman (1966). We assumed after Townsend that poverty is an inability to meet the standards of a given

society. Although Townsend's definition refers to a relative dimension, it is simply reflected in individuals' subjective expectations, especially given their aspirations.

It is worth noting that the definition of subjective poverty that we adopted limits the fraction of poor people only to those who have a feeling of unmet needs, while leaving out those who do not have this feeling. In conceptualizing subjective poverty, we thus found that behavioral factors are extremely important. This emphasis allows us to assume that subjective poverty is influenced by the respondents' circumstances. We assumed that the sense of poverty is influenced by the feeling of deprivation in relation to the environment, i.e., the situation of the surveyed individual and how he or she perceives his or her own well-being. To quote John Stuart Mill (1907), "Men do not desire merely to be rich, but to be richer than other men." This relativism of thinking at the same time encourages the formation of subjective assessments of one's own position in relation to the environment. A question arises – what is this environment? Who is this benchmark for respondents' assessments? Without much error, it can be assumed that they are people closely related to the respondents (family, friends) or other people they know (neighbors, co-workers). However, without being sure of who constitutes the comparison group, one should be cautious in this regard.

According to Haveman (2015), "the process of measuring poverty and analyzing its causes and consequences has advanced social science research in several areas, including identifying the underlying causes of poverty, understanding social mobility, attainment, and income dynamics, and measuring the behavioral effects of antipoverty policy interventions." A problematic issue in all the measures indicated is the feeling of deprivation of needs in relation to expectations and, consequently, the estimation of one's own line of prosperity. This leads to measurement errors. In deciding to create a synthetic measure, we therefore wanted it to be the result of the evaluation of financial situation and material conditions of one's own household, as well as the perception of one's own income compared to the income of other households. We also wanted the proposed measure to be based on a subjective sense of the standard of living of household members and a sense of helplessness against the risk of poverty. We believe that it is not only the moment of the pandemic that is important, but also the past situation and the anticipation of future changes. It should be emphasized that our innovation in research consists in the use of the time dimension in research, including the past, present, and future. We propose a procedure for constructing a synthetic measure based on repeated surveys (in this case from three periods) conducted using the CAWI method. The comprehensive procedure we propose is a hybrid MCDM approach based on fuzzy methods that extend the approach proposed by Łuczak and Kalinowski (2022). The key elements of the methodology, i.e. determination of the indicator-weighting system and calculation of synthetic measures, are based on the fuzzy hierarchical analysis (FHA) and the fuzzy technique for order of preference by similarity to ideal solution (FTOPSIS), respectively. In addition, we propose our own compactness measure to examine the homogeneity of the created groups of objects.

As the main objective of the research, we adopted a presentation of a unconventional procedure for the construction of a synthetic measure of subjective household poverty in the context of poverty types and household types based on a hybrid multi-criteria decision-making approach in a fuzzy environment. The proposed approach was used to study the perception of subjective poverty by households in Poland during the COVID-19 pandemic. The research was carried out on the basis of three-stage primary research in April, June and September 2020. This paper consists of five parts. In addition to the introduction, section 2 provides a detailed description of the proposed multiple-criteria decision-making method. Section 3 describes the results of empirical research on the evaluation of subjective household poverty in Poland during the COVID-19 pandemic. Chapter 4 discusses the proposed research procedure and the results obtained. The conclusions are presented in Chapter 5.

2. Literature review

Poverty is a multidimensional phenomenon, the definition and measurement of which raises a lot of controversy and discussion. In research on poverty, the lines of poverty separating relatively well off (non-poor) people from poor people are most often used (Golinowska 1997, Broda-Wysocki 2012). They are criticized in many studies because they cause a dichotomous division of society. Generally, there are two approaches to determining the poverty line – economic and multidimensional (Figure 1). The objective approach is determined both on the basis of normative and parametric lines. The first are absolute, while the second are relative. Determining the normative lines consists in determining the value of income necessary to satisfy a certain group of needs (Booth 1889, Rowntree 1901, Orshansky 1969). They are based on various types of standards (expert or political) regarding the fulfillment of needs (Kalinowski, 2015).

Relatively the least important in the measurements is the poverty threshold based on official lines. Its minor importance results, on the one hand, from a certain underestimation, and on the other from overestimation. This is due to several factors (Kalinowski, 2015):

1) lowering the statistics contributes to the apparent reduction of the poverty threshold without its actual elimination, which may lead to a lack of valorization of the number entitled to receive benefits,
2) for fear of being stigmatized some people consciously do not want receive social welfare benefits, thus they are not included in the assistance systems, and as a result they are not treated as poor, even though they cannot meet their needs,
3) some people receive benefits, although they are not formally entitled to them (e.g. working illegally),
4) lack of international comparability.

The subjective measures of poverty are also important (cf. Hagenaars, van Praag 1985, Kapteyn, van Praag, van Herwaarden 1978, Goedhart et al. 1977). These are

considered the most democratic methods of defining poverty, which results from the individual setting of the limit of deprivation.

Measurement of poverty is often limited to objective, one-dimensional indicators (e.g. income or expenses). However, when assessing poverty, its subjective dimension is also important, as it shows the perceptions of the poor. The growing contrast between the rich and the poor only increases the level of feeling poverty. There are many levels of poverty, from no poverty to extreme poverty. It should be noted that poverty is not always immediately noticeable, and those that are visible are not always felt by the respondents. Hence the problem of subjective poverty measurement is important, as it identifies various degrees of poverty perception among respondents and often depends on the point of reference (on the people to whom the respondents compare themselves, e.g. family, friends, neighbors). For these reasons, research on the measurement of subjective poverty was undertaken. The study of subjective poverty allows for the identification of the diversity of the respondents' perceptions of poverty.

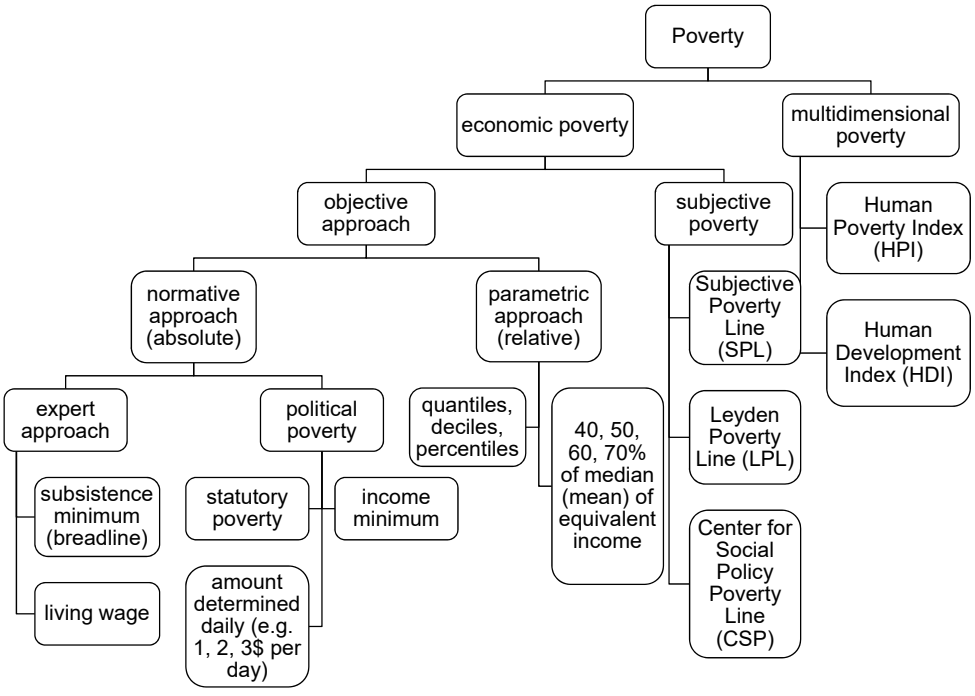


Figure 1: Methods of determining the poverty.
Source: Kalinowski (2015).

Existing definitions of poverty are characterized by a high degree of subjectivity and individual interpretation by individual researchers. This is why some of them have a broad scope, others are narrower. Due to this, in many cases, it is difficult to make comparisons because adopting a different understanding of the definition often means that the researcher had a research sample that was different in terms of quality.

Nevertheless, in many cases, one can note that despite the differences in the approach to particular definitions, the core is similar and many elements remain common (Kalinowski, 2015). Thus we defined subjective poverty as a conscious sense of the lack of sufficient resources to meet one's needs in relation to the "socioeconomic status (income and current financial situation, level of education and profession, place of residence, lifestyle and leisure activities) and one's own aspirations to achieve and maintain the desired standard of living" (Łuczak & Kalinowski, 2022).

To complete the picture of measurement of poverty, it is necessary to add an observation of problems that need to be taken into account when assessing subjective poverty. They are related to the selection of variables, survey design, measurement errors, frames of reference, idiosyncratic characteristics of respondents, and differences in their personality and tastes (Ravallion & Lokshin, 2002; Ravallion, 2012; Ravallion et al., 2016). Some of these can be solved by conducting research which is well grounded in theory and practice. However, some of them are unmeasurable and elusive in nature, regardless of the research procedure adopted.

We would like to emphasize the fact that objective and subjective dimensions of poverty are equally important, just as in well-being analyses (cf. Stiglitz et al., 2009). Instead of treating them as substitutes, they should be regarded as complementary. The picture of reality should be created by juxtaposing various approaches. Only then will it be possible to draw the correct conclusions.

3. Methodological approach

There are different approaches to assessing poverty based on fuzzy sets theory (Cerioli & Zani, 1990; Chiappero-Martinetti, 1994; Betti et al., 2008; Montrone et al., 2010; Belhadj, 2011; Neff, 2013; Betti et al., 2017; Belhadj & Limam, 2012; Aristondo & Ciommi, 2017; Ciani et al., 2019). However, our proposed composite-index approach goes far beyond what has been proposed so far, describing the subjective evaluation of household poverty as a multi-dimensional self-evaluation of respondents using multiple-criteria decision-making methods i.e. the fuzzy hierarchical analysis (FHA) and the fuzzy technique for order of preference by similarity to ideal solution (FTOPSIS). In this paper we introduce the time dimension to the poverty measure and propose a triple reference-point approach. This is based on the respondents' past, present and future feelings. Each step of the proposed procedure is described in detail below.

Step 1: Preparation of and conducting a survey on subjective poverty. In this step, we assume that households are characterized by three criteria: perceptions of the present situation, perception of the past, and future projections. In typical measures of subjective poverty, participants are asked to assess their financial situation or standard of living in relation to other families. Individual prosperity lines are constructed on this basis. Without going into the details of the creation of these lines, they can be reduced

to a number of commonly used ways of measuring poverty on the basis of subjective perceptions. They are related to:

a) an evaluation of one's own income situation (Hagenaars, van Praag, 1985),
b) a feeling of being poor – a minimum income (Kepteyn et al. 1988),
c) evaluation of one's life in verbal terms, e.g.: "very bad", "bad", "sufficient" and "good", "very good" (Van Praag, 1971, Van Praag et al. 1980). Such questions treat poverty as a more general concept than just income poverty and often approach terms such as subjective well-being, satisfaction with life and happiness,
d) assessment of the possibility of "making ends meet" (often referred to as the Deleeck question) or difficulties in making the necessary payments (Deleeck & Van Den Bosch, 1990; Ghiatis, 1990).

On one hand, households' perception of their own poverty may affect self-evaluation in the future, even if objective poverty decreases. On the other hand, previous experience of poverty may also result in a household currently having a sensation of a higher level of income than it actually does and vice versa (Ravallion & Lokshin, 2002). Thus the hysteresis in the perception of subjective poverty by households occurs. It should be added that the perceived condition of the household is also influenced by the actual dynamics of poverty (Alem et al., 2014).

Each of these criteria is described by k_i (i= 1, 2, 3) indicators, $k = k_1 + k_2 + k_3$. Households are subject to self-evaluation within each indicator using an ordinal measurement scale and verbal descriptions. The measurement scales used in the study have m_j categories ($j = 1, 2, \ldots, k$), where 1 is the most optimistic response in relation to the criterion of subjective poverty, and m_j is the most pessimistic. In other words, the higher the evaluation, the worse the perception with regard to the level of subjective poverty. So there are $(m_j - 1)/2$ positive and negative responses. In the case of an ordinal scale with inverted categories, these should be re-coded to the form described above.

Step 2: The selection of indicators of subjective poverty. A set of indicators[3] (attributes) is used to describe subjective poverty, characterizing it in terms of: an assessment of the financial situation and material conditions of the household, the perception of one's own income against the income of other households, the household's standard of living, feeling helpless in the face of poverty.

The collected indicator values are summarized in the data matrix:

$$\mathbf{X} = \left[x_{ij}\right] \tag{1}$$

where: x_{ij} – is the value of j-th indicator in i-th household, $i = 1, \ldots, n$; n – the number of households; $j = 1, \ldots, k$, k – the number of indicators.

Step 3: Determination of the nature of the indicators in relation to the main criterion. The direction of indicator preferences in relation to the criterion in question

[3] An indicator (variable) is a quantitative or a qualitative measure that can show value of characteristics or their level for an objects. On the other hand, aggregated indicators are an index.

is determined, i.e. their division into benefit and cost indicators. A benefit indicator contributes to increasing the level of a phenomenon, whereas a cost indicator is a variable that reduces the level of that phenomenon. We assumed that all indicators were benefit indicators, because when measuring complex phenomena (i.e. the level of subjective poverty) using surveys, the criteria are usually selected so that they are positively correlated with the phenomenon (the higher the evaluation of an indicator, the higher the level of subjective poverty)

Table 1: Formulas for determining the parameters of triangular fuzzy numbers.

Categories	Triangular fuzzy number parameters		
	a_{ij}	b_{ij}	c_{ij}
1	0	0	$1/[2(m_j-1)]$
2	$1/[2(m_j-1)]$	$1/(m_j-1)$	$3/[2(m_j-1)]$
…	…	…	…
m_j-1	$(2m_j-5)/[2(m_j-1)]$	$(m_j-2)/(m_j-1)$	$(2m_j-3)/[2(m_j-1)]$
m_j	$(2m_j-3)/[2(m_j-1)]$	1	1

Step 4: Conversion of ordinal categories of indicators to triangular fuzzy numbers. Indicator variants are transformed into triangular numbers $(a,\ b,\ c)$ in the form of three evaluations (parameters). Table 1 shows the formulas for determining the parameters of triangular fuzzy numbers. The parameters of triangular fuzzy numbers can be scaled by a selected fixed value freely determined by the researcher. The triangular fuzzy numbers obtained are presented in the form of fuzzy data matrix:

$$\widetilde{\mathbf{X}} = \left[\tilde{x}_{ij}\right] \tag{2}$$

where: $\tilde{x}_{ij} = \left(a_{ij}, b_{ij}, c_{ij}\right),\ i = 1,\ \ldots,\ n;\ n = n_1 + n_2 + n_3; n_1,\ n_2,\ n_3$ – number of households in stages I, II and III respectively; $j = 1,\ \ldots,\ k,\ k$ – number of indicators.

Step 5: Determination of the indicator-weighting system. One of the most commonly used methods of determining the weighting system is equal treatment of all indicators (Aaberge & Brandolini, 2015). This is the case, for example, with the Human Development Index. However, it should be noted that indicators under each criterion have different impacts on the level of subjective poverty, so a differentiated indicator-weighting system should be introduced. In our research, we used one version of the fuzzy analytical hierarchical process – the Fuzzy Hierarchical Analysis (FHA) – to determine the weighting system. This is an extension of the analytical hierarchical process (AHP) and also applies when there are difficulties in presenting the evaluations of comparisons of pairs of elements in the hierarchy in the form of real numbers. In our paper, we calculated the weighting system $\widetilde{w}_j = \left(w_j^L,\ w_j^M,\ w_j^U\right), j = 1, \ldots, k; k = k_1 + k_2 + k_3$ using fuzzy hierarchical analysis (see Csutora & Buckley 2001, Buckley et al. 2001, Łuczak & Wysocki 2008).

Step 6: Normalization of indicator values. Normalization of indicators with a nature of stimulants:

$$\tilde{z}_{ij} = \left(a_{ij}^{(z)}, b_{ij}^{(z)}, c_{ij}^{(z)}\right) = \left(\frac{a_{ij}}{c_j^+}, \frac{b_{ij}}{c_j^+}, \frac{c_{ij}}{c_j^+}\right) \qquad (i = 1, 2, \ldots, n;\ j \in P_s) \quad (3)$$

where $c_j^+ = \max_i(c_{ij})$, $c_j^+ \neq 0$; P_s – a set of stimulant indices.

for the destimulants:

$$\tilde{z}_{ij} = \left(a_{ij}^{(z)}, b_{ij}^{(z)}, c_{ij}^{(z)}\right) = \begin{cases} \left(\frac{a_j^-}{c_{ij}}, \frac{a_j^-}{b_{ij}}, \frac{a_j^-}{a_{ij}}\right) \text{for } a_{ij}, b_{ij}, c_{ij} \neq 0 \\ (0,0,0) \qquad \text{for } a_{ij}, b_{ij} = 0 \end{cases} \quad (4)$$

$$(i = 1, 2, \ldots, n;\ j \in P_D)$$

where $a_j^- = \min_i(a_{ij})$; P_D – a set of destimulant indices.

Structure of the weighted normalized fuzzy data matrix:

$$\widetilde{\mathbf{R}} = [\tilde{r}_{ij}] \quad (5)$$

where $\tilde{r}_{ik} = \tilde{z}_{ij}(\cdot)\widetilde{w}_j = \left(a_{ij}^{(z)}, b_{ij}^{(z)}, c_{ij}^{(z)}\right)(\cdot)\left(w_j^L,\ w_j^M,\ w_j^U\right) =$

$= \left(a_{ij}^{(z)} w_j^L, b_{ij}^{(z)} w_j^M, c_{ij}^{(z)} w_j^U\right) = \left(a_{ij}^{(r)},\ b_{ij}^{(r)}, c_{ij}^{(r)}\right)$, $(\cdot)$ is the fuzzy numbers multiplication.

Step 7: Calculating the pattern and antipattern. Determination of a fuzzy pattern $\tilde{A}^+$ (cf. Hwang & Yoon 1981, Chen 2000):

$$\tilde{A}^+ = \left(\max_i(\tilde{r}_{i1}), \max_i(\tilde{r}_{i2}), \ldots, \max_i(\tilde{r}_{ik})\right) = (\tilde{r}_1^+, \tilde{r}_2^+, \ldots, \tilde{r}_k^+) \quad (6)$$

where $\tilde{r}_j^+ = \left(a_{ij}^{(r)+},\ b_{ij}^{(r)+}, c_{ij}^{(r)+}\right),\ j = 1,\ \ldots,\ k.$

and fuzzy antipattern $\tilde{A}^-$:

$$\tilde{A}^- = \left(\min_i(\tilde{r}_{i1}), \min_i(\tilde{r}_{i2}), \ldots, \min_i(\tilde{r}_{ik})\right) = (\tilde{r}_1^-, \tilde{r}_2^-, \ldots, \tilde{r}_k^-) \quad (7)$$

where $\tilde{r}_j^- = \left(a_{ij}^{(r)-},\ b_{ij}^{(r)-}, c_{ij}^{(r)-}\right),\ j = 1,\ \ldots,\ k.$

Step 8: Calculation of the distance of each object from the pattern and antipattern. Calculation of the distance between fuzzy indicator values for the evaluated objects and the pattern is performed using the following formula (Chen 2000):

$$d_i^+ = \sum_{i=1}^k \sqrt{\frac{1}{3}\left[\left(a_{ij}^{(r)} - a_{ij}^{(r)+}\right)^2 + \left(b_{ij}^{(r)} - b_{ij}^{(r)+}\right)^2 + \left(c_{ij}^{(r)} - c_{ij}^{(r)+}\right)^2\right]} \quad (8)$$

and from the antipattern:

$$d_i^- = \sum_{i=1}^k \sqrt{\frac{1}{3}\left[\left(a_{ij}^{(r)} - a_{ij}^{(r)-}\right)^2 + \left(b_{ij}^{(r)} - b_{ij}^{(r)-}\right)^2 + \left(c_{ij}^{(r)} - c_{ij}^{(r)-}\right)^2\right]} \quad (9)$$

Step 9: Calculation of synthetic measures of the level of subjective poverty for households at different research stages.

Calculation of the value of the synthetic measure (index) for each household i = 1, 2, …, n using the following formula of TOPSIS (Hwang and Yoon 1982):

$$S_i = d_i^- / (d_i^+ + d_i^-) \quad (10)$$

The higher the value S_i, the higher the level of subjective poverty of the household. The measure S_i is normalized to the range [0,1] and S_i becomes 0 for the antipattern object and 1 for the pattern object.

Step 10: Identification of subjective poverty types for households according to selected criteria and research stages.

Averaging of the standard values within the researched criteria:

$$S_i^{csv} = \underset{i \in P_{csv}}{\text{med}}(S_i) \tag{11}$$

where P_{csv} – a set of household indices within the s-th category of the c-th criterion at the v-th stage of survey ($v = 1, 2, 3$). Three categories were adopted: for the whole country, divided into village and city, or village, small town with less than 20,000 residents, urban area with 20,000-99,000 residents, urban area with 100,000-499,000 residents, urban area with 500,000 or more residents.

Table 2: Subjective poverty index values and theoretical types of poverty – poverty profiles

S_i^{csv}	Level of index	Type of household poverty	Type of household
[0.00; 0.10)	very extreme low	no poverty	prosperous
[0.10; 0.20)	extremely low	very mild poverty	
[0.20; 0.30)	very low	at risk of poverty	relatively prosperous/ coping/ resourceful
[0.30; 0.40)	low	indistinct poverty	
[0.40; 0.50)	medium-low	moderate low poverty	endangered by poverty
[0.50; 0.60)	medium-high	moderate high poverty	
[0.60; 0.70)	high	strong advancing poverty	poor
[0.70; 0.80)	very high	severe poverty	
[0.80; 0.90)	extremely high	very severe poverty	extremely poor
[0.80; 1.00]	very extreme low	utter poverty	

Source: own elaboration.

The identification of subjective poverty level types can be carried out arbitrarily. Theoretical (hypothetical) poverty types – poverty profiles (Table 2) were also proposed on the basis of synthetic measure S_i^{csv}. Poverty is not dichotomous; households cannot be divided into poor or non-poor. There are many shades within the limits of the lack of poverty up to extreme poverty. Households may therefore be characterized by various levels of poverty (cf. Cerioli & Zani, 1990; Betti et al., 2008; Montrone et al., 2010; Belhadj & Limam, 2012; Ciani et al., 2019).

The authors' indicators of LK_{csv} compactness were also calculated as part of the s-th category of the c-th criterion at the v-th stage of survey:

$$LK_{csv} = \frac{\sum_{i=1}^{n_{csv}} \left(S_i^{csv} - \underset{i \in P_{csv}}{\text{med}}(S_i) \right)}{n_{csv} \cdot \max_i \left(1 - \underset{i \in P_{csv}}{\text{med}}(S_i); \underset{i \in P_{csv}}{\text{med}}(S_i) \right)} \tag{12}$$

where n_{csv} – the number of households within the s- th category of the $(s = 1, 2, 3)$ c-th criterion $(c = 1, .., n_s)$ at the stage v $(v = 1, 2, 3)$. The indicators are normalized within the range [0, 1]. The lower the measure of the compactness index, the more homogeneous is the group. The degrees of compactness according to the gradation given in Table 3 can be assumed.

Table 3: Degrees of compactness

LK_{csv}	[0.00; 0.20)	[0.20; 0.40)	[0.40; 0.60)	[0.60; 0.80)	[0.80; 1.00]
Degree of compactness	very high	high	medium	low	very low

Source: own elaboration.

4. Conducting research and results

The analyses used data from primary household research in Poland, during which the CAWI (Computer-Assisted Web Interview) method was used. The research was conducted in three stages: April, 2020 (1st research stage), June, 2020 (2nd research stage), September, 2020 (3rd research stage). In each of the three stages, the sample is a quota sample according to the key size category of the place of usual residence and covers 458 households.

The research included variables describing the subjective situation of households according to three criteria:

- perceptions of the present situation: feeling of being satisfied with life (x_1), degree of present satisfaction of household needs through income earned (x_2), evaluation of household income compared to other households, evaluation of the change in food needs during the pandemic period compared to previous years (x_3), evaluation of own household situation (x_4), whether it is possible to "make ends meet" with current income (x_5),
- future projections: perception of the degree of possibility of deterioration of one's own household's situation in the near future (x_6), feeling concerning the degree of potential for loss of income (x_7), perception of the degree of potential loss of financial stability (x_8), perception of the degree of possibility of losing work (x_9), evaluation of the possibility of change in one's own household's financial situation within the next 12 months (x_{10}),
- perceptions of past situations: degree of satisfaction of one's own household's needs through income (before the epidemic) (x_{11}), past feelings of being poor (x_{12}).

The variables adopted for the research define three unique time dimensions, not taken into account in research on poverty, in which it manifests itself, i.e.: the past (past fillings), the present (current subjective state) and the future (perceptions of future projections). We assumed that all indicators were stimulants. We assumed this because when measuring complex phenomena (i.e. the level of subjective poverty) by surveys,

usually the criteria are selected in such a way that they are positively correlated with this phenomenon. The higher the partial assessment, the higher is the level of subjective poverty. We adopted a system of differentiated fuzzy weights for indicators (Table 4).

Table 4: Fuzzy weights system for indicators

Indicator category	Indicators	Triangular fuzzy number		
		a	*b*	*c*
Perceptions of the present situation	$x_1 - x_5$	0.079	0.155	0.269
Future projections	$x_6 - x_{10}$	0.014	0.027	0.054
Perception of the past situation	x_{11}, x_{12}	0.043	0.045	0.097

As shown in Figure 2, the levels of perceived poverty at the different stages of the research suggest that there has been a shift from panic to adaptation. Figure 3 shows box-plots for levels of subjective household poverty, in which even greater disparities can be observed in the evaluation of subjective poverty between the 1st and 2nd and 3rd stages of the research; a relatively large increase in optimism can be observed in Poland between the 1st and 2nd stages of the research (the decrease in the index value from 0.387 to 0.354). At the third stage, despite the increase in disease incidence, the subjective evaluation of poverty remained almost unchanged (0.348). This may indicate that constant fear stimulation has become a factor of coronavirus becoming more common. Effective metaphors, war comparisons or post-apocalyptic language have become an adaptive factor to a "new normality" (Kalinowski, 2020a). This weakened the negative perception of one's own socio-economic situation.

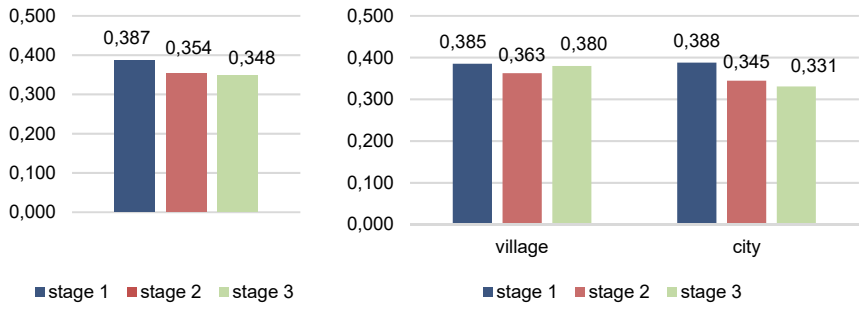


Figure 2: Levels of subjective household poverty by research stages and division into the village and the city.

Although, as indicated in public discourse, COVID-19 is treated more as an urban disease, studies indicate that it is a reason for rural residents' unfavorable assessment of their own situation to a greater extent. Although in the first stage, poverty perceived among rural residents (0.385) was almost at the same level as among urban residents (0.388), in subsequent stages, stratification to the disadvantage of rural areas occurred (Figure 2). As many studies show, it is rural areas that suffer greater economic and social consequences of poverty.

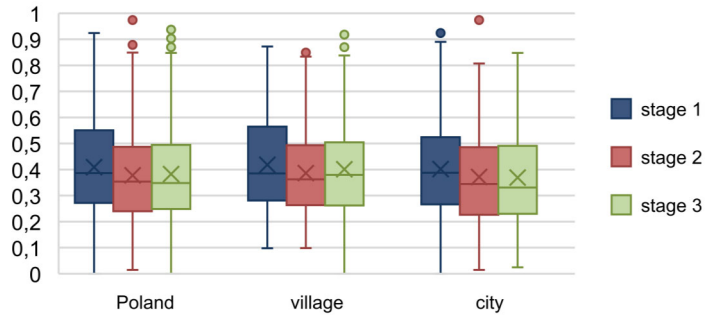


Figure 3: The box-plot for levels of subjective household poverty by research stages and division into village and city

Note: A box based on: median, and the first and third quartiles. Above the third quartile, a distance of 1.5 times the interquartile range (IQR) is measured and a whisker is drawn to the largest observed point from the set of data that falls within this distance. Similarly, a distance of 1.5 times the IQR is measured below the lower quartile, and the whisker is drawn to the bottom observed point from the set of data that falls within this distance. All other observed points are plotted as outliers.

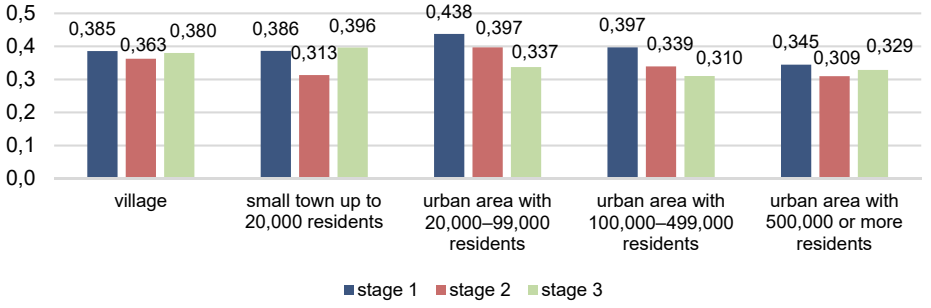


Figure 4: Levels of subjective household poverty by research stages and the class of the locality of the household head

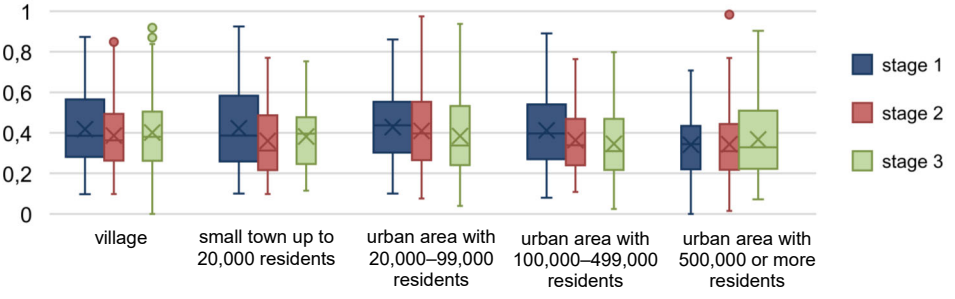


Figure 5: Box plot for levels of subjective household poverty by research stages and class of the locality of the household head

It is worth noting that the village and city categories are a certain mental construct. Just as there is no one village (Stanny et al., 2018), it is difficult to speak of a unified city.

It is worth noting that residents of small towns (up to 20,000 residents) and urban areas with 20,000-100,000 residents evaluate the level of poverty much below residents of medium or large cities (over 100,000 residents). The level of subjective poverty of small towns and villages was similar at all stages of the research (Figure 4). Interestingly, in the largest cities, the poverty-perception level increased again at the third stage. On the one hand, this may result from the ongoing lockdown, but also from rising expectations and discouragement, which fostered negative evaluations during surveys. However, despite a fairly significant increase in negative evaluation, the largest cities – next to the medium-sized ones – were still at the lowest risk of subjective poverty (Figures 4 and 5).

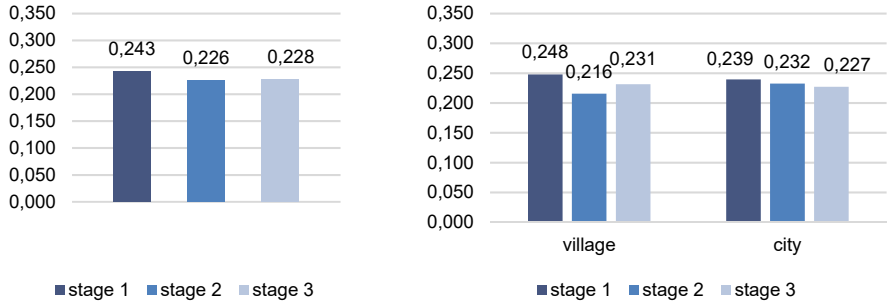


Figure 6: Compactness indices (*LK*) of the synthetic measure of subjective household poverty in Poland by research stages and rural-urban division

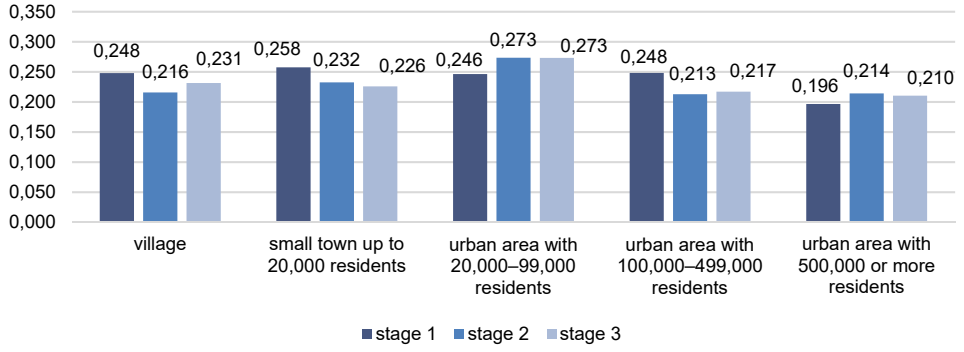


Figure 7: Compactness indices *(LK)* of locality classes in terms of a measure of the subjective household poverty level by stages

It is also worth mentioning that the groups of areas studied were characterized by high compactness of synthetic measures (Figures 6 and 7), as evidenced by the values of the LK index, which ranged from 0.193 (for large cities with more than 500,000 residents at stage I of the research) to 0.277 (for cities with 20,000 to 99,000 residents at stages II and III of the research) (Figure 7). This confirms that large cities with more than 500,000 residents are more homogeneous in their perception of poverty than smaller cities.

5. Conclusion

The multiplicity of subjective poverty indicators raises the following question – what is the purpose of establishing an additional synthetic measure of it? In our view, decomposing poverty using self-assessments of unidimensional indicators and then constructing new synthetic measure is justified for several reasons. First, this makes it possible to show the impact of many factors on the changes in the socioeconomic situation of the population, especially during epidemics. Changes of the synthetic index illustrate the trajectory of switching from panic to negation or trying to cope with the situation or alternatively switching to the state of irritation. Second, the proposed synthetic measure takes account of several overlapping factors related to both income security, deprivation, job security and the expectation of changes in them in the future, all of which are extremely important for assessing subjective poverty. Third, in the analysis of this index we took into account the fact that the assessment of one's own situation is influenced by expectations and aspirations. The index is therefore designed to take these aspects into account as well. Fourth, realizing that the current assessment of one's own socioeconomic situation is also influenced by past events, as well as predictions of changes in that situation in the future, we also took these into account.

Given the above aspects of perceived subjective poverty, our proposed synthetic measure allows us to easily compare various aspects of subjective poverty during the periods studied. The number of variables offsets the risk that a change in one factor will significantly alter the entire index. At the same time, the total level of perceived subjective poverty is affected by a number of variables that amplify or offset its magnitude.

By constructing the synthetic index, we would like to show that the measurement of poverty is a complex issue. Our contribution to research into poverty is to show that the synthetic measure capturing factors combining the future, present and past makes it easier to grasp the feeling of deprivation. It is useful for studying changes in the level of poverty perception over time under the influence of unpredictable phenomena, in this case, during the coronavirus period, without going into detail about the factors causing it. The proposed procedure could be used for conducting official statistics with regularly repeated surveys.

The indicators used are static. Both the LPL (Leyden Poverty Line) and the SPL (Subjective Poverty Line) or the CSP (The Center for Social Policy Poverty Line) are based on individual welfare lines, defining the situation at a given point in time. Such estimates ignore projections both of the future situations and take limited account of events from the distant past. It should be emphasized here that our innovative methodology for constructing a subjective measure of poverty takes into account indicators describing the past, present, and future. In addition, the previously mentioned indicators existing in the literature also have the disadvantage that a significant segment of the population cannot estimate the income that separates the

poor from the wealthy or the income that allows them to live at a certain acceptable level. Individual prosperity lines also have the disadvantage of focusing solely on income while ignoring behavioral aspects, or those related to the socioeconomic environment.

Knowledge of subjective poverty makes it possible to define the elements that influence the sense of poverty. It helps to bridge the gap between its objective and subjective dimensions. In the subjective dimension, the research also makes it possible to pay attention to the nature of inequality. Our study is in line with Aristondo and Ciommi's (2017) observation that "the recent literature on poverty measurement stresses the importance of an index to take into account intensity, incidence and inequality." By emphasizing subjective poverty, we wanted to highlight the importance of maximizing individual wealth, because, as Pouw (2020) argues, it translates into an increase in the prosperity of society as a whole. It is also worth adding, quoting Mowafi (2004), that "studies can only be justified if their conclusions are conscientiously used to inform the development of an adequate and accurate definition of poverty – a definition that not only withstands the rigors of science, but also reflects the realities of the poor."

To summarize our discussion of the construction of a measure of subjective household poverty, several facts should be noted. First, using a fuzzy approach to assessing subjective poverty allows us to identify individual indicators more precisely than with a standard poverty measurement. To the best of our knowledge, nearly all existing approaches to studying household poverty self-assessment are based on a dichotomous division of respondents into poor or non-poor. The advantage of our method is to determine the degree of poverty of the households studied. For these reasons, our work goes beyond a conventional poverty study. We confirm the opinion of Betti et al. (2017) that "the conventional approach presents a serious limitation: poverty is not an attribute that characterises an individual in terms of its presence or absence, but is rather a predicate that manifests itself in different shades and degrees."

Second, the subjective poverty index that we constructed is an attempt to explain poverty from the perspective of the poor. By estimating the level of subjective poverty for each household studied, the index we propose can be used to create a truly individual measure of poverty, taking account of a multi-faceted perceptions of feelings regarding the household's current situation, but also its past situation and predictions for the future.

Third, the subjective picture of the economic stratification of the population is reflected in the aggregate subjective poverty index for each class of locality. A comparison of the dynamics of population indicators revealed their multidirectional dynamics. This may indicate that either people are gradually getting used to the pandemic and are no longer bothered by it that much, or that they are adapting to the new circumstances.

In conclusion, our methodological proposal opens the door to new opportunities for research and applications of multidimensional subjective poverty. Quantitative measurement of subjective poverty at the micro (household) level is an important tool for evaluating anti-poverty policies. At the same time, research over time helps to explain changes occurring in households. In addition, the subjective poverty index can also be viewed as a measure of vulnerability to poverty and can provide a basis for formulating poverty-alleviation policies and strategies.

References

Aaberge, R., Brandolini, A., (2015). Multidimensional poverty and inequality. In A. B. Atkinson, & F. Bourguignon (Eds.), Handbook of income distribution (Vol. 2, pp. 141–216). Elsevier. https://doi.org/10.1016/B978-0-444-59428-0.00004-7.

Ahuvia, A., (2008). Wealth, consumption and happiness. In A. Lewis (Ed.), The Cambridge handbook of psychology and economic behaviour (pp. 199–226). *Cambridge University Press*.

Alem, Y., Köhlin, G. and Stage, J., (2014). The persistence of subjective poverty in urban Ethiopia. *World Development*, 56, pp. 51–61. https://doi.org/10.1016/j.worlddev.2013.10.017.

Aristondo, O., Ciommi, M., (2017). The orness value for rank-dependent welfare functions and rank-dependent poverty measures. *Fuzzy Sets Systems*, 325, pp. 114–136. https://doi.org/10.1016/j.fss.2017.04.003.

Asfaw, A. A., (2021). The effect of income support programs on job search, workplace mobility and COVID-19: International evidence. *Economics & Human Biology*, 41, Article 100997. https://doi.org/10.1016/j.ehb.2021.100997.

Belhadj, B., (2011). New fuzzy indices of poverty by distinguishing three levels of poverty. *Research in Economics*, 65(3), pp. 221–231. https://doi.org/10.1016/j.rie.2010.10.002.

Belhadj, B., Limam, M., (2012). Unidimensional and multidimensional fuzzy poverty measures: New approach. *Economic Modelling*, 29(4), pp. 995–1002.https://doi.org/10.1016/j.econmod.2012.03.009.

Betti, G., Cheli, B., Lemmi, A. and Verma, V., (2008). The fuzzy set approach to multidimensional poverty: The case of Italy in the 1990s. In N. Kakwani, & J. Silber (Eds.), Quantitative approaches to multidimensional poverty measurement (pp. 30–48). *Palgrave Macmillan*. https://doi.org/10.1057/9780230582354_2.

Betti, G., Mangiavacchi, L. and Piccoli, L., (2017). Individual poverty measurement using a fuzzy intrahousehold approach (*IZA Discussion Paper*, No. 11009). https://ftp.iza.org/dp11009.pdf.

Blank, R. M., (1997). It takes a nation: A new agenda for fighting poverty. Russell Sage Foundation, *Princeton University Press*.

Blank, R. M., (2003). Selecting among anti-poverty policies: Can an economics be both critical and caring? *Review of Social Economy*, 61(4), pp. 447–471. https://doi.org/10.1080/0034676032000160949.

Boender, C. G. E., De Graan, J. G. and Lootsma, F. A., (1989). Multi-criteria decision analysis with fuzzy pairwise comparisons. *Fuzzy Sets and Systems*, 29, pp. 133–143. https://doi.org/10.1016/0165-0114(89)90187-5.

Bradshaw, T. K., (2007). Theories of poverty and anti-poverty programs in community development. *Community Development*, 38(1), pp. 7–25. https://doi.org/10.1080/15575330709490182.

Brady, D., (2019). Theories of the causes of poverty. *Annual Review of Sociology*, 45(1), pp. 155–175. https://doi.org/10.1146/annurev-soc-073018-022550.

Brandt, L., (1908). The causes of poverty. *Political Science Quarterly*, 23(4), 637-651. https://doi.org/10.2307/2140866.

Buckley, J. J., Feuring, T. and Hayashi, Y., (2001). Fuzzy hierarchical analysis revisited. *European Journal of Operational Research*, 129(1), pp. 48–64. https://doi.org/10.1016/S0377-2217(99)00405-1.

Cerioli, A., Zani, S., (1990). A fuzzy approach to the measurement of povert y. In C. Dagum, & M. Zenga (Eds.), Income and wealth distribution, inequality and poverty (pp. 272–284). *Springer*. https://doi.org/10.1007/978-3-642-84250-4_18.

Chen, C. T., (2000), Extensions of the TOPSIS for group decision-making under fuzzy environment. *Fuzzy Set Syst*, 114, pp. 1–9. https://doi.org/10.1016/S0165-0114(97)00377-1.

Chen, S.-M., (1996). Evaluating weapon systems using fuzzy arithmetic operations. *Fuzzy Sets and Systems*, 77(3), pp. 265–276. https://doi.org/10.1016/0165-0114(95)00096-8.

Chiappero-Martinetti, E. (1994). A new approach to evaluation of well-being and poverty by Fuzzy Set Theory. *Giornale degli Economisti e Annali di Economia. Nuova Serie*, 53(7–9), pp. 367–388.

Chou C.-H., Liang G.-S. and Chang H.-C., (2013). A fuzzy AHP approach based on the concept of possibility extent. *Quality & Quantity*, 47, pp. 1–14. https://doi.org/10.1007/s11135-011-9473-6.

Ciani, M., Gagliardi, F., Riccarelli, S. and Betti, G., (2019). Fuzzy measures of multidimensional poverty in the Mediterranean Area: A focus on financial dimension. Sustainability, 11(1), *Article 143*. https://doi.org/10.3390/su11010143.

Csutora, R., Buckley, J. J., (2001). Fuzzy hierarchical analysis: The Lambda-Max method. *Fuzzy Sets and Systems*, 120(2), pp. 181–195. https://doi.org/10.1016/S0165-0114(99)00155-4.

Deleeck, H., Van den Bosch, K., (1990). The measurement of poverty in a comparative context: Empirical evidence and methodological evaluation of four poverty lines in seven EC countries. In R. Teekens, & B. M. S. van Praag (Eds.), Analysing poverty in the European Community: Policy issues, research options and data sources (pp. 153–186), (*Eurostat News spec. ed.*, 1–1990). Paper presented at the seminar 'Poverty statistics in the European Community'. https://op.europa.eu/en/publication-detail/-/publication/a74627b5-3b6f-4ec3-992f-5a104b68dccf.

Dudek, H., (2008). Subjective aspects of economic poverty: Ordered response model approach (*Working Paper*, No. 8–08). Department of Applied Econometrics Warsaw School of Economics. https://ssl-kolegia.sgh.waw.pl/pl/KAE/struktura/IE/struktura/ZES/Documents/Working_Papers/aewp08-08.pdf.

Dudek, H., Szczesny, W., (2021). Multidimensional material deprivation in Poland: A focus on changes in 2015–2017. *Quality & Quantity*, 55(2), pp. 741–763. https://doi.org/10.1007/s11135-020-01024-3.

Easterlin, R. A., (1974). Does economic growth improve the human lot? Some empirical evidence. In P. A. David, & M. W. Reder (Eds.), Nations and households in economic growth: Essays in honor of Moses Abramovitz (pp. 89–125). *Academic Press*. https://doi.org/10.1016/B978-0-12-205050-3.50008-7.

Ghiatis, A., (1990). Low income groups obtained by enhanced processing of the Household Budget Surveys in the EC: Summary figures for Italy and The Netherlands. In R. Teekens, & B. M. S. van Praag (Eds.), Analysing poverty in the European Community: Policy issues, research options and data sources (pp. 117–137), (*Eurostat News spec. ed.*, 1–1990). Paper presented at the seminar 'Poverty statistics in the European Community'. https://op.europa.eu/en/publication-detail/-/publication/a74627b5-3b6f-4ec3-992f-5a104b68dccf.

Gupta, J., Bavinck, M., Ros-Tonen, M., Asubonteng, K., Bosch, H., Van Ewijk, E., Hordijk, M.,Van Leynseele, Y., Lopes Cardozo, M., Miedema, E., Pouw, N., Rammelt, C., Scholtens, J., Vegelin, C. and Verrest, H., (2021). COVID-19, poverty and inclusive development. World Development, 145. *Article 105527*. https://doi.org/10.1016/j.worlddev.2021.105527.

Hagenaars, A. J. M., Van Praag, B. M. S., (1985). A synthesis of poverty line definitions. *Review of Income and Wealth*, 31, pp. 139–153. https://doi.org/10.1111/j.1475-4991.1985.tb00504.x.

Haveman, R. H., (2015). Poverty: Measurement and analysis. In J. D. Wright (Ed.), International encyclopedia of the social & behavioral sciences (2nd ed., pp. 741–746). *Elsevier*. https://doi.org/10.1016/B978-0-08-097086-8.71035-X.

Hwang C. L.; Yoon K., (1981): Multiple attribute decision making: methods and applications. *Springer-Verlag*, New York.

Jennings, J., (1999). Persistent poverty in the United States: Review of theories and explanations. In L. Kushnick, & J. Jennings (Eds.), A new introduction to poverty: The role of race, power, and politics (pp. 13-38). *New York University Press*.

Kalinowski, S., (2015). Poziom życia ludności wiejskiej o niepewnych dochodach [The living standards of the rural population with uncertain income]. *PWN*.

Kalinowski, S., (2020a). Od paniki do negacji: Zmiana postaw wobec COVID-19 [From panic to denial: Changing attitudes towards Covid-19]. *Wieś i Rolnictwo*, 3(188), pp. 45–65. https://doi.org/10.7366/wir032020/03.

Kalinowski, S., (2020b). Poverty in rural areas: An outline of the problem. Acta Scientiarum Polonorum. *Oeconomia*, 19(4), pp. 69–78. https://doi.org/10.22630/ASPE.2020.19.4.42.

Kalinowski, S., Wyduba, W., (2020). Moja sytuacja w okresie koronawirusa: Raport końcowy z badań [My situation during the coronavirus period: Final research report]. *Instytut Rozwoju Wsi i Rolnictwa PAN*. https://doi.org/10.53098/9788389900609.

Kapteyn, A., Kooreman, P. and Willemse, R., (1988). Some methodological issues in the implementation of subjective poverty definitions. *Journal of Human Resources*, 23(2), pp. 222–242. https://doi.org/10.2307/145777.

Kordos, M., Skwarczyński, M. and Zawadowski, W., (Eds.) (1993). Leksykon matematyczny [Mathematical lexicon]. *Wydawnictwo Wiedza Powszechna*.

Łuczak A., Kalinowski S. (2022). Measuring Subjective Poverty: Methodological and Application Aspects. In: Jajuga, K., Dehnel, G., Walesiak, M. (eds) Modern Classification and Data Analysis. SKAD 2021. Studies in Classification, Data Analysis, and Knowledge Organization. *Springer*, Cham. https://doi.org/10.1007/978-3-031-10190-8_18.

Łuczak, A., Wysocki, F., (2008). Wykorzystanie rozmytej metody TOPSIS opartej na zbiorach α-poziomów do porządkowania liniowego obiektów [Application of fuzzy TOPSIS method based on α-level sets to linear ordering of objects]. In K. Jajuga, & M. Walesiak (Eds.), Klasyfikacja i analiza danych – teoria i zastosowania (pp. 337–345), (Taksonomia 15; Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu 1207). *Uniwersytet Ekonomiczny we Wrocławiu*.

Michoń, P., (2010). Ekonomia szczęścia: Dlaczego ludzie odmawiają wpasowania się w modele ekonomiczne [Happiness economics: Why people refuse to fit into economic models]. *Dom Wydawniczy Harasimowicz*. https://www.wbc.poznan.pl/dlibra/publication/242598.

Mill, J. S., (1907). On social freedom: Or the necessary limits of individual freedom arising out of the conditions of our social life. *The Oxford and Cambridge Review*, 1, pp. 57–83. https://archive.org/details/the-oxford-and-cambridge-review 1/page/n13/mode/2up.

Montrone, S., Campobasso, F., Perchinunno, P. and Fanizzi, A., (2010). A fuzzy approach to the small area estimation of poverty in Italy. In G. Phillips-Wren, L. C. Jain, K. Nakamatsu, & R. J. Howlett (Eds.), Advances in intelligent decision technologies (pp. 309–318), (Smart Innovation, Systems and Technologies 4). *Springer*. https:// doi.org/10.1007/978-3-642-14616-9_30.

Mowafi, M., (2004). *The meaning and measurement of poverty: A look into the global debate*. http://www.sas.upenn.edu/~dludden/Mowafi_Poverty_Measurement_Debate.pdf.

Neff, D., (2013). Fuzzy set theoretic applications in poverty research. *Policy and Society*, 32(4), 319-331. https://doi.org/10.1016/j.polsoc.2013.10.004.

Nettle, D., (2005). Happiness: The science behind your smile. *Oxford University Press*.

Orshansky, M., (1969). How poverty is measured. *Monthly Labor Review*, 92(2), pp. 37–41. http://www.jstor.org/stable/41837556.

Pouw, N., (2020). Wellbeing economics: How and why economics needs to change? *Amsterdam University Press*.

Ravallion, M., (2012). Poor, or just feeling poor? On using subjective data in measuring poverty (Policy Research Working Paper No. 5968). *The World Bank*. https://ssrn.com/abstract=2004930.

Ravallion, M., Lokshin, M., (2002). Self-rated economic welfare in Russia. *European Economic Review*, 46(8), pp. 1453–1473. https://doi.org/10.1016/S0014-2921(01)00151-9.

Ravallion, M., Himelein, K. and Beegle, K., (2016). Can subjective questions on economic welfare be trusted? *Economic Development and Cultural Change*, 64(4), pp. 697–726. https://doi.org/10.1086/686793.

Rayo, L., Becker, G. S., (2007). Evolutionary efficiency and happiness. *Journal of Political Economy*, 115(2), pp. 302–337. https://doi.org/10.1086/516737.

Rodgers, H. R., Jr., (2000). American poverty in a new era of reform. M. E. Sharp.

Rowntree, B. S., (1901). Poverty: A study of town life. *Macmillan*.

Runciman, W. G., (1966). Relative deprivation and social justice: A study of attitudes to social inequality in twentieth-century England. *Routledge and Kegan Paul*.

Saaty, T. L., (1980). The analytic hierarchy process: planning, priority setting, resource allocation. *McGraw- Hill International Book Company*.

Schiller, B. R., (1989). The economics of poverty and discrimination (5th ed.). *Prentice Hall*.

Shaw, W., (1996). The geography of United States poverty: Patterns of deprivation, 1980-1990. *Garland Publishing*.

Stanny, M., Rosner, A. and Komorowski, Ł., (2018). Monitoring rozwoju obszarów wiejskich. Etap III. Struktury społeczno-gospodarcze, ich przestrzenne zróżnicowanie i dynamika [Monitoring of rural areas developmnet. III Stage. Socio-economic structures, their spatial diversity and dynamics]. Europejski Fundusz Rozwoju Wsi Polskiej, *IRWiR PAN*.

Stiglitz, J. E., Sen, A. and Fitoussi, J. P., (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress. *CMEPSP*. https://www.socioeco.org/bdf_fiche-document-2339_en.html.

Thurow, L. C., (1967). The causes of poverty. *The Quarterly Journal of Economics*, 81(1), pp. 39–57. https://doi.org/10.2307/1879672.

Townsend, P., (1979). Poverty in the United Kingdom: A survey of household resources and standards of living. *Penguin Books*. https://www.poverty.ac.uk/system/files/townsend-book-pdfs/PIUK/piuk-whole.pdf.

Valentine, C. A., (1968). Culture and poverty: Critique and counter-proposals. *University of Chicago Press*.

van Praag, B. M. S., (1971). The welfare function of income in Belgium: An empirical investigation. *European Economic Review*, 2(3), pp. 337–369. https://doi.org/10.1016/0014-2921(71)90045-6.

van Praag, B., Ferrer-i-Carbonell, A., (2008). A multidimensional approach to subjective poverty. In N. Kakwani, & J. Silber (Eds.), Quantitative approaches to multidimensional poverty measurement (pp. 135–154). *Palgrave Macmillan*. https://doi.org/10.1057/9780230582354_8.

van Praag, B., Goedhart, T. and Kapteyn, A., (1980). The poverty line – a pilot survey in Europe. *The Review of Economics and Statistics*, 62(3), pp. 461–465. https://doi.org/10.2307/1927116.

Wang, J.-W., Cheng, C.-H. and Kun-Cheng, H., (2009). Fuzzy hierarchical TOPSIS for supplier selection. *Applied Soft Computing*, 9(1), pp. 377–386. https://doi.org/10.1016/j.asoc.2008.04.014.

Zadeh, L. A., (1975a). The concept of a linguistic variable and its application to approximate reasoning – I. *Information Sciences*, 8(3), pp. 199–249. https://doi.org/10.1016/0020-0255(75)90036-5.

Zadeh, L. A., (1975b). The concept of a linguistic variable and its application to approximate reasoning – II. *Information Sciences*, 8(4), pp. 301–357. https://doi.org/10.1016/0020-0255(75)90046-8.

Zadeh, L. A., (1975c). The concept of a linguistic variable and its application to approximate reasoning – III. *Information Sciences*, 9(1), pp. 43–80. https://doi.org/10.1016/0020-0255(75)90017-1.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 93–115, https://doi.org/10.59139/stattrans-2025-006
Received – 22.02.2024; accepted – 13.12.2024

Type I heavy-tailed family of generalized Burr III distributions: properties, actuarial measures, regression and applications

Wilbert Nkomo[1], Broderick Oluyede[2], Fastel Chipepa[3]

Abstract

This study introduces a new family of distributions (FoD) called type I heavy-tailed odd Burr III-G (TI-HT-OBIII-G) distribution. Several statistical properties of the family are derived along with actuarial risk measures. The maximum likelihood estimation (MLE) approach is adopted in the parameter estimation process. The estimates are evaluated centered on mean square errors and average bias via the Monte Carlo simulation framework. A regression model is formulated and the residual analysis is investigated. Members of the new FoD are applied to heavy-tailed data sets and compared to some well-known competing heavy-tailed distributions. The practicality, flexibility and importance of the new distribution in modeling is empirically proven using three data sets.

Key words: type I heavy-tailed-G, odd Burr III-G, parameter estimation, regression, actuarial measures.
Mathematics Subject Classification: 62E99, 60E05.

1. Introduction

Heavy-tailed distributions have high variances due to outliers with very high values. Modeling data with high variances using standard distributions has deficiencies since they lack flexibility in providing the good fit to heavy-tailed data sets. In a similar vein to rare occurrences such as earthquakes and cyclones, financial risks, such as insurance losses, often exhibit right-skewed data with heavy tails. This characteristic poses challenges for modeling such data using conventional methods.

[1] Department of Mathematics and Statistical Sciences, Botswana International University of Science and Technology, Botswana & Department of Applied Statistics, Manicaland State University of Applied Sciences, Zimbabwe. E-mail: nw22100009@studentmail.biust.ac.bw. ORCID: https://orcid.org/0009-0006-0277-3981.

[2] Department of Mathematics and Statistical Sciences, Botswana International University of Science and Technology, Botswana. E-mail: oluyedeo@biust.ac.bw. ORCID: https://orcid.org/0000-0002-9945-2255.

[3] Department of Mathematics and Statistical Sciences, Botswana International University of Science and Technology, Botswana. E-mail: chipepaf@biust.ac.b. ORCID: https://orcid.org/0000-0001-6854-8740.

Several authors have proposed generalized distributions to curb this inadequacy, for instance, Zhao et al. (2020) in their research proposed the type I heavy-tailed Weibull (TI-HT-W) distribution using the transformed-transformer (T-X) approach. They evaluated its suitability for analyzing the prevention of HIV progression with two antiretroviral drugs and compared it to the Weibull distribution. The findings demonstrated that the TI-HT-W distribution outperformed the Weibull distribution. The study aimed to enhance understanding of treatment strategies for HIV by providing insights into the effectiveness of different approaches contributing to advancements in HIV prevention and management. In their study, Dey et al. (2019) introduced a statistical distribution called the alpha power transformed inverse Lindley (APTIL) distribution. This distribution incorporates the inverse Lindley distribution and utilizes the alpha power transformation (APT), resulting in a versatile model with both scale and shape parameters. They found that the density function exhibited a single peak, indicating unimodality, and the hazard rate function (hrf) displayed a bathtub-shaped pattern, hence the distribution was found effective in analyzing lifetime data. Descheemaeker et al. (2021) studied complex ecological communities using stochastic Lotka-Volterra models with heavy-tailed abundance distributions. Their research focused on explaining how numerous species coexist within these communities and why rare species tend to dominate.

In situations where events exhibit very high deviations from the mean, surpassing what is expected based on the available baseline distributions, the application of heavy-tailed distributions becomes necessary. Heavy-tailed distributions are utilized to handle exceptional or uncommon events that defy explanation by conventional distributions. The literature presents diverse heavy-tailed distributions, offering mathematical models tailored to capture the distinctive characteristics of these events and provide more accurate probability estimates. Some recent advancements in the field of heavy-tailed distributions encompass various contributions including the heavy-tailed exponential by Affify et al. (2020), type II half logistical odd Fréchet by Alyami et al. (2022), alpha power Topp-Leone Weibull by Benkhelifa (2022) and heavy-tailed log-logistic by Teamah et al. (2021), among others. The motivations behind the development of this heavy-tailed distribution include:

(i) extending existing distributions using the TI-HT and the OBIII-G FoD.
(ii) expanding the parental distribution's adaptability in terms of density and hazard rate forms.
(iii) controlling the magnitude or influence of the tails in a parental distribution.
(iv) modeling and representing diverse data sets across multiple domains.

The paper follows the subsequent organization. We develop and present the new TI-HT-OBIII-G distribution including its sub-families and special cases in Section 2. Several statistical properties including moments, Rényi entropy and order statistics are

presented. Section 3 presents parameter estimation. Section 4 discusses risk measures and their numerical simulations. Section 5 discusses simulations and findings. The regression model is formulated in Section 6 while Sections 7 and 8 cover applications and conclusions, respectively.

2. The generalized distribution

The type I heavy-tailed odd Burr III-G (TI-HT-OBIII-G) family is developed in this section.

Zhao et al. (2020) proposed the type I heavy-tailed (TI-HT-G) FoD. The cumulative distribution function (cdf) of the TI-HT FoD is

$$F_{TI-HT-G}(x;\theta,\Omega) = 1 - \left(\frac{1-G(x;\Omega)}{1-(1-\theta)G(x;,\Omega)}\right)^{\theta} \tag{1}$$

and the probability density function (pdf) is

$$f_{TI-HT-G}(x;\theta,\Omega) = \frac{\theta^2 g(x;\Omega)\{1-G(x;\Omega)\}^{\theta-1}}{\{1-(1-\theta)G(x;\Omega)\}^{\theta+1}}, \tag{2}$$

for θ, $x > 0$, where Ω denotes the parameter vector from the baseline distribution $G(.)$.

Alizadeh et al. (2017) presented the odd Burr III-G (OBIII-G) FoD. The OBIII-G cdf is

$$F_{OBIII-G}(x;\, c,\, k,\, \Psi) = [1 + B_G(x;\, c,\, \Psi)]^{-k} \tag{3}$$

and the pdf is

$$f_{OBIII-G}(x;\theta,\Omega) = ckg(x;\Psi)\frac{[1-G(x;\Psi)]^{c-1}}{[G(x;\Omega)]^{c+1}}[1 + B_G\,(x;\; c,\Psi)]^{-k-1}, \tag{4}$$

for c, k, $x > 0$, and parameter vector Ψ, where $B_G\,(x;\; c,\Psi) = \left(\frac{1-G(x;\Psi)}{G(x;\Psi)}\right)^{c}$.

Replacing the baseline cdf in Equation (1) with the OBIII-G FoD yields the new FoD called TI-HT-OBIII-G with cdf

$$F(x;\; c,k,\theta,\Psi) = 1 - \left(\frac{1-[1 + B_G\,(x;\, c,\Psi)]^{-k}}{1-(1-\theta)[1 + B_G\,(x;\, c,\Psi)]^{-k}}\right)^{\theta} \tag{5}$$

and pdf

$$f(x;c,k,\theta,\Psi) = \frac{\theta^2 ckg(x;\Psi)\left[1+\left(\frac{1-G(x;\Psi)}{G(x;\Psi)}\right)^{c}\right]^{-k-1}(1-G(x;\Psi))^{c-1}}{(G(x;\Psi))^{c+1}}\left\{1-\left[1+\left(\frac{1-G(x;\Psi)}{G(x;\Psi)}\right)^{c}\right]^{-k}\right\}^{\theta-1}$$

$$\times\left\{1-(1-\theta)\left[1+\left(\frac{1-G(x;\Psi)}{G(x;\Psi)}\right)^{c}\right]^{-k}\right\}^{-(\theta+1)} \tag{6}$$

for $c, k, \theta > 0$ and baseline parameter vector Ψ. The model contains many sub-families by letting some of the parameters equal to unit.

2.1. Quantile function

The quantile function of the TI-HT-OBIII-G FoD is

$$Q_{X(u)} = G^{-1}\left[\left(1+\left[\left(\frac{1-(1-u)^{\frac{1}{\theta}}}{1-(1-u)^{\frac{1}{\theta}}[1-\theta]}\right)^{\frac{-1}{k}}-1\right]^{\frac{1}{c}}\right)\right]^{-1} \tag{7}$$

for $0 \le u \le 1$. To determine the quantile values of the TI-HT-OBIII-G FoD, the process involves solving Equation (7) and providing the baseline distribution G(.). The quantile function relies on the baseline cdf G(.), and the quantile values can be obtained by employing numerical methods in R software to solve the nonlinear equation. For more detailed derivations, please refer to the web **appendix**.

2.2. Linear representation

This subsection seeks to expand the density of the TI-HT-OBIII-G FoD. The density of the TI-HT-OBIII-G FoD can be represented in the form $f(x;\ c,k,\theta,\Psi) = \sum_{w=0}^{\infty}\eta_{w+1}g_{w+1}\ (x;\ \Psi)$ (8)

where $g_{w+1}(x; \Psi) = (w + 1)\, g\,(x; \Psi)\, G^{w}\,(x; \Psi)$ defines the exponentiated-G (Expon-G) distribution, $(w + 1)$ is the power parameter and

$$\eta_{w+1} = \sum_{r,s,t,v=0}^{\infty}\ (-1)^{r+s+v+w}(1-\theta)^{r}\mathrm{ck}\,\theta^{2}\binom{-(\theta+1)}{r}\binom{\theta-1}{s}\binom{\frac{1}{w+1}}{} \times\binom{-[1+k(s+r+1)]}{t}\binom{-[1+c(t+1)]}{v}\binom{v+c(t+1)-1)}{w}. \tag{9}$$

As a result, the pdf of the TI-HT-OBIII-G FoD can be presented as an unbounded linear mixture of the exponentiated-G (Expon-G) densities. This representation allows for the direct deduction of structural properties associated with the TI-HT-OBIII-G FoD. The web **appendix** contains detailed derivations of the density expansion and explores structural properties, including moments, order statistics, and Rényi entropy.

2.3. Special cases

This subsection presents some special cases of TI-HT-OBIII-G FoD by specifying $G\,(x; \Psi)$ and $g\,(x; \Psi)$ in Equation (5) and Equation (6).

2.3.1. Type I heavy-tailed odd Burr III-Weibull distribution

Considering the Weibull distribution with cdf $G\,(x; \lambda) = 1-\exp(-x^{\lambda})$ and pdf

$g(x;\lambda) = \lambda x^{\lambda-1}\exp(-x^{\lambda})$, for $\lambda, x > 0$, as the baseline distribution, we have the Type I heavy-tailed odd Burr III-Weibull (TI-HT-OBIII-W) distribution defined by the cdf

$$F(x;\theta,\lambda,c,k) = 1 - \left(\frac{1-[1+B_1(x;c,\lambda)]^{-k}}{1-(1-\theta)[1+B_1(x;c,\lambda)]^{-k}}\right)^{\theta},$$

and pdf

$$f(x;\theta,\lambda,c,k) = \frac{\theta^2 ck\lambda x^{\lambda-1}[1+B_1(x;c,\lambda)]^{-k-1}\exp\left(-x^{\lambda}\right)^{c}}{\left(1-\exp(-x^{\lambda})\right)^{c+1}(1-(1-\theta)[1+B_1(x;c,\lambda)]^{-k})^{(\theta+1)}} \times\{1-[1+B_1(x;c,\lambda)]^{-k}\}^{\theta-1}$$

for $\theta,\lambda,c,k > 0$, where $B_1(x;c,\lambda) = \left(\frac{\exp(-x^{\lambda})}{1-\exp(-x^{\lambda})}\right)^{c}$.

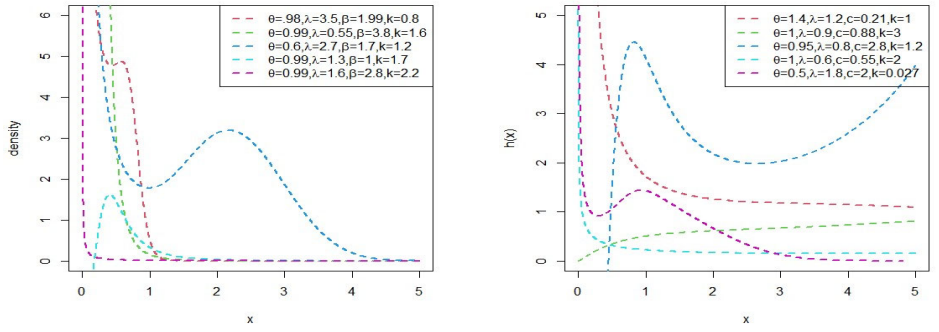


Figure 1: TI-HT-OBIII-W distribution's density and hrf plots

Figure 1 illustrates several density plots and hrf plots for the TI-HT-OBIII-W distribution. The TI-HT-OBIII-W density is capable of handling data that is right-skewed as well as reversed-J shaped. The hrf exhibits a variety of geometric configurations, such as decreasing, increasing, and bathtub-shaped followed by inverted bathtub-shaped patterns, as well as inverted bathtub-shaped followed by bathtub-shaped patterns.

2.3.2. Type I heavy-tailed odd Burr III-Kumaraswamy distribution

The type I heavy-tailed odd Burr III-Kumaraswamy (TI-HT-OBIII-Kum) distribution can be considered by utilizing the Kumaraswamy distribution as the baseline. The cdf and pdf of Kumaraswamy distribution are characterized by $G(x; a, b) = 1 - (1 - x^{a})^{b}$ and $g(x; a, b) = abx^{a-1}(1 - x^{a})^{b-1}$ respectively, for $a, b, x > 0$. The cdf of TI-HT-OBIII-Kum distribution is

$$F(x;b,a,\theta,c,k) = 1 - \left(\frac{1-[1+B_2(x;b,a,c)]^{-k}}{1-(1-\theta)[1+B_2(x;b,a,c)]^{-k}}\right)^{\theta}$$

and the pdf is

$$f(x;b,a,\theta,c,k) = \frac{\theta^2 abck(x)^{a-1}[1+B_2(x;b,a,c)]^{-k-1}[(1-x^a)^b]^{c-1}}{(1-(1-x^a)^b)^{c+1}(1-(1-\theta)[1+B_2(x;b,a,c)]^{-k})^{(\theta+1)}} \times \{1-[1+B_2(x;b,a,c)]^{-k}\}^{\theta-1},$$

for $a,b,c,k,\theta,x>0$, where $B_2(x;b,a,c) = \left(\frac{(1-x^a)^b}{1-(1-x^a)^b}\right)^c$.

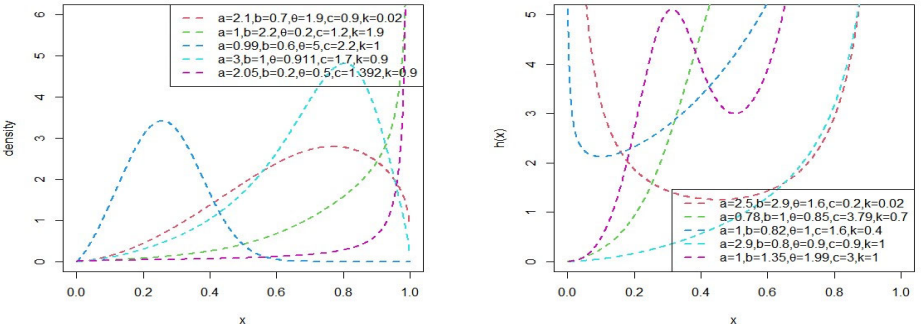


Figure 2: Density and hrf plots for the TI-HT OBIII-Kum distribution.

Figure 2 reveals the density and hrf plots corresponding to the TI-HT-OBIII-Kum distribution. The pdf of the TI-HT-OBIII-Kum distribution is suitable for analyzing data that display positive skewness, negative skewness, and J-shaped pattern. The hrf exhibits both monotonically increasing and nonmonotonically increasing patterns.

2.3.3. Type I-heavy tailed odd Burr III-Pareto distribution

The cdf and pdf of the Pareto (type I) distribution are $G(x;\psi) = 1-\left(\frac{\alpha}{x}\right)^\gamma$ and $g(x;\psi) = \frac{\gamma\alpha^\gamma}{x^{\gamma+1}}$ respectively, where $\gamma > 0$ and $x \geq \alpha$. Setting the Pareto (type I) distribution as the baseline, we have the type I heavy-tailed odd Burr III-Pareto (TI-HT-OBIII-P) distribution with cdf

$$F(x;\theta,\gamma,\alpha,c,k) = 1-\left(\frac{1-[1+B_3(x;a,c,\gamma)]^{-k}}{1-(1-\theta)[1+B_3(x;a,c,\gamma)]^{-k}}\right)^\theta$$

and the pdf is

$$f(x;\theta,\gamma,\alpha,c,k) = \frac{\theta^2 ck\frac{\gamma\alpha^\gamma}{x^{\gamma+1}}[1+B_3(x;a,c,\gamma)]^{-k-1}\left[\left(\frac{\alpha}{x}\right)^\gamma\right]^{c-1}}{\left[1-\left(\frac{\alpha}{x}\right)^\gamma\right]^{c+1}[1-(1-\theta)[1+B_3(x;a,c,\gamma)]^{-k}]^{(\theta+1)}} \times \{1-[1+B_3(x;a,c,\gamma)]^{-k}\}^{\theta-1}$$

respectively for $\theta,\gamma,\alpha,c,k>0$ and $B_3(x;a,c,\gamma) = \left(\frac{\alpha^\gamma}{x^\gamma-\alpha^\gamma}\right)^c$.

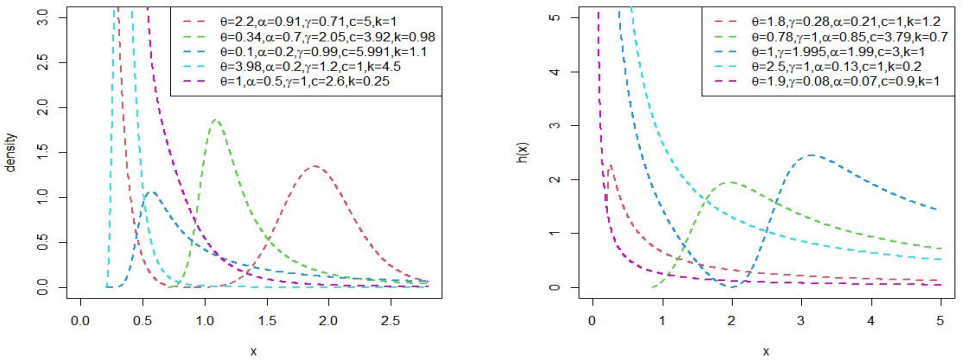


Figure 3: Pdf and hrf plots for the TI-HT-OBIII-P distribution

Density and hrf plots for the TI-HT-OBIII-P distribution are depicted in Figure 3. The TI-HT-OBIII-P density is capable of handling data that is positively-skewed, reversed-J and almost symmetric. The distribution's hrf displays inverted bathtub, decreasing and bathtub followed by inverted bathtub shapes.

3. Maximum Likelihood Estimation

The process of estimating unknown parameters for the TI-HT-OBIII-G FoD was carried out using the maximum likelihood estimation (MLE) technique. Consider a parameter vector $\Delta = (c, k, \theta, \Psi)^T$ and $X_i \sim$ TI-HT-OBIII-G. Using $\ell = \ell(\Delta)$ to denote the log-likelihood function, we have

$$\begin{aligned}\ell(\Delta) = {} & 2n\ln(\theta) + n\ln(c) + n\ln(k) + \sum_{i=1}^{n} \ln[g(x_i;\Psi)] \\ & -(k+1)\sum_{i=1}^{n} \ln\left[1+\left(\frac{1-G(x_i;\Psi)}{G(x_i;\Psi)}\right)^c\right] \\ & +(c-1)\sum_{i=1}^{n} \ln\big(1-G(x_i;\Psi)\big) - (c+1)\sum_{i=1}^{n} \ln\big(G(x_i;\Psi)\big) \\ & +(\theta-1)\sum_{i=1}^{n} \ln\left\{1-\left[1+\left(\frac{1-G(x_i;\Psi)}{G(x_i;\Psi)}\right)^c\right]^{-k}\right\} \\ & -(\theta+1)\ln\sum_{i=1}^{n}\left\{1-(1-\theta)\left[1+\left(\frac{1-G(x_i;\Psi)}{G(x_i;\Psi)}\right)^c\right]^{-k}\right\}.\end{aligned}$$

The MLEs of $(c, k, \theta$ and $\Psi_k)$ are obtained by solving a system of non-linear equations $\left(\frac{\partial \ell}{\partial c}, \frac{\partial \ell}{\partial k}, \frac{\partial \ell}{\partial \theta}, \frac{\partial \ell}{\partial \Psi_k}\right)^T = \mathbf{0}$ using iterative methods in R. See web **appendix** for individual components of the score vector.

4. Risk measures

Risk measures are statistical tools and formulae used by actuaries to evaluate market risk in prospective investments. These metrics encompass the value at risk (VaR), tail variance (TV), tail value at risk (TVaR), and tail variance premium (TVP).

4.1. VaR

VaR quantifies the magnitude of prospective financial losses over a specified time frame. VaR_q for the TI-HT-OBIII-G FoD is calculated from

$$X_q = G^{-1}\left(\left\{1+\left[\left(\frac{1-(1-u)^{\frac{1}{\theta}}}{1-(1-u)^{\frac{1}{\theta}}[1-\theta]}\right)^{\frac{-1}{k}}-1\right]^{\frac{1}{c}}\right\}\right)^{-1}, \tag{10}$$

where $q \in (0,1)$ specifies the significance level.

4.2. TVaR

TVaR computes expected loss value, considering the occurrence of an event exceeding a predefined probability threshold. TVaR for the TI-HTOBIII-G FoD is

$$\begin{aligned} TVaR_q &= E\left(X \mid X > x_q\right) = \frac{1}{1-q}\int_{VaR_q} xf(x)dx \\ &= \frac{1}{1-q}\sum_{w=0}^{\infty} \eta_{w+1}\int_{VaR_q}^{\infty} xg_{w+1}(x;\Psi)dx \end{aligned} \tag{11}$$

where η_{w+1} is provided by Equation (9), $g_{w+1}(x;\Psi)$ represents the Expon-G pdf and $(w+1)$ is the power parameter.

4.3. TV

TV captures the extent of variability in losses given that they exceed a predefined VaR threshold with a specific probability, denoted as p. The TV of the TI-HT-OBIII-G FoD is

$$\begin{aligned} TV_q &= E\left(X^2 \mid X > x_q\right) - \left(TVaR_q\right)^2 \\ &= (1-q)^{-1}\int_{VaR_q}^{\infty} x^2 f(x)dx - \left(TVaR_q\right)^2 \\ &= (1-q)^{-1}\sum_{w=0}^{\infty} \eta_{w+1}\int_{VaR_q}^{\infty} x^2 g_{w+1}(x;\Psi)dx - \left(TVaR_q\right)^2, \end{aligned} \tag{12}$$

where η_{w+1} is defined by Equation (9) and $g_{w+1}(x;\Psi)$ defines the Expon-G distribution. Hence, the TV of TI-HT-OBIII-G FoD can be derived from those of Expon-G distributions.

4.4. TVP

Risk professionals are fretful about risks exceeding certain thresholds. Such situations are common in insurance, for example, in policies involving deductibles and reinsurance contracts. Tail value premium answers demands to these circumstances. The TVP of the TI-HT-OBIII-G FoD is expressed as

$$TVP_q = TVaR_q + \delta(TV_q), \tag{13}$$

for $0 < \delta < 1$. The TI-HT-OBIII-G FoD TVP is found by incorporating Equations (11) and (12) into Equation (13).

4.5. Numerical analysis of risk measures

We provide findings from numerical simulations for the risk measures associated with the TI-HT-OBIII-W distribution. These risk measures were then compared among various distributions, including the type-I heavy-tailed Weibull (TI-HT-W), the two-parameter Weibull, and the one-parameter Weibull distributions. The simulation results were derived by implementing the following procedure:

(1) for each of the distributions under consideration, 100 random samples were generated, and the MLE method was used to estimate the parameters.

(2) 1000 replications were made in computing the risk measures these distributions.

Table 1: Numerical simulation results for risk measures

Distribution	Risk measure	0.70	0.75	0.80	0.85	0.90	0.95	0.99
TI-HT-OBIII-W ($\theta = 0.77, \lambda = 1.3, c = 1.4, k = 0.5$)	VaR	20.4389	22.5339	24.9911	28.0229	32.0982	38.6792	52.7591
	TVaR	29.4737	31.3723	33.6538	36.5297	40.4637	46.8901	48.7331
	TV	128.0030	130.5836	132.0992	134.9406	136.9434	293.2358	317.0861
	TVP	117.6757	124.0600	129.7331	134.2292	136.7127	135.4641	371.6070
TI-HT-W ($\theta = 0.77, \lambda = 1.3, \gamma = 0.04$)	VaR	18.4988	20.4513	22.7442	25.5754	29.3817	35.5251	48.6520
	TVaR	27.3904	29.1343	31.2249	33.8559	37.4539	39.0051	41.024
	TV	94.5879	99.0216	105.7756	112.1283	115.8558	124.8394	207.4612
	TVP	94.3019	99.2408	103.8454	107.9151	111.0971	185.2480	316.6550
Weibull ($\lambda = 1.3, \gamma = 0.04$)	VaR	13.6867	15.2406	17.0794	19.3654	22.4561	27.4667	38.1923
	TVaR	4.2572	5.1467	5.5777	5.9509	6.4436	6.4884	6.6304
	TV	79.8871	86.1114	102.2430	107.9056	110.8282	160.7040	166.8418
	TVP	66.0646	74.7844	81.3720	89.8165	93.7026	101.9418	112.4681
Weibull ($\lambda = 1.3$)	VaR	0.0701	0.0790	0.0896	0.1029	0.1212	0.1514	02182
	TVaR	0.0945	0.1021	0.1345	0.2310	0.2410	0.2610	0.3102
	TV	0.0213	0.0246	0.0294	0.0362	0.0483	0.0777	0.2197
	TVP	0.0149	0.0184	0.0234	0.0307	0.0435	0.0738	0.2175

Table [1] shows the findings of the risk metrics for the three heavy-tailed distributions. The model exhibiting elevated values of the risk measures implies that the model has a more pronounced tail. We can infer from the comparison that the TI-HT-OBIII-W distribution exhibits a heavier tail compared to both the TI-HT-W distribution and the Weibull distributions. As a result, the TI-HT-OBIII-W distribution is considered appropriate for modeling data sets with heavy-tail characteristics.

5 Simulation Study

We seek to weigh the efficiency of MLEs by carrying out a simulation study. Table 2 gives simulation results. We simulated for n= 35, 70, 140, 280, 560, 1120 and 2240 for N=3000 from the TI-HT-OBIII-W distribution. Average bias (AvBIAS) and root mean square error (RMSErr) for an estimated parameter, say (β), are computed as follows:

$AvBIAS(\hat{\beta}) = \frac{\sum_{i=1}^{N} \hat{\beta}_i}{N} - \beta,$ and $RMSErr(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{N} (\hat{\beta}_i - \beta)^2}{N}}$, respectively.

Regarding the displayed data in Tables 2, it is evident that the average estimated parameter values converge towards the true parameter values. Additionally, both the RMSErr and AvBIAS decrease towards zero across all parameters as we increase the sample size. This shows that the TIHT-OBIII-W distribution produces consistent and efficient parameter estimates.

Table 2: Monte Carlo Simulation Results

Parameter	n	(0.8, 1.1, 1.1, 0.6)			(0.9, 1.0, 0.8, 1.2)		
		Mean	RMSErr	AvBias	Mean	RMSErr	AvBias
θ	35	1.4503	1.6864	0.6503	1.4685	1.1435	0.5685
	70	1.2697	1.0268	0.4697	1.3473	1.0018	0.4473
	140	1.0496	0.6277	0.2496	1.2106	0.9649	0.3106
	280	0.9591	0.4094	0.1591	1.1390	0.7247	0.2390
	560	0.8828	0.2049	0.0828	1.0796	0.4049	0.1796
	1120	0.8551	0.1012	0.0551	0.9789	0.1716	0.0789
	2240	0.8310	0.0714	0.0310	0.9176	0.0380	0.0176
λ	35	2.0024	2.2259	1.4024	1.1118	0.2362	0.2118
	70	1.9122	1.6588	0.9122	1.0757	0.1140	0.1757
	140	1.7680	1.5527	0.6680	1.0566	0.0965	0.1466
	280	1.6235	1.1358	0.5235	1.0419	0.0604	0.1219
	560	1.5244	0.9048	0.4244	1.0285	0.0372	0.1085
	1120	1.3954	0.6954	0.2954	1.0227	0.0166	0.0927
	2240	1.2982	0.4760	0.1982	1.0205	0.0078	0.0605

Table 2: Monte Carlo Simulation Results

Parameter	n	(0.8, 1.1, 1.1, 0.6)			(0.9, 1.0, 0.8, 1.2)		
		Mean	RMSErr	AvBias	Mean	RMSErr	AvBias
c	35	1.3164	1.5108	0.2164	1.0165	0.8932	0.2069
	70	1.2170	0.4721	0.1170	0.9508	0.5928	0.2052
	140	1.2118	0.3613	0.1118	0.9332	0.4495	0.2027
	280	1.2012	0.2484	0.1012	0.9235	0.3788	0.1765
	560	1.1893	0.2168	0.0893	0.8573	0.3268	0.1668
	1120	1.1691	0.1764	0.0691	0.8348	0.2596	0.1492
	2240	1.1434	0.1180	0.0634	0.8131	0.2013	0.0835
k	35	1.5605	2.6207	0.9605	2.1491	3.2232	0.9491
	70	1.2535	1.2710	0.6535	1.9198	1.6246	0.7198
	140	0.9593	0.8515	0.3593	1.7307	1.2020	0.5307
	280	0.8355	0.5619	0.2355	1.6073	1.0181	0.4073
	560	0.7250	0.2926	0.1250	1.5180	0.6697	0.3180
	1120	0.6515	0.1475	0.0815	1.3500	0.4386	0.1500
	2240	0.6437	0.1162	0.0737	1.2286	0.2114	0.0486

6. The TI-HT-OBIII-W regression model

The process of conducting a regression analysis on lifetime data entails determining the distribution of a variable X based on a set of covariates $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_z)^T$. Within this context, we present a regression model for the TIHT-OBIII-W distribution which is designed to handle both censored and uncensored data. By letting c=1, we establish a relationship between the parameter λ and the covariates using a log-linear link function $\lambda_j = \exp(\mathbf{u}^T{}_j \beta)$ for $j = 1, \ldots, n$ and $\beta = (\beta_1, \beta_2, \beta_z)^T$ represents a vector comprising the regression coefficients. The survival function of $X \mid u$ is

$$S(X \mid u) = \left(\frac{1-\left[1+\left(\frac{\exp\left(-x^{\exp\left(u^T\beta\right)}\right)}{1-\exp\left(-x^{\exp\left(u^T\beta\right)}\right)}\right)\right]^{-k}}{1-(1-\theta)\left[1+\left(\frac{\exp\left(-x^{\exp\left(u^T\beta\right)}\right)}{1-\exp\left(-x^{\exp\left(u^T\beta\right)}\right)}\right)\right]^{-k}} \right)^{\theta}. \tag{14}$$

The parametric regression model, known as the TI-HT-OBIII-W, is denoted by Equation (14). If we consider A and S as two separate groups of individuals, where x_j represents the lifetime of individuals in set A and S represents the censoring information, the overall log-likelihood function for the parameter vector $\Delta = (\theta, \beta^T, k)^T$ derived from Equation (14) can be expressed in the form $\ell(\Delta) = \sum_{j \in A} \ell_j(\Delta) + \sum_{j \in S} \ell_j^s(\Delta)$,

where $\ell_j(\Delta) = \log(f(x_j \mid u_j))$, $\ell^s_j(\Delta) = \log(S(x_j \mid u_j))$ and $f(x_j \mid u_j)$, $S(x_j \mid u_j)$ represent the density and survival functions of X, respectively. Let ρ take the value 0 if censoring occurs and 1 if failure is observed. Then the expression for the log-likelihood function for Δ can be expressed as

$$\begin{aligned}
\ell(\Delta) = {} & 2n\rho \ln(\theta) + n\ln(k) + \sum_{j\in A}^{n} \ln\left[\exp\left(u_j^T\beta\right) x_j^{\exp\left(u_j^T\beta\right)-1} \exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)\right] \\
& - (k+1)\sum_{j\in A}^{n} \ln\left[1 + \left(\frac{\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}{1-\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}\right)\right] + 2\sum_{j\in A}^{n} \ln\left(1 - \exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)\right) \\
& + (\theta-1)\sum_{j\in A}^{n} \ln\left\{1 - \left[1 + \left(\frac{\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}{1-\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}\right)\right]^{-k}\right\} \\
& - (\theta+1)\ln\sum_{j\in A}^{n} \left\{1 - (1-\theta)\left[1 + \left(\frac{\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}{1-\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}\right)\right]^{-k}\right\} \\
& + \theta\sum_{j\in S} \log(1-\rho)\left[\left(\frac{1 - \left[1 + \left(\frac{\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}{1-\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}\right)\right]^{-k}}{1-(1-\theta)\left[1 + \left(\frac{\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}{1-\exp\left(-x_j^{\exp\left(u_j^T\beta\right)}\right)}\right)\right]^{-k}}\right).
\end{aligned} \tag{15}$$

The MLE $\hat{\Delta}$ of the vector of unknown parameters can be obtained by maximizing Equation (15) using R software. We also consider residuals for the TI-HT-OBIII-W regression model. By plotting the deviance residuals against the index (numerical identifier assigned to each observation in the dataset), one can effectively identify and validate the appropriateness of the fitted model for a typical observation. The deviance residual, which serves as a measure of the disparity between the observed values and the predicted values, can be mathematically defined as

$$r_{Di} = sign(r_{Mi})\,(-2[r_{Mi} + \rho_i \log(\rho_i) - r_{Mi}])^{0.5},$$

where r_{Mi} is the martingale residual defined by

$$r_{M_i} = \begin{cases} 1 + log\left[\left(\dfrac{1-\left[1+\left(\dfrac{exp\left(-x_j^{\,exp\left(u_j{}^T\beta\right)}\right)}{1-exp\left(-x_j^{\,exp\left(u_j{}^T\beta\right)}\right)}\right)\right]^{-k}}{1-(1-\theta)\left[1+\left(\dfrac{exp\left(-x_j^{\,exp\left(u_j{}^T\beta\right)}\right)}{1-exp\left(-x_j^{\,exp\left(u_j{}^T\beta\right)}\right)}\right)\right]^{-k}}\right)^{\theta}\right] & if\ \rho_{i=1}, \\ log\left[\left(\dfrac{1-\left[1+\left(\dfrac{exp\left(-x_j^{\,exp\left(u_j{}^T\beta\right)}\right)}{1-exp\left(-x_j^{\,exp\left(u_j{}^T\beta\right)}\right)}\right)\right]^{-k}}{1-(1-\theta)\left[1+\left(\dfrac{exp\left(-x_j^{\,exp\left(u_j{}^T\beta\right)}\right)}{1-exp\left(-x_j^{\,exp\left(u_j{}^T\beta\right)}\right)}\right)\right]^{-k}}\right)^{\theta}\right] & if\ \rho_{i=0}. \end{cases}$$

and sign (.) assigns the values of +1 when the argument is positive, and −1 when the argument is negative.

In the work by Atkinson (1985), a technique was proposed to create envelopes that facilitate the enhanced analysis of the normal probability plot of residuals. These envelopes, sometimes referred to as simulated confidence bands, are constructed to encompass the residuals. The anticipation is that when the model is suitably fitted, most data points will align within these specified ranges and demonstrate a random distribution.

7. Applications

Real data examples are fitted to the TI-HT-OBIII-W distribution and compared to several non-nested models including some known heavy-tailed distributions and equiparameter models. The TI-HT-OBIII-W distribution is compared to the transmuted exponentiated generalized Weibull distribution (TExGW) proposed by Yousof et al. (2015), the type I heavy-tailed Weibull distribution (TI-HT-W) introduced by Zhao et al. (2020), the heavy-tailed beta power transformed Weibull distribution (HTBPTW) introduced by Zhao et al. (2021), the Weibull Lomax distribution (WL) by Tahir et al. (2014), Kumaraswamy Weibull distribution (KW) introduced by Cordeiro et al. (2010) and the exponential Lindley odd log-logistic Weibull (ELOLLW) proposed by Korkmaz et al. (2018). Visit the web **appendix** for the pdfs of distributions used in the comparisons.

We presented the goodness-of-fit (Gof) statistics: -2 log-likelihood (-2log(L)), Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (CAIC), Bayesian Information Criterion (BIC), Cramér-von Mises (W^*) and Andersen-Darling (A^*). These statistics are used to verify the best-fitting model for a given data set. Reduced values of these metrics indicate that the model is a better fit compared to other competing models.

Gof was also assessed by the Kolmogorov-Smirnov (K-S) statistic, its associated p-value, and the sum of squares (SS) derived from probability plots. The model that exhibits the smallest K-S value and the highest p-value for the K-S statistic is considered as the best-fitting model. Furthermore, graphical presentations of the fitted densities and probability-probability (PP) plots, empirical cumulative distribution function (ECDF), Kaplan-Meier (K-M) survival, total time on test (TTT) plots and hrf plots were presented.

7.1. Stress-rupture life data

The initial dataset showcases the stress-rupture life of Kevlar 49/epoxy strands under continuous sustained pressure at a 90% stress level until failure. This practical example was originally presented by Cooray and Ananda (2008), and subsequently reported by Cordeiro et al. (2014). The data can be found in the **appendix**.

The asymptotic confidence intervals at a 95% confidence level for the model parameters are as follows:
$\theta \in [0.7055 \pm 0.4243]$, $\lambda \in [0.9911 \pm 0.3556]$, $c \in 1.2942 \pm 0.9120]$ and $k \in [0.5288 \pm 0.60409]$.

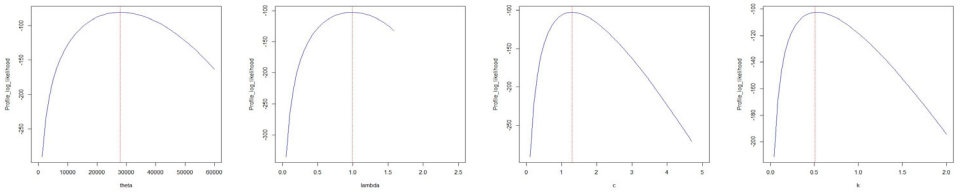
The estimated variance-covariance matrix is

$$\begin{bmatrix} 0.0469 & 0.0030 & -0.0830 & 0.0647 \\ 0.0030 & 0.0329 & -0.0473 & 0.0009 \\ -0.0830 & -0.0473 & 0.2165 & -0.1141 \\ 0.0647 & 0.0009 & -0.1141 & 0.0950 \end{bmatrix}.$$

Table 3: Parameter estimates on stress-rupture life data

Model	Estimates			
TI-HT-OBIII-W	θ	λ	c	k
	0.7055	0.9911	1.2942	0.5288
	(0.2165)	(0.1814)	(0.4653)	(0.3082)
TExGW	λ	a	b	β
	0.0011	0.8113	0.7930	1.0604
	(2.7424)	(2.0878)	(0.6052)	(1.0756)
TI-HT-W	α	θ	γ	
	0.8435	0.6277	1.9662	- -
	(0.0959)	(0.2431)	(0.9963)	
HBPTW	α	γ	k	
	0.8840	1.1239	1.7284	--
	(0.0995)	(0.2008)	(0.6056)	
WL	α	b	α	β
	0.2506	0.7860	1.3581	0.3303
	(0.4173)	(0.1804)	(0.4580)	(0.6282)
KW	α	b	α	β
	0.7280	0.3323	2.4974	1.0514
	(0.2699)	(0.5544)	(4.3669)	(0.1639)
ELOLLW	β	λ	θ	γ
	6.7617	0.1641	7.0720	0.8374
	(0.3733)	(0.0853)	(0.2268)	(0.0144)

The MLEs together with their standard errors (SEs) (in parenthesis) for the models on stress-rupture life data are given in Table 3 and Gof statistics are presented in Table 4. The profile plots in Figure [4] clearly show that the TI-HT-OBIII-W parameters on stress-rupture life data are global maximums and are identifiable.



(a) (b) (c) (d)

Figure 4: Profile plots for the TI-HT-OBIII-W parameters: Stress-rupture life data

Table 4: Gof statistics on stress-rupture life data

Model	2log(L)	AIC	CAIC	BIC	W^*	A^*	K-S	p-value
TI-HT-OBIII-W	204.7661	212.7661	213.1827	223.2266	0.1255	0.7790	0.0679	0.7409
TExW	205.5743	213.5743	213.9910	224.0348	0.1652	0.9586	0.0844	0.4681
TI-HT-W	207.9245	213.9245	214.7905	219.9504	0.1703	0.9742	0.0786	0.5612
HBPTW	205.5669	211.5669	211.8144	219.4123	0.1864	1.0524	0.0854	0.4526
WL	205.1976	213.1976	213.6143	223.6581	0.1440	0.5627	0.0787	0.5587
KW	205.3001	213.3001	213.7183	223.7621	0.1414	0.1423	0.0773	0.5815
ELOLLW	205.1945	213.1945	213.6112	223.6550	0.1666	0.9601	0.0816	0.5126

Figure 4 shows profile plots for stress-rupture data. The plots illustrate that the TI-HT-OBIII-W parameters are global maximums and are identifiable. Figure 5 supports the dominance of TI-HT-OBIII-W model over the non-nested models on stress-rupture life data. Gof statistics and the p-values obtained on stress-rupture life data show that the TI-HT-OBIII-W model outperforms the various non-nested models that were evaluated.

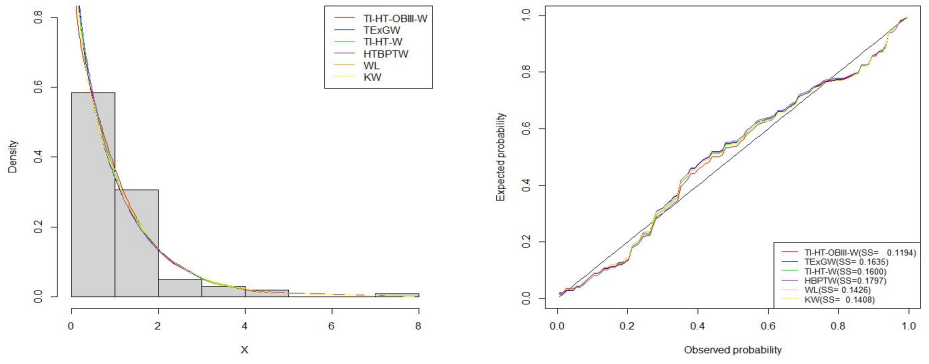


Figure 5: Graphical representations of fitted density functions and probability plots for stress-rupture life data

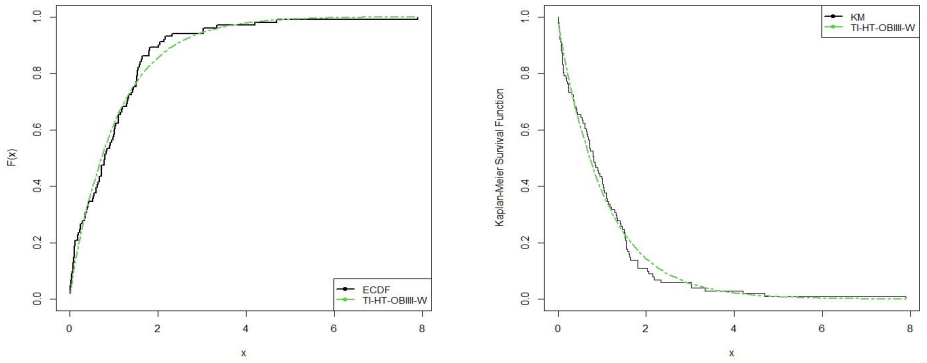


Figure 6: Fitted ECDF curve and K-M plots for stress-rupture life data

Figure 6 presents fitted and observed ECDF and K-M survival curves for stress-rupture life data. The plots show that the TI-HT-OBIII-W distribution closely follows the ECDF and K-M survival curves. The TTT scaled and hrf plots in Figure 7 generally show a sequence of a bathtub followed by an inverted bathtub hazard rate shapes for the stress-rupture life data.

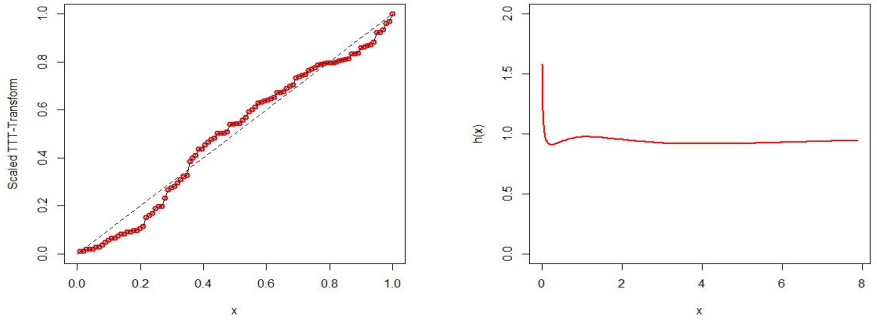


Figure 7: TTT scaled and hrf plots for stress-rupture life data

7.2. Turbocharger data

The second data set was taken from Xu et al. (2003), describing a reliability study on turbochargers in diesel engines. The data is presented in the **appendix**. The estimated MLEs accompanied by their corresponding SEs (in brackets), and GoF statistics for the turbochargers data are displayed in Table 5.

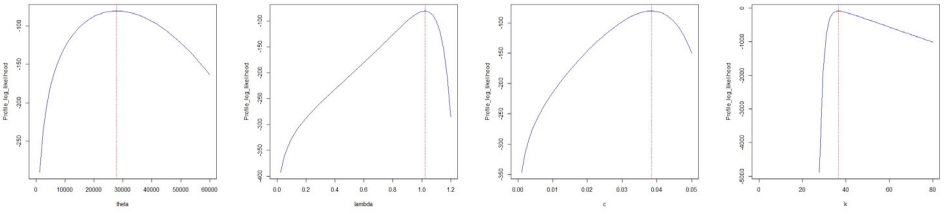
The asymptotic confidence intervals at a 95% confidence level for the model parameters are as follows: $\theta \in [\ 2.64\times10^{4}\pm\ 4.70\times10^{-4}]$, $\lambda \in [1.0278\pm0.6633]$,

$c \in [3.81\times10^{-2}\pm 6.90\times10^{-2}]$ and $k \in [36.3600 \pm 4.2571]$. The estimated variance-covariance matrix is

$$\begin{bmatrix} 5.76\times10^{-8} & 7.12\times10^{-5} & -7.83\times10^{-5} & 5.21\times10^{-4} \\ 7.12\times10^{-5} & 1.15\times10^{-1} & -1.18\times10^{-2} & -6.44\times10^{-1} \\ -7.83\times10^{-6} & -1.18\times10^{-2} & 1.24\times10^{-3} & 7.09\times10^{-2} \\ 5.21\times10^{-4} & -6.44\times10^{-1} & 7.09\times10^{-2} & 4.7200 \end{bmatrix}.$$

Table 5: Estimates on turbochargers data

Model	Estimates			
TI-HT-OBIII-W	θ	λ	c	k
	2.64×10^{4}	1.0278	5.76×10^{-8}	36.3600
	(2.40×10^{-4})	(0.3384)	(3.52×10^{-2})	(2.1720)
TExGW	λ	a	b	β
	0.0321	6.90×10^{-5}	0.6720	4.7712
	(5.21×10^{-10})	(1.31×10^{-5})	(9.07×10^{-10})	(1.85×10^{-9})
TI-HT-W	α	θ	γ	
	3.5513	0.6514	0.0019	- -
	(0.6241)	(0.3243)	(0.0033)	
HBPTW	α	γ	k	
	3.2828	0.0026	0.1490	- -
	(0.7116)	(0.0042)	(0.1766)	
WL	α	b	α	β
	0.7306	2.8721	2.12×10^{4}	1.98×10^{5}
	(0.1329)	(0.3790)	(1.69×10^{-5})	(1.82×10^{-6})
KW	α	b	α	β
	0.5021	89.1240	$4.51\times10^{4-2}$	7.6849
	(6.68×10^{-2})	$(2.65\times10^{4-6})$	$(6.72\times10^{4-3})$	$(4.36\times10^{4-3})$
ELOLLW	β	λ	θ	γ
	1.944	0.1393	1.7917	3.4997
	(1.0760)	(0.0351)	(1.1969)	(0.6303)



(a) (b) (c) (d)

Figure 8: Profile plots for the TI-HT-OBIII-W parameters: Turbochargers data

The Gof statistics and the K-S p-values obtained on reliability of turbochargers in diesel engines data also show that the TI-HT-OBIII-W model is superior to the several non-nested models that were considered. Figure 8 shows profile plots for turbochargers data. The plots illustrate that the TI-HT-OBIII-W parameters are global

maximums and are identifiable. Figure 9 shows that the TI-HT-OBIII-W model outperforms the non-nested models that were considered.

Table 6: Gof statistics on turbochargers life data

Model	2log(L)	AIC	CAIC	BIC	W^*	A^*	K-S	p-value
TI-HT-OBIII-W	160.4252	168.4252	169.5681	175.1808	0.0350	0.2581	0.0752	0.9775
TExW	160.7456	168.7456	171.8885	177.5011	0.0579	0.4441	0.1069	0.7511
TI-HT-W	164.2436	170.2436	170.9103	175.3102	0.0674	0.5073	0.1005	0.8143
HBPTW	163.5073	169.5073	170.1740	174.5739	0.0592	0.4493	0.0977	0.8396
WL	162.3950	170.3950	171.5379	177.1505	0.0522	0.4018	0.1001	0.8173
KW	164.7952	172.7952	173.9381	179.5507	0.0755	0.5633	0.1076	0.7430
ELOLLW	163.7752	171.7752	172.9181	178.5308	0.0636	0.4815	0.1017	0.8028

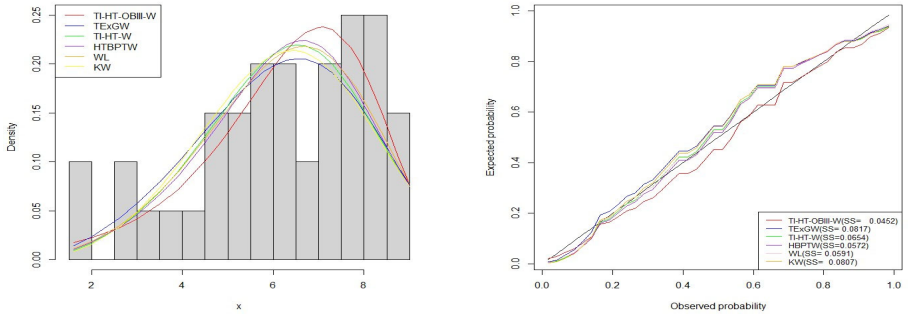


Figure 9: Visualizations of fitted density functions and probability plots for turbochargers data.

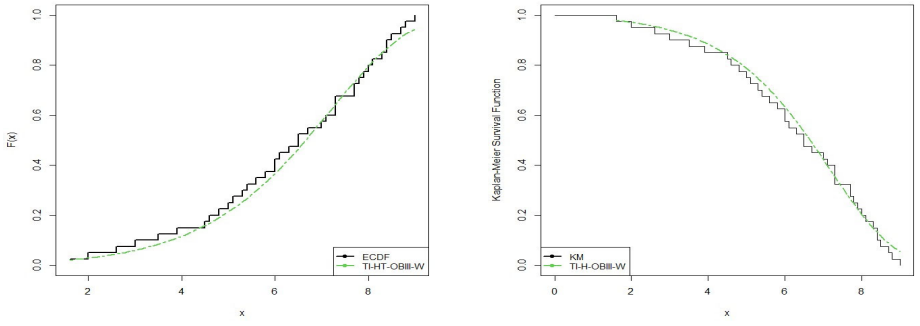


Figure 10: Fitted ECDF curve and K-M plots for turbochargers data

Figure 10 exhibits the convergence between the empirical and the fitted ECDF and K-M survival curves, for turbocharger data. We can see that the TI-HT-OBIII-W demonstrates a remarkable alignment with both the ECDF and K-M survival curves, indicating a close correspondence between the observed and fitted data. The TTT scaled

plot and hrf plots depicted in Figure 11 provide clear evidence that the data shows an increasing hazard rate.

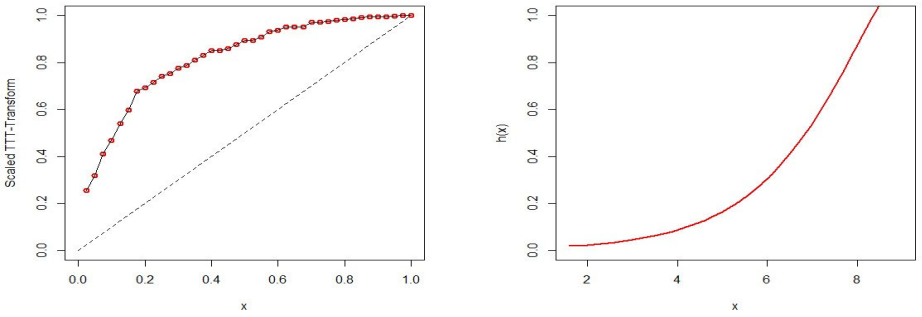


Figure 11: TTT scaled and hrf plots for turbochargers data

7.3. Regression model for transformer turns data

This real data set is sourced from Nelson (2004) and represents a life testing of transformers conducted at high voltage, resulting in multiple censored data. The failures observed in the study were turn-to-turn failures of the primary insulation with 13% censored observations. The data includes observations at three different voltage levels (35.4kV, 42.4kV, and 46.7kV), and the symbol $(*)$ indicates instances of censored data. The data set can be found in the **appendix** section. Firstly, we examine the TI-HT-OBIII-W, TI-HT-W and ELOLLW distributions for the transformer turns data presented in Table 8. MLEs and their SEs (in parentheses) are provided for these distributions. Additionally, we report the -2log(L), AIC, and BIC Gof statistics associated with the fitted models.

The variables considered in this study are x_j=time of failure in hours of the transformer, j = 1, 2, ...,30, and three voltage levels defined by (35.4kV, 42.4kV and 35.4kV).

Table 7: Estimates and Gof statistics transformer turns data

Distribution	Parameter estimates			
TI-HT-OBIII-W	θ	λ	c	k
	2.3871	0.2362	0.9342	11.6409
	0.1361	0.0307	0.3875	0.2855
ELOLLW	β	λ	θ	γ
	0.0011	0.0241	0.7767	0.7185
	(0.2866)	(0.0144)	(6.24×10^{-4})	(0.1034)
TI-HT-W	α	θ	γ	k
	0.7105	1.17×10^{-3}	47.4190	11.6409
	(0.1024)	(6.12×10^{-4})	(1.93×10^{-6})	0.2855

Table 8: Gof statistics on transformer turns data.

Distribution	Gof Statistics				
	−2log(L)	AIC	CAIC	BIC	*SS*
TI-HT-OBIII-W	267.4544	212.7661	213.1827	223.2266	0.0455
ELOLLW	270.6383	213.5743	213.991	224.0348	0.1052
TI-HT-W	270.2332	211.5669	211.8144	219.4123	0.0623

We consider the following structured regression

$$\lambda_j = \exp(\beta_{10} + \beta_{11}u_{j1} + \beta_{12}u_{j2}),$$

where u_{j1} and u_{j2} are the covariates representing the predictor variables for j = 1,2,3, ...,30, to maximize the log-likelihood function in Equation (17) to obtain the MLEs of the parameters of the proposed model. We provide the parameter estimates, SEs, and the significance of the MLEs in Table 9. The findings presented in Table 9 offer convincing empirical support, at a significance level of 5%, indicating a substantial disparity between the 35kV level and the 46.7 kV level.

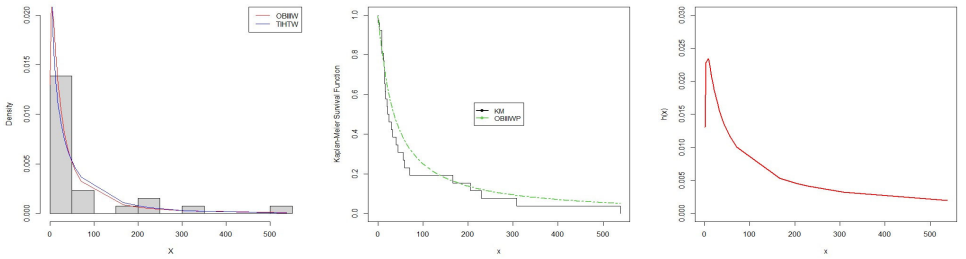


Figure 12: Fitted models: Histogram, K-M curve and hrf plot for transformer turns data

Figure 12 displays the histograms and fitted densities of the TI-HT-OBIII-W and TI-HT-W distributions. The K-M survival curve demonstrates that the TIHT-OBIII-W closely corresponds to the transformer turns data. The hrf curve indicates an inverted hazard rate shape.

Table 9: MLEs for regression model fitted to transformer turns

Parameters	Estimate	SE	p-value
θ	0.5000	0.1515	--
k	3.7576	7.33×10^{-4}	--
β_{10}	7.43×10^{-4}	0.0104	0.3109
β_{11}	0.0659	0.0162	<0.00001
β_{12}	-0.0269	0.2687	0.0978

7.4. Goodness-of-fit

We present the results of the residual analysis in Figure 13, specifically focusing on the deviance component residual r_{Di} discussed in Section 6. In the deviance residual vs index plot, we observe a prominent outlier, indicating a significant deviation from the expected pattern. However, when considering the normal probability plot along with the generated envelope, we find that approximately 93% of the data points fall within the envelope. This suggests that the proposed TI-HT-OBIII-W regression model is indeed appropriate for these data, as only a single observation lies outside the expected range.

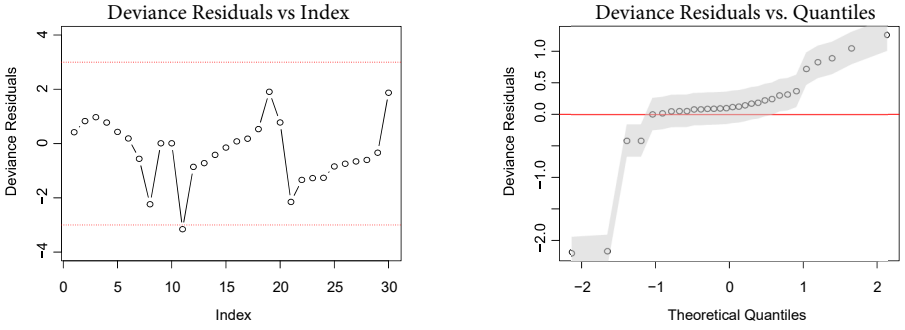


Figure 13: Plot of deviance residuals against index and normal probability plot with envelopes for the deviance component residuals.

8. Summary

We developed and presented a new heavy-tailed FoD called the type I heavy-tailed odd Burr III-G (TI-HT-OBIII-G) distribution. Statistical properties of this new FoD were derived and presented. The MLE technique was utilized in the estimation of parameters. Actuarial risk measures were computed. Numerical comparisons of actuarial measures with other distributions were conducted and the results were presented. The regression model and the analysis of residuals were examined in the context of the new distribution. Finally, the superiority of the TI-HT-OBIII-G FoD was illustrated by the Kevlar 49/epoxy strands and turbocharger data sets. We recommend bivariate regression models and different parameter estimation techniques for future research.

***To access the appendix, kindly click the link provided below*:**
https://drive.google.com/file/d/1KEtYcoHXU3FhE4OK8DvdFSg5v01_Qi1M/view?usp=sharing

References

Afify, A. Z., Gemeay, A. M. and Ibrahim, N. A., (2020). The Heavy-Tailed Exponential Distribution: Risk Measures, Estimation, and Application to Actuarial Data. *Mathematics*, 8(8), 1276.

Alizadeh, M., Cordeiro, G., Nascimento, A. and M. Ortega, E. M. M. (2017). Odd Burr Generalized Family of Distributions with some Applications. *Journal of Statistical Computation and Simulation*, 87, pp. 367–389.

Alyami, S. A., Babu, M. G., Elbatal, I., Alotaibi, N. and Elgarhy, M., (2022). Type II Half Logistical Odd Fréchet Class of Distributions: Statistical Theory and Applications. *Symmetry*, 14(6), 1222. https://doi.org/10.3390/sym14061222

Atiknson, A. C., (1985). Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. *Clarendon Press Oxford*. https://doi.org/10.1007/s40300-013-0007

Benkhelifa, L., (2022). Alpha Power Topp-Leone Weibull Distribution: Properties, Characterizations, Regression Modeling and Applications. *Journal of Statistics and Management Systems*, 25(8), pp. 1–26.

Cooray, K., Ananda, M. M., (2008). A Generalization of the Half-normal Distribution with Applications to Lifetime Data. *Communications in Statistics Theory and Methods*, 37(9), pp. 1323–1337.

Cordeiro, G. M., Alizadeh, M. and Ortega, E. M., (2014). The Exponentiated Half-Logistic Family of Distributions: Properties and Applications. *Journal of Probability and Statistics*. https://doi.org/10.1155/2014/864396

Cordeiro, G. M., Ortega, E. M. and Nadarajah, S., (2010). The Kumaraswamy Weibull Distribution with Application to Failure Data. *Journal of the Franklin Institute*, 347(8), pp.1399–1429.

Dey, S., Nassar. M. and Kumar, D., (2019). Alpha Power Transformed Inverse Lindley Distribution. A Distribution with an Upside-down Bathtub shaped Hazard Function. *Journal of Computational and Applied Mathematics*, 348, pp. 130-145.

Descheemaeker, L., Grilli. J. and de Buyl, S., (2021). Heavy-tailed Abundance Distributions from Stochastic Lotka-Volterra models. *American Physical Society*, 104(38), pp. 034404-034413.

Korkmaz, M. C., Yousof, H. M. and Hamedani, G. G., (2018). The Exponential Lindley odd Log-Logistic-G Family: Properties, Characterizations and Applications. *Journal of Statistical Theory and Applications*, 17(3), pp. 554–571.

Nelson, W. B., (2004). Accelerated Testing: Statistical Models, Test Plans, and Data Analysis. *John Wiley and sons*.

Rényi, A., (1960). On Measures of Entropy and Information. Proceedings of the Fourth Berkeley. *Symposium on Mathematical Statistics and Probability*.

Shannon, C. E., (1951). Prediction and Entropy of Printed English. *The Bell System Technical Journal*.

Tahir, M., Cordeiro, G. M. and Zubair, M., (2014). The Weibull-Lomax distribution: Properties and Applications. *Hacettepe University Bulletin of Natural Sciences and Engineering Series B: Mathematics and Statistics*, 10, pp. 147-465.

Teamah, A. E. A., Elbanna, A. A. and Gemeay, A. M., (2021). Heavy-tailed Log-Logistic Distribution: Properties, Risk Measures and Applications. *Statistics, Optimization, and Information Computing*, 9(4), pp. 910-941.

Xu, K., Xie, M., Tang, Ching, L. and Ho, S. L., (2003). Application of Neural Networks in Forecasting Engine Systems Reliability. *Applied Soft Computing*, 2, pp. 255-268.

Yousof, H. M., Afify, A. Z., Alizadeh, M., Butt, N. S., Hamedani, G. and Ali, M. M., (2015). The Transmuted Exponentiated Generalized-G Family of Distributions. *Pakistan Journal of Statistics and Operation Research*, 11(4), pp. 441-464.

Zhao, W., Khosa, S. K., Ahmad, Z., Aslam M., and Afify, A. Z., (2020). Type-I Heavy-Tailed Family with Applications in Medicine, Engineering, and Insurance. *PLoS ONE*, 15(8). https://doi.org/10.1371/journal.pone.0237462.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 117–134, https://doi.org/10.59139/stattrans-2025-007
Received – 07.01.2021; accepted – 09.10.2024

On a family that unifies the generalized Marshall-Olkin and Poisson-G family of distributions

Laba Handique[1], Farrukh Jamal[2], Subrata Chakraborty[3]

Abstract

The aim of the article is to propose a unification of the generalized Marshall-Olkin (GMO) and Poisson-G (P-G) distributions into a new family of distributions. The density and survival function are expressed as infinite mixtures of an exponentiated-P-G family. The quantile function, asymptotes, shapes, stochastic ordering and Rényi entropy are derived. The paper presents a maximum likelihood estimation with large sample properties. A Monte Carlo simulation is used to examine the pattern of the bias and the mean square error of the maximum likelihood estimators. The utility of the proposed family is illustrated through its comparison with some important models and sub models of the family in terms of modeling real data.

Key words: GMO family, Poisson-G family, stochastic ordering, MLE, AIC.

1. Introduction

Generalized classes of univariate continuous distributions through introduction of additional shape parameter(s) to a baseline distribution have attracted a lot of attention in recent times. With the basic motivation to bring in more flexibility in the modelling different types of data, a preferred area of research in the field of probability distribution is that of generating new distributions starting with a baseline distribution by inducing one or more additional parameters through various methodologies. A number of useful continuous univariate-G families have been added in the literature in recent times. Notable families introduced since 2017 are Poisson-G family (Abouelmagd et al., 2017), Marshall-OlkinKumaraswamy-G family (Handique et al.,

[1]Corresponding author. Department of Statistics, Darrang College, Tezpur, Assam-784001, India. E-mail:handiquelaba@gmail.com. ORCID: https://orcid.org/0000-0001-9255-2918

[2] Department of Statistics, The Islamia University, Bahawalpur 63100, Pakistan. E-mail: drfarrukh1982@gmail.com. ORCID: https://orcid.org/0000-0001-6192-9890.

[3] Department of Statistics, Dibrugarh University, Dibrugarh-786004, India. E-mail: subrata_stats@dibru.ac.in. ORCID: https://orcid.org/0000-0002-6405-1486.

2017), Generalized Marshall-Olkin Kumaraswamy-G family (Chakraborty and Handique, 2017), Exponentiated generalized-G Poisson family (Gokarna and Haitham, 2017), Beta Kumaraswamy-G family (Handique et al., 2017), Beta generated Kumaraswamy Marshall-Olkin-G family (Handique and Chakraborty, 2017a), Beta generalized Marshall-Olkin Kumaraswamy-G family (Handique and Chakraborty, 2017b), Beta generated Marshall-Olkin Kumaraswamy-G (Chakraborty et al., 2018), Kumaraswamy generalized Marshall-Olkin-G family (Chakraborty and Handique, 2018), Odd modified exponential generalized family (Ahsan et al., 2018), Zografos-Balakrishnan Burr XII family (Emrah et al., 2018), Exponentiated generalized Marshall-Olkin-G family by (Handique et al., 2019), Beta-G Poisson family (Gokarna et al., 2019), Zero truncated Poisson family (Abouelmagd et al., 2019), Extended generalized Gompertz family (Thiago et al., 2019), Generalized modified exponential-G family (Handique et al., 2020), Odd Half-Cauchy family (Chakraborty et al., 2021), Poisson Transmuted-G family (Handique et al., 2021), Beta Poisson-G family (Handique et al., 2022), Kumaraswamy Poisson-G family (Chakraborty et al., 2022), Complementary Geometric-Topp-Leone-G family (Handique et al., 2023), generalized Marshall-Olkin Transmuted-G family (Handique et al., 2024) and Truncated Cauchy Power Kumaraswamy-G family (Ibrahim et al., 2024), among others.

In this article, a new family of continuous probability distribution called the Generalized Marshall-Olkin Poisson-G ($GMOP-G(\theta,\alpha,\lambda)$) is introduced to unify generalized Marshall-Olkin (GMO) of Jayakumar and Mathew, (2008) and the Poisson-G (P-G) family of distribution (Tahir et al., 2016). Now, we briefly describe the GMO and P-G family and then introduce GMOP-G in the next section.

1.1. Generalized Marshall-Olkin (GMO) family

Jayakumar and Mathew (2008) proposed a new generalization of the Marshall-Olkin family (Marshall and Olkin, 1997) of distributions called the generalized Marshall-Olkin (GMO) family of distributions. The survival function (sf) and probability distribution function (pdf) of the GMO distribution are given respectively by

$$\bar{F}^{\mathrm{GMO}}(t\,;\theta\,,\alpha)=\left[\frac{\alpha\bar{F}(t)}{1-\bar{\alpha}\bar{F}(t)}\right]^{\theta}\text{ and }f^{\mathrm{GMO}}(t\,;\theta\,,\alpha)=\frac{\theta\alpha^{\theta}f(t)\bar{F}(t)^{\theta-1}}{[1-\bar{\alpha}\bar{F}(t)]^{\theta+1}},\tag{1}$$

where $-\infty<t<\infty,\alpha>0$ ($\bar{\alpha}=1-\alpha$), $\theta>0$ is an additional shape parameter and $\bar{F}(t)$ and$f(t)$ is the baseline sf and pdf respectively.

When $\theta=1$, $\bar{F}^{\mathrm{GMO}}(t\,;\theta\,,\alpha)=\bar{F}^{\mathrm{MO}}(t\,;\alpha)$ and for $\alpha=\theta=1$, $\bar{F}^{\mathrm{GMO}}(t\,;\theta\,,\alpha)=\bar{F}(t)$.

1.2. Poisson-G (P-G) family

The Poisson-G (P-G) family of distributions with survival function and cdf is given by (see Kumaraswamy Poisson-G family, Chakraborty et al., 2022)

$$\bar{G}^{\mathrm{PG}}(t;\lambda) = \frac{e^{-\lambda G(t)} - e^{-\lambda}}{1 - e^{-\lambda}}$$

$$and\ \ G^{\mathrm{PG}}(t;\lambda) = \frac{1-e^{-\lambda G(t)}}{1-e^{-\lambda}},\ \lambda \in R - \{0\}; n = 1, 2, \ldots \tag{2}$$

The corresponding pdf of (2) is given by

$$g^{\mathrm{PG}}(t;\lambda) = (1 - e^{-\lambda})^{-1}\lambda g(t)e^{-\lambda G(t)}, \lambda \in R - \{0\}; -\infty < t < \infty. \tag{3}$$

where $G(t)$and$g(t)$ is the baseline distribution.

The article is arranged in the following 5 sections. In Section 2, we introduce the proposed family along with its physical basis and list of some important sub models and also define some mathematical properties. In Section 3, a linear representation of the sf and pdf of the proposed family is discussed along with some statistical properties of the proposed family. In Section 4, maximum likelihood methods of estimation of parameters and simulation are presented. The data fitting applications are presented in Section 5. Final conclusion is provided in Section 6.

2. Generalized Marshall-Olkin Poisson-G family

In this section we introduce the $GMOP - G(\theta, \alpha, \lambda)$ family and also provide its special cases and a statistical genesis.

The sf, cdf, pdf and hrf of this distribution are given respectively by:

$$\bar{F}^{\mathrm{GMOPG}}(t;\theta,\alpha,\lambda) = \left[\frac{\alpha(e^{-\lambda G(t)} - e^{-\lambda})}{1-\alpha e^{-\lambda} - \bar{\alpha} e^{-\lambda G(t)}}\right]^{\theta}, F^{\mathrm{GMOPG}}(t;\theta,\alpha,\lambda) = 1 - \left[\frac{\alpha(e^{-\lambda G(t)} - e^{-\lambda})}{1-\alpha e^{-\lambda} - \bar{\alpha} e^{-\lambda G(t)}}\right]^{\theta} \tag{4}$$

$$f^{\mathrm{GMOPG}}(t;\theta,\alpha,\lambda) = \frac{\theta\lambda\alpha^{\theta}(1-e^{-\lambda})g(t)e^{-\lambda G(t)}(e^{-\lambda G(t)} - e^{-\lambda})^{\theta-1}}{(1-\alpha e^{-\lambda} - \bar{\alpha} e^{-\lambda G(t)})^{\theta+1}}, \tag{5}$$

and $$h^{\mathrm{GMOPG}}(t;\theta,\alpha,\lambda) = \frac{\theta\lambda(1-e^{-\lambda})g(t)e^{-\lambda G(t)}(e^{-\lambda G(t)} - e^{-\lambda})^{-1}}{1-\alpha e^{-\lambda} - \bar{\alpha} e^{-\lambda G(t)}}. \tag{6}$$

In particular, we get for

(i) $\theta = 1$, the $MOP - G(\alpha, \lambda)$ distribution.

(ii) $\theta = \alpha = 1$, the $P - G(\lambda)$ distribution.

(iii) $\lambda \to 0$, the $GMO(\theta, \alpha)$ distribution.

(iv) $\theta = 1, \lambda \to 0$, the $MO(\alpha)$ distribution.

Proposition 1 Let $T_{i1}, T_{i2}, \ldots, T_{iN}$, $i = 1,2,\cdots,\theta$ be a sequence of θN *i.i.d.* random variables from Poisson-G distribution and $W_i = min(T_{i1}, T_{i2}, \ldots, T_{iN})$ and $V_i = max(T_{i1}, T_{i2}, \ldots, T_{iN})$. Then

(i) $\min_i W_i$ follows $GMOP - G(\theta, \alpha, \lambda)$ if $N \sim G\,eometric(\alpha)$ and

(ii) $\max_i V_i$ follows $GMOP - G(\theta, \alpha, \lambda)$ if $N \sim G\,eometric(1/\alpha)$.

Proof: Case (i) When $0 < \alpha \leq 1$, considering N has a geometric distribution with parameter α, we get

$$P[min\{W_1, W_2, \ldots, W_\theta\} > t] = P[W_1 > t]P[W_2 > t]\ldots P[W_\theta > t]$$

$$= \prod_{i=1}^{\theta} P[W_i > t] = [\bar{F}^{MOPG}(t; \alpha, \lambda)]^\theta = \left[\frac{\alpha(e^{-\lambda G(t)} - e^{-\lambda})}{1 - \alpha e^{-\lambda} - \bar{\alpha} e^{-\lambda G(t)}}\right]^\theta.$$

Case (ii) For $\alpha > 1$, considering N has a geometric distribution with parameter $1/\alpha$, we get

$$P[min\{V_1, V_2, \ldots, V_\theta\} > t] = P[V_1 > t]P[V_2 > t]\ldots P[V_\theta > t]$$

$$= \prod_{i=1}^{\theta} P[V_i > t] = [\bar{F}^{MOPG}(t; \alpha, \lambda)]^\theta = \left[\frac{\alpha(e^{-\lambda G(t)} - e^{-\lambda})}{1 - \alpha e^{-\lambda} - \bar{\alpha} e^{-\lambda G(t)}}\right]^\theta.$$

In what follows we investigate some general properties, parameter estimation and real life applications.

2.1. Special model and shape of the density and hazard function

In this section we have plotted the pdf and hrf of the $GMOP - E(\theta, \alpha, \lambda, \beta)$ for some chosen values of the parameters in Figure 1 and Figure 2 respectively to show the variety of shapes assumed by the family.

The pdf and hrf of the $GMOP - E(\theta, \alpha, \lambda, \beta)$ are as follows:

- The GMOP-Exponential (GMOP-E) distribution.

Considering the Exponential distribution with parameters $\beta > 0$ having pdf and cdf $g(t) = \beta e^{-\beta t}$ and $G(t) = 1 - e^{-\beta t}$ respectively we get the pdf and hrf of $GMOP - E(\theta, \alpha, \lambda, \beta)$ distribution as

$$f^{\text{GMOPE}}(t\,; \theta\,, \alpha, \lambda, \beta) = \frac{\theta\lambda\alpha^\theta(1-e^{-\lambda})\beta e^{-\beta t} e^{-\lambda(1-e^{-\beta t})}(e^{-\lambda(1-e^{-\beta t})} - e^{-\lambda})^{\theta-1}}{(1 - \alpha e^{-\lambda} - \bar{\alpha} e^{-\lambda(1-e^{-\beta t})})^{\theta+1}},$$

and $h^{\text{GMOPE}}(t\,; \theta\,, \alpha, \lambda, \beta) = \frac{\theta\lambda\alpha^\theta(1-e^{-\lambda})\beta e^{-\beta t} e^{-\lambda(1-e^{-\beta t})}(e^{-\lambda(1-e^{-\beta t})} - e^{-\lambda})^{-1}}{1 - \alpha e^{-\lambda} - \bar{\alpha} e^{-\lambda(1-e^{-\beta t})}}$.

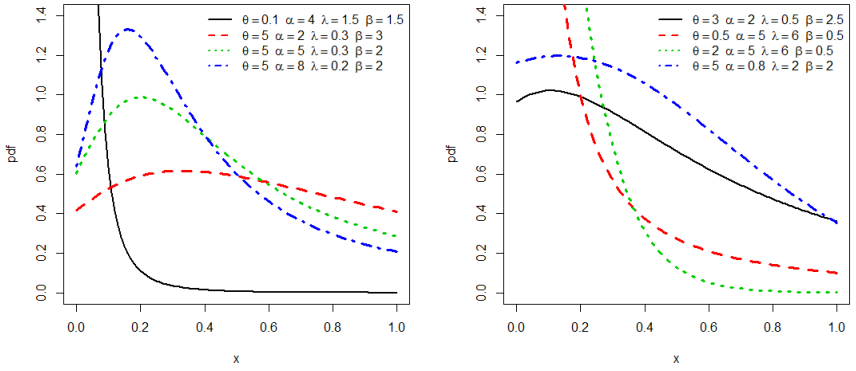


Figure 1: pdf plots of the $GMOP-E(\theta,\alpha,\lambda,\beta)$ distribution

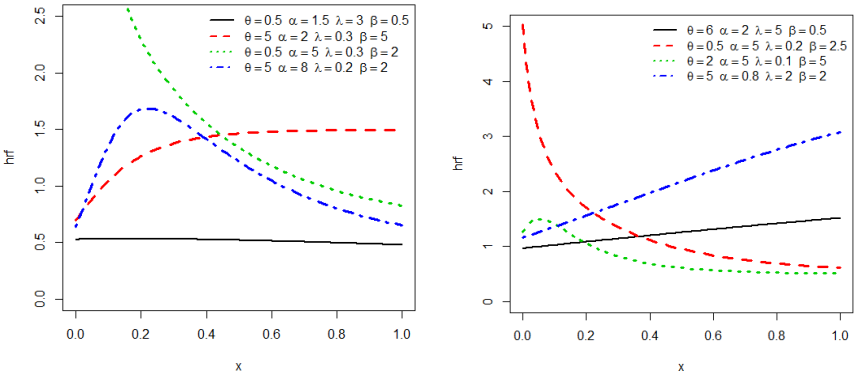


Figure 2: hrf plots of the $GMOP-E(\theta,\alpha,\lambda,\beta)$ distribution

Remark 1. From Figures 1 and 2 it can be seen that the proposed family of distributions is very flexible and can offer different types of shapes for density and hazard like increasing, decreasing and right skewed.

Quantile and related measures

The p^{th} quantile t_p for $GMOP-G(\theta,\alpha,\lambda)$ can be easily obtained by solving the equation $F^{\text{GMOPG}}(t)=p$ as

$$t_p = G^{-1}\left[-\frac{1}{\lambda}log\left[\frac{\alpha e^{-\lambda}+(1-\alpha e^{-\lambda})(1-F(t))^{1/\theta}}{\alpha+\bar{\alpha}(1-F(t))^{1/\theta}}\right]\right].$$

Here the flexibility of skewness and kurtosis of $GMOP-G(\theta,\alpha,\lambda)$ is checked by plotting Galton skewness (S) that measures the degree of the long tail and Moors (1988) kurtosis (K) that measures the degree of tail heaviness in Figure 3 for the $GMOP-$

$E(\theta, \alpha, \lambda, \beta)$ distribution for some values of parameters. These are respectively defined by

$$S = \frac{Q(6/8)-2Q(4/8)+Q(2/8)}{Q(6/8)-Q(2/8)} \text{ and } K = \frac{Q(7/8)-Q(5/8)+Q(3/8)-Q(1/8)}{Q(6/8)-Q(2/8)}.$$

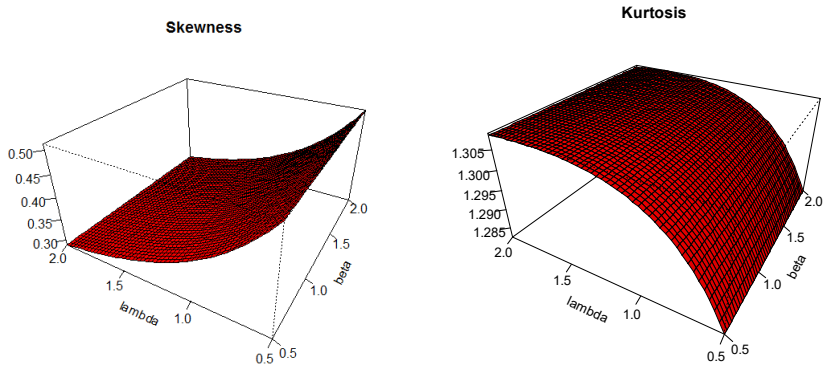


Figure 3: Plots of the Galton skewness S and the Moor kurtosis K for the GMOP-E distribution with parameters $\theta = 3, \alpha = 2, 0.2 < \lambda, \beta < 2$

2.2. Asymptotes and shapes

Two propositions regarding asymptotes of the proposed family are discussed here.

Proposition 2 The asymptotes of pdf, cdf and hrf of $GMOP - G(\theta, \alpha, \lambda)$ as $t \to 0$ are given by

$$f^{\text{GMOPG}}(t\,;\theta\,,\alpha,\lambda) \sim \frac{\theta\lambda g(t)}{\alpha(1-e^{-\lambda})},$$
$$F^{\text{GMOPG}}(t\,;\theta\,,\alpha,\lambda) \sim 0 \text{ and}$$
$$h^{\text{GMOPG}}(t\,;\theta\,,\alpha,\lambda) \sim \frac{\theta\lambda g(t)}{\alpha(1-e^{-\lambda})}.$$

Proposition 3 The asymptotes of pdf, cdf and hrf of $GMOP - G(\theta, \alpha, \lambda)$ as $t \to \infty$ are given by

$$f^{\text{GMOPG}}(t\,;\theta\,,\alpha,\lambda) \sim \theta\,\lambda\alpha^{\theta}e^{-\lambda}g(t)(e^{-\lambda G(t)} - e^{-\lambda})^{\theta-1}/(1-e^{-\lambda})^{\theta},$$
$$F^{\text{GMOPG}}(t\,;\theta\,,\alpha,\lambda) \sim 1 - \alpha^{\theta}(e^{-\lambda G(t)} - e^{-\lambda})^{\theta}/(1-e^{-\lambda})^{\theta} \text{ and}$$
$$h^{\text{GMOPG}}(t\,;\theta\,,\alpha,\lambda) \sim \theta\,\lambda e^{-\lambda}g(t)(e^{-\lambda G(t)} - e^{-\lambda})^{-1}.$$

Analytically the shapes of the pdf and hazard rate function can be stated through critical points. The critical points of the pdf are the roots of the equation

$$\frac{g'(t)}{g(t)} - \lambda g(t) - (\theta-1)\frac{\lambda e^{-\lambda G(t)}g(t)}{e^{-\lambda G(t)}-e^{-\lambda}} - (\theta+1)\frac{\bar{\alpha}\lambda e^{-\lambda G(t)}g(t)}{1-\alpha e^{-\lambda}-\bar{\alpha}e^{-\lambda G(t)}} = 0. \qquad (7)$$

The critical point of $GMOP-G(\theta,\alpha,\lambda)$ family hazard rate are the roots of the equation

$$\frac{g'(t)}{g(t)}-\lambda g(t)+\frac{\lambda e^{-\lambda G(t)}g(t)}{e^{-\lambda G(t)}-e^{-\lambda}}-\frac{\bar{\alpha}\lambda e^{-\lambda G(t)}g(t)}{1-\alpha e^{-\lambda}-\bar{\alpha}e^{-\lambda G(t)}}=0. \tag{8}$$

Equations (7) and (8) may have multiple solutions. If $t=t_0$ is a root then it is a local maximum, a local minimum or a point of inflexion if $\psi(t_0)<0,\psi(t_0)>0 or \psi(t_0)=0$ and for (8) if $\omega(t_0)<0,\omega(t_0)>0 or \omega(t_0)=0$ where $\psi(t)=(d^2/dt^2)\,log[\,f(t)]$ and $\omega(t)=(d^2/dt^2)\,log[\,h(t)]$

2.3. Stochastic orderings

Let X and Y be two random variables with cdfs F and G, respectively, corresponding pdf's f and g. Then X is said to be smaller than Y in the likelihood ratio order ($X\leq_{lr}Y$) if $f(t)/g(t)$ is decreasing in $t\geq 0$. Here we present a result of likelihood ratio ordering.

Theorem 1 Let $X\sim GMOPG\rightleftarrows(\theta,\alpha_1,\lambda)$ and$Y\sim GMOPG(\theta,\alpha_2,\lambda)$. If $\alpha_1<\alpha_2$, then $X\leq_{lr}Y$

Proof: $\frac{f(t)}{g(t)}=\left(\frac{\alpha_1}{\alpha_2}\right)^{\theta}\left[\frac{1-\alpha_2e^{-\lambda}-\bar{\alpha}_2e^{-\lambda G(t)}}{1-\alpha_1e^{-\lambda}-\bar{\alpha}_1e^{-\lambda G(t)}}\right]^{\theta+1}$

$$\frac{d}{dt}(f(t)/g(t))=(\theta+1)(\frac{\alpha_1}{\alpha_2})^{\theta}(\alpha_1-\alpha_2)\frac{[1-\alpha_2e^{-\lambda}-\bar{\alpha}_2e^{-\lambda G(t)}]^{\theta}\lambda e^{-\lambda G(t)}g(t)(1-e^{-\lambda})}{[1-\alpha_1e^{-\lambda}-\bar{\alpha}_1e^{-\lambda G(t)}]^{\theta+2}}.$$

Now this is always less than 0, since $\alpha_1<\alpha_2$. Hence, $f(t)/g(t)$ is decreasing in t. That is $X\leq_{lr}Y$.

3. Linear representation

Linear representation of sf and pdf, etc. in terms of corresponding functions of known distributions is an important tool for further mathematical properties. In this section we present some important results for the proposed family.

3.1. Expansions of the survival and density functions as infinite linear mixture

Here the sf and pdf of the $GMOP-G(\theta,\alpha,\lambda)$ are expressed as linear mixture of the corresponding functions of exponentiated-$P-G(\lambda)$ distribution.

Consider the series representation

$$(1-z)^{-k}=\sum_{j=0}^{\infty}\frac{\Gamma(k+j)}{\Gamma(k)j!}z^j=\sum_{j=0}^{\infty}\frac{(j+k-1)!}{(k-1)!j!}z^j,|z|<1 \text{ and } k>0, \tag{9}$$

where $\Gamma(.)$is the gamma function.

Using equation (9) in equation (4), for $\alpha \in (0,1)$ we obtain

$$\bar{F}^{GMOPG}(t;\theta,\alpha,\lambda) = \alpha^{\theta}\{\bar{G}^{\mathrm{PG}}(t;\lambda)\}^{\theta}\sum_{j=0}^{\infty}\frac{(j+\theta-1)!}{(\theta-1)!\,j!}(1-\alpha)^{j}\{\bar{G}^{\mathrm{PG}}(t;\lambda)\}^{j}$$

$$= \sum_{j=0}^{\infty}\eta'_{j}\,[\bar{G}^{\mathrm{PG}}(t;\lambda)]^{j+\theta}. \tag{10}$$

Differentiating in equation (10) with respect to 't' we get

$$f^{GMOPG}(t;\theta,\alpha,\lambda) = g^{PG}(t;\lambda)\sum_{j=0}^{\infty}\eta_{j}[\bar{G}^{PG}(t;\lambda)]^{j+\theta-1} \tag{11}$$

$$= -\sum_{j=0}^{\infty}\eta'_{j}\frac{d}{dt}[\bar{G}^{\mathrm{PG}}(t;\lambda)]^{j+\theta} \tag{12}$$

where $\eta'_{j} = \eta'_{j}(\alpha) = \binom{j+\theta-1}{j}(1-\alpha)^{j}\alpha^{\theta}, \quad \eta_{j} = \eta_{j}(\alpha) = (j+\theta)\eta'_{j}$.

We have presented a numerical evaluation result of mean, variance, skewness and kurtosis of the $GMOP-E(\theta,\alpha,\lambda,\beta)$ distribution for some selected parameter values in Table 1.

Table 1: Mean, variance, skewness and kurtosis of the $GMOP-E(\theta,\alpha,\lambda,\beta)$ distribution with different values of θ, α, λ and β

θ	α	λ	β	Mean	Variance	Skewness	Kurtosis
10	10	2	2	0.1572	0.0163	1.2096	4.6452
10	10	1	2	0.2224	0.0306	1.0761	4.0834
10	10	0.5	2	0.2648	0.0408	0.9749	3.7373
10	10	0.1	2	0.3033	0.0502	0.8840	3.4698
10	10	2	1	0.3145	0.0652	1.2096	4.6452
10	10	2	0.5	0.6291	0.2610	1.2096	4.6452
10	10	0.5	0.5	1.0592	0.6528	0.9749	3.7373
10	10	0.1	0.1	6.0675	20.1022	0.8840	3.4698
10	5	2	2	0.0918	0.0066	1.5144	6.0241
10	2	2	2	0.0419	0.0016	1.9171	8.4744
10	0.5	2	2	0.0115	0.0001	2.4163	12.8206
10	0.5	0.5	0.5	0.0834	0.0077	2.3537	12.1190
5	10	2	2	0.2692	0.0424	1.0978	4.3497
5	5	2	2	0.1676	0.0206	1.4448	5.8193
2	5	2	2	0.3615	0.0914	1.5105	6.3958
2	2	2	2	0.2049	0.0427	2.1968	10.7924
1	2	2	2	0.4180	0.19284	2.3136	11.2913
5	0.1	0.1	0.1	0.2309	0.0804	3.6320	30.5202
5	5	5	3	0.0489	0.0016	1.4027	5.7915
5	5	3	5	0.04882	0.0017	1.5071	6.2776
5	5	5	8	0.0183	0.0002	1.3967	5.7764
5	5	10	10	0.0069	0.00001	1.1168	2.1041

3.2. Rényi entropy

Entropy of a random variable is a measure of uncertainty and has been used in various situations in science and engineering. The Rényi entropy (see details, Song, 2001) is defined by

$$I_R(\delta) = (1-\delta)^{-1} log\left(\int_{-\infty}^{\infty} f(t)^{\delta} dt\right), \text{ where } \delta > 0 \text{ and } \delta \neq 1.$$

Thus the Rényi entropy of $GMOP - G(\theta, \alpha, \lambda)$ distribution can be obtained as

$$I_R(\delta) = (1-\delta)^{-1} log\left(\sum_{j=0}^{\infty} \mu_j \int_{-\infty}^{\infty} [g^{\mathrm{PG}}(t;\lambda)\bar{G}^{\mathrm{PG}}(t;\lambda)^{\theta-1}]^{\delta} [\bar{G}^{\mathrm{PG}}(t;\lambda)]^{j} dt\right),$$

where $\mu_j = \mu_j(\alpha) = \{\theta^{\delta}\alpha^{\delta\theta}(1-\alpha)^{j}\Gamma[\delta(\theta+1)+j]\}/\{\Gamma[\delta(\theta+1)]j!\}$.

Table 2 shows the values of numerical values of Rényi entropy $GMOP - E(\theta, \alpha, \lambda, \beta)$ distribution for some selected parameter values. As expected, the Rényi entropy turns out to be non-increasing with δ.

Table 2: Rényi entropy $GMOP - E(\theta, \alpha, \lambda, \beta)$ distribution with different values of θ, α, λ and β

Parameter				δ					
θ	α	λ	β	**0.2**	**0.5**	**1.5**	**2**	**3**	**5**
10	10	**2**	**2**	-0.2550	-0.6403	-0.9916	-1.0647	-1.1549	-1.2490
5	5	0.5	0.5	1.6816	1.3053	0.9722	0.9032	0.8176	0.7280
5	5	2	0.5	1.3469	0.8661	0.4519	0.3687	0.2669	0.1621
3	3	2	0.5	1.6090	1.0220	0.5252	0.42839	0.3113	0.1924
1.5	1.5	2	0.5	2.1288	1.3773	0.6945	0.5640	0.4097	0.2571
2	0.5	0.5	0.5	1.7182	0.8594	0.4390	-0.1133	-0.2982	-0.4789

4. Estimation

Here, we consider the parameter estimation of $GMOP - G(\theta, \alpha, \lambda)$ via the maximum likelihood (ML) method.

4.1 Maximum likelihood method

Let $T = (t_1, t_2, \ldots, t_n)$ be a random sample of size n from $GMOP - G(\theta, \alpha, \lambda)$ with parameter vector $\rho = (\theta, \alpha, \lambda, \xi)$, where $\xi = (\xi_1, \xi_2, \ldots, \xi_q)$ is the parameter vector of G. Then, the log-likelihood function for ρ is given by

$$\ell = \ell(\boldsymbol{\rho}) = n\log(\theta\lambda\alpha^{\theta}) - n\log(1-e^{-\lambda}) + \sum_{i=1}^{n}\log[g(t_i,\boldsymbol{\xi})] - \lambda\sum_{i=1}^{n}[G(t_i,\boldsymbol{\xi})]$$

$$+ (\theta-1)\sum_{i=1}^{n}\log(e^{-\lambda G(t_i,\boldsymbol{\xi})} - e^{-\lambda}) - (\theta+1)\sum_{i=1}^{n}\log(1-\alpha e^{-\lambda} - \overline{\alpha}e^{-\lambda G(t_i,\boldsymbol{\xi})}).$$

Due to its complex form, this function cannot be solved precisely, but it can be numerically maximized by using optimization methods available with the software R.

We obtain the components of the score vector $U_{\boldsymbol{\rho}} = (U_\theta, U_\alpha, U_\lambda, U_{\boldsymbol{\xi}})$ by taking the partial derivatives of the log-likelihood function with respect to θ, α, λ and $\boldsymbol{\xi}$.

The asymptotic variance-covariance matrix of the MLEs of parameters is obtained by inverting the Fisher information matrix $I(\rho)$ derived using the second partial derivatives of the log-likelihood function with respect to each parameter. The ij^{th} elements of $I_n(\rho)$ are given by

$$I_{ij} = -E[\partial^2 l(\rho)/\partial\rho_i \partial\rho_j],\ i,j = 1,2,\cdots,3+q.$$

In practice one can estimate $I_n(\rho)$ by the observed Fisher's information matrix $\hat{I}_n(\hat{\rho}) = (\hat{I}_{ij})$ defined as

$$\hat{I}_{ij} \approx \left(-\partial^2 l(\rho)/\partial\rho_i \partial\rho_j\right)_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}},\ i,j = 1,2,\cdots,3+q.$$

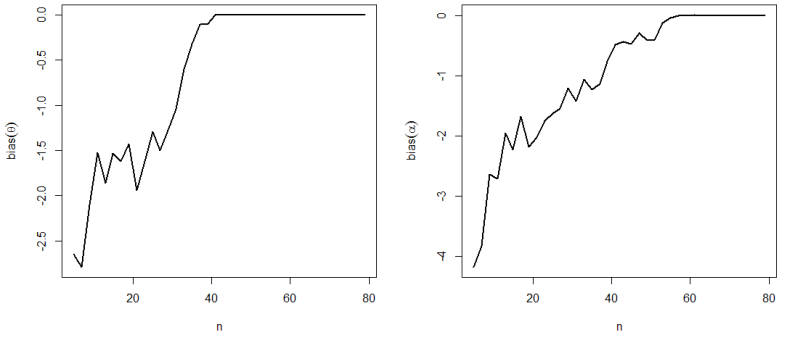
From the asymptotic theory of MLEs under some regularity conditions on the parameters as $n \to \infty$ the asymptotic distribution of $\sqrt{n}(\hat{\rho} - \rho)$ is $N_k(0, V_n)$ where $V_n = (v_{jj}) = I_n^{-1}(\rho)$. This holds even if V_n is replaced by $\hat{V}_n = \hat{I}^{-1}(\hat{\rho})$. Using this result large sample standard errors of j^{th} parameter ρ_j is given by $\sqrt{\hat{v}_{jj}}$.

4.2 Simulation

Here, a Monte Carlo simulation study is conducted to compare the performance of the different estimators of the unknown parameters for the $GMOP-E(\theta,\alpha,\lambda,\beta)$ distribution using R program. We generate $N = 3000$ samples of size $n = 5\text{to}80$from GMOP-E distribution with true parameter values $\theta = 2, \alpha = 8, \lambda = 5, \beta = 0.5$, and calculate the bias and mean square error (MSE) of the MLEs empirically by

$Bias_h = \frac{1}{N}\sum_{i=1}^{N}(\hat{h}_i - h)$ and $MSE_h = \frac{1}{N}\sum_{i=1}^{N}(\hat{h}_i - h)^2$respectively (for$h = \theta, \alpha, \lambda, \beta$).

Results of this simulation study are presented graphically in Figures 4 and 5 and tell us that as the sample sizes increases the biases and MSE's approach to 0 in all caess, which is consistent with the theoretical properties of the MLE and hence appropriate for estimating the $GMOP-E(\theta,\alpha,\lambda,\beta)$ distribution parameters.



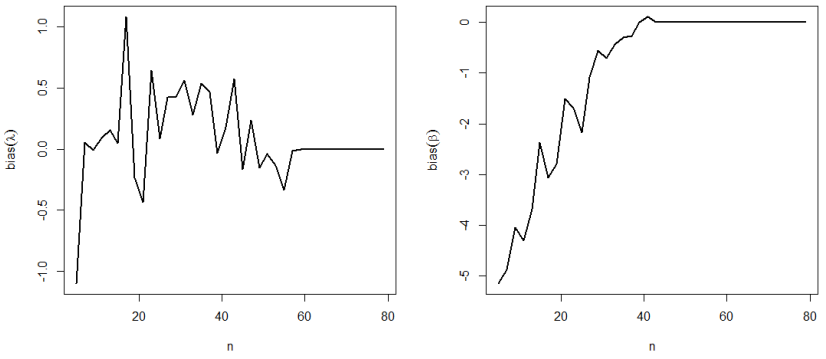


Figure 4: The Biases for the parameter values $\theta = 2, \alpha = 8, \lambda = 5,\ \beta = 0.5$ for $GMOP - E(\theta, \alpha, \lambda, \beta)$ distribution

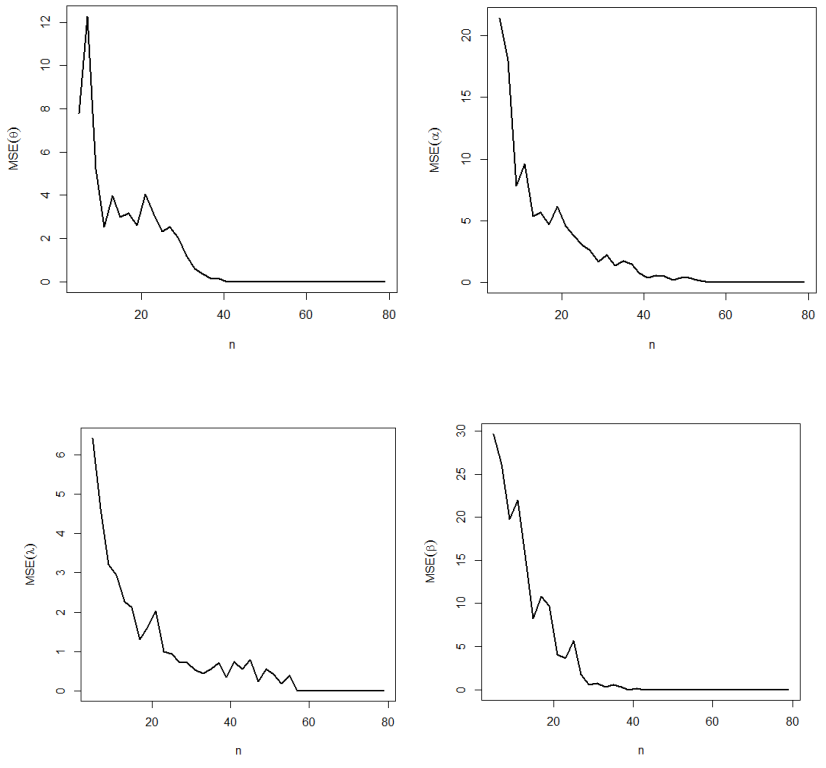


Figure 5: The MSEs for the parameter values $\theta = 2, \alpha = 8, \lambda = 5, \beta = 0.5$ for $GMOP - E(\theta, \alpha, \lambda, \beta)$ distribution

5. A Real Data Application

Here, we consider modelling of the one failure time data set to illustrate the suitability of the $GMOP-E(\theta,\alpha,\lambda,\beta)$ distribution in comparison to some existing distributions by estimating the parameters by numerical maximization of log-likelihood functions. The data set consists of survival time of 72 guinea pigs infected with virulent tubercle bacilli, reported by Bjerkedal (1960). The descriptive statistics about the data set shown in Table 3 reveal that the data set is positively skewed as expected from the nature of life time data and has higher kurtosis.

Table 3: Descriptive Statistics for the guinea pigs survival time's data set

Data Set	n	Min.	Mean	Median	s.d.	Skewness	Kurtosis	1st Qu.	3rd Qu.	Max.
I	72	0.100	1.851	1.560	1.200	1.788	4.157	1.080	2.303	7.000

We have compared the $GMOP-E(\theta,\alpha,\lambda,\beta)$ distribution with exponential (Exp), moment exponential (ME), transmuted exponential (T-E), Marshall-Olkin exponential (MO-E) (Marshall and Olkin, 1997), generalized Marshall-Olkin exponential (GMO-E) (Jayakumar and Mathew, 2008) and Marshall-Olkin transmuted exponential (MOT-E), Kumaraswamy exponential (Kw-E) (Cordeiro and de Castro, 2011), Beta exponential (BE) (Eugene et al., 2002), Marshall-Olkin Kumaraswamy exponential (MOKw-E) (Handique et al., 2017), Kumaraswamy Marshall-Olkin exponential (KwMO-E) (Alizadeh et al., 2015), beta Poisson exponential (BP-E) (Handique et al., 2022) and Kumaraswamy Poisson exponential (KwP-E) (Chakraborty et al., 2022) distributions for the failure time data set.

Model with the lowest AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), CAIC (Consistent Akaike Information Criterion), and HQIC (Hannan-Quinn Information Criterion) is chosen as the best. Also, to verify which distribution fits better these data goodness-of-fit tests, Anderson-Darling (A), Cram′er-von Mises (W) and Kolmogorov-Smirnov (K-S) statistics are applied. Asymptotic standard errors of the MLEs for each competing model are also provided. The best fitted density and the fitted cdf are plotted with the corresponding observed histograms and ogives in Figure 7, which indicates that the proposed distributions provide a close fit to this data set.

To check the shape of the observed hazard function the total time on test (TTT) plot Aarset, (1987) is used. A straight diagonal line indicates constant hazard for the data set, whereas a convex (concave) shape implies decreasing (increasing) hazard. The TTT plots for the data set Figure 6 indicate that the data set has increasing hazard rate. We also provide the box plot of the data to summerize the minimum, first quartile, median, third quartile, and maximum, where a box is shown from the first quartile to the third quartile with a vertical line going through the box at the median.

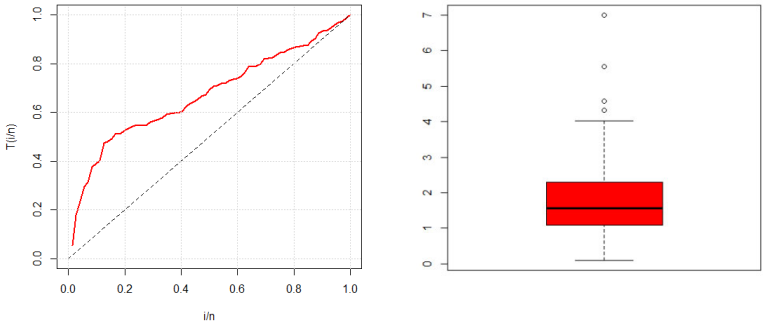


Figure 6: TTT and Box plot for the failure time data set

Table 4: MLEs, standard errors (in parentheses) values for the guinea pigs survival time's data set

Models	$\hat{\theta}$	$\hat{\alpha}$	$\hat{a}$	$\hat{b}$	$\hat{\lambda}$	$\hat{\beta}$
Exp	---	---	---	---	---	0.540
(β)						(0.063)
ME	---	---	---	---	---	0.925
(β)						(0.077)
T-E	---	---	---	---	-0.812	1.041
(λ, β)					(0.038)	(0.105)
MO-E	---	8.778	---	---	---	1.379
(α, β)		(3.555)				(0.193)
GMO-E	0.179	47.635	---	---	---	4.465
(θ, α, β)	(0.070)	(44.901)				(1.327)
MOT-E	---	3.245	---	---	-0.696	1.354
(α, λ, β)		(1.863)			(0.137)	(0.125)
Kw-E	---	---	3.304	1.100	---	1.037
(a, b, β)			(1.106)	(0.764)		(0.614)
B-E	---	---	0.807	3.461	---	1.331
(a, b, β)			(0.696)	(1.003)		(0.855)
MOKw-E	---	0.008	2.716	1.986	---	0.099
(α, a, b, β)		(0.002)	(1.316)	(0.784)		(0.048)
KwMO-E	---	0.373	3.478	3.306	---	0.299
(α, a, b, β)		(0.136)	(0.861)	(0.779)		(1.112)
BP-E	---	---	3.595	0.724	0.014	1.482
(a, b, λ, β)			(1.031)	(1.590)	(0.010)	(0.516)
KwP-E	---	---	3.265	2.658	4.001	0.177
(a, b, λ, β)			(0.991)	(1.984)	(5.670)	(0.226)
GMOP-E	0.333	12.584	---	---	0.054	2.858
$(\theta, \alpha, \lambda, \beta)$	(0.151)	(7.696)			(1.376)	(0.959)

Table 5: Log-likelihood, AIC, BIC, CAIC, HQIC, A, W and KS (p-value) values for the guinea pigs survival times data set

Models	AIC	BIC	CAIC	HQIC	A	W	KS (p-value)
Exp	234.63	236.91	234.68	235.54	6.53	1.25	0.27
(β)							(0.06)
ME	210.40	212.68	210.45	211.30	1.52	0.25	0.14
(β)							(0.13)
T-E	209.94	214.50	210.11	211.74	0.98	0.19	0.10
(λ, β)							(0.17)
MO-E	210.36	214.92	210.53	212.16	1.18	0.17	0.10
(α, β)							(0.43)
GMO-E	210.54	217.38	210.89	213.24	1.02	0.16	0.09
(θ, α, β)							(0.51)
MOT-E	208.26	215.10	208.61	210.96	0.86	0.15	0.10
(α, λ, β)							(0.47)
Kw-E	209.42	216.24	209.77	212.12	0.74	0.11	0.08
(a, b, β)							(0.50)
B-E	207.38	214.22	207.73	210.08	0.98	0.15	0.11
(a, b, β)							(0.34)
MOKw-E	209.44	218.56	210.04	213.04	0.79	0.12	0.10
(α, a, b, β)							(0.44)
KwMO-E	207.82	216.94	208.42	211.42	0.61	0.11	0.08
(α, a, b, β)							(0.73)
BP-E	205.42	214.50	206.02	209.02	0.55	0.08	0.09
(a, b, λ, β)							(0.81)
KwP-E	206.63	215.74	207.23	210.26	0.48	0.07	0.09
(a, b, λ, β)							(0.79)
GMOP-E	204.24	213.36	204.83	207.84	0.44	0.04	0.07
$(\theta, \alpha, \lambda, \beta)$							(0.83)

MLEs of parameters with standard errors for all the fitted models and AIC, BIC, CAIC, HQIC, A, W and K-S statistic with p-value for the failure time data set are presented respectively in Tables 4 and 5. It is obvious from these results that the $GMOP-E$ distribution is not only a better model than the entire sub models but is also better than most of the recently introduced three or four parameters models. The plots in Figure 7 also indicate that the proposed distribution provides a close fit to the data set considered here.

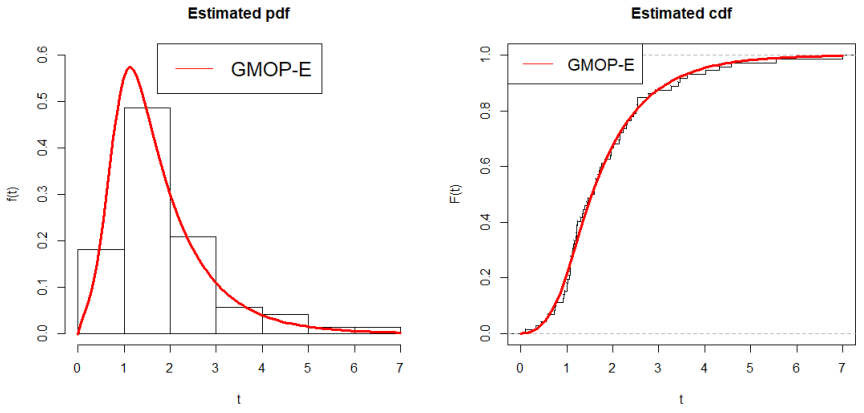


Figure 7: Plots of the observed histogram and estimated pdf on left and on right the observed ogive and estimated cdf for failure time data set for the GMOP-E model

6. Conclusions

In this work, we propose a new family of continuous distributions called the *Generalized Marshall-Olkin Poisson -G* family of distributions. Several new models can be generated by considering special distributions for G. We demonstrate that the pdf of any GMOPG distribution can be expressed as a linear combination of exponentiated-G density functions, which allowed us to derive some of its mathematical and statistical properties. The estimations of the model parameters are obtained by the maximum likelihood method. One application of the proposed family empirically proves its flexibility to model real data sets. In particular, we verified that a special case of the GMOPG family can provide better fits than its sub models and other models generated from well-known families.

Conflicts of Interest

Authors have no conflict of interest.

References

Aarset, M. V., (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability,* 36, pp. 106–108.

Abouelmagd, T. H. M., Hamed, M. S. and Ebraheim, A. N., (2017). The Poisson-G family of distributions with applications. *Pakistan Journal of Statistics and Operation Research*, XIII, pp. 313–326.

Abouelmagd, T. H. M., Hamed, M. S., Handique, L., Goual, H., Ali, M. M., Yousof, H. M. and Korkmaz, M. C., (2019). A New Class of Continuous Distributions Based

on the Zero Truncated Poisson distribution with Properties and Applications. *The Journal of Nonlinear Sciences and Applications*, 12, pp. 152–164.

Ahsan, A. L., Handique, L. and Chakraborty, S., (2018). The odd modified exponential generalized family of distributions: its properties and applications. *International Journal of Applied Mathematics and Statistics,* 57, pp. 48–62.

Alizadeh, M., Tahir, M. H., Cordeiro, G.M., Zubai, M. and Hamedani, G. G., (2015). The Kumaraswamy Marshal-Olkin family of distributions. *Journal of the Egyptian Mathematical Society,* 23, pp. 546–557.

Bjerkedal, T., (1960). Acquisition of resistance in Guinea pigs infected with different doses of virulent tubercle bacilli. *American Journal of Hygiene*, 72, pp. 130–148.

Chakraborty, S., Handique, L., (2017). The generalized Marshall-Olkin-Kumaraswamy-G family of distributions. *Journal of Data Science*, 15, pp. 391–422.

Chakraborty, S., Handique, L. and Ali, M. M., (2018). A new family which integrates beta Marshall-Olkin-G and Marshall-Olkin-Kumaraswamy-G families of distributions. *Journal of Probability and Statistical Science,* 16, pp. 81–101.

Chakraborty, S., Handique, L., (2018). Properties and data modelling applications of the Kumaraswamy generalized Marshall-Olkin-G family of distributions. *Journal of Data Science*, 16, pp. 605–620.

Chakraborty, S., Alizadeh, M., Handique, L., Altun, E. and Hamedani, G. G., (2021). A New Extension of Odd Half-Cauchy Family of Distributions: Properties and Applications with Regression Modeling. *Statistics in Transition New Series,* 22, pp. 77–100.

Chakraborty, S., Handique, L. and Jamal, F., (2022). The Kumaraswamy Poisson-G family of distribution: its properties and applications. *Annals of Data Science*, 9, pp. 229–247.

Cordeiro, G. M., De Castro, M., (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, 81, pp. 883–893.

Eugene, N., Lee, C. and Famoye, F., (2002). Beta-normal distribution and its applications. *Communication Statistics Theory and Methods*, 31, pp. 497–512.

Emrah, A., Yousuf, H. M., Chakraborty, S. and Handique, L., (2018). The Zografos-Balakrishnan Burr XII Distribution: Regression Modeling and Applications. *International Journal of Mathematics and Statistics*, 19, pp. 46–70.

Gokarna, R. A., Haitham, M. Y., (2017). The exponentiated generalized-G Poisson family of distributions. *Stochastics and Quality Control,* 32, pp. 7–23.

Gokarna, R. A., Sher, B. C., Hongwei, L. and Alfred, A. A., (2019). On the Beta-G Poisson family. *Annals of Data Science*, 6, pp. 361–389.

Handique, L., Chakraborty, S. and Hamedani, G. G., (2017). The Marshall-Olkin-Kumaraswamy-G family of distributions. *Journal of Statistical Theory and Applications*, 16, pp. 427–447.

Handique, L., Chakraborty, S. and Ali, M. M., (2017). The Beta generated Kumaraswamy-G family of distributions. *Pakistan Journal of Statistics*, 33, pp. 467–490.

Handique, L., Chakraborty, S., (2017a). A new beta generated Kumaraswamy Marshall-Olkin-G family of distributions with applications. *Malaysian Journal of Science*, 36, pp. 157–174.

Handique, L., Chakraborty, S., (2017b). The Beta generalized Marshall-Olkin Kumaraswamy-G family of distributions with applications. *International Journal of Agricultural and Statistical Sciences*, 13, pp. 721–733.

Handique, L., Chakraborty, S. and Thiago, A. N., (2019). The exponentiated generalized Marshall-Olkin family of distributions: Its properties and applications. *Annals of Data Science*, 6, pp. 391–411.

Handique, L., Ahsan, A. L. and Chakraborty, S., (2020). Generalized Modified exponential-G family of distributions: its properties and applications. *International Journal of Mathematics and Statistics*, 21, pp. 1–17.

Handique, L., Chakraborty, S. Eliwa, M. S. and Hamedani, G. G., (2021). Poisson Transmuted-G family of distributions: Its properties and application. *Pakistan Journal of Statistics and Operation research*, 17, pp. 309–332.

Handique, L., Chakraborty, S. and Jamal, F., (2022). Beta Poisson-G family of distribution: Its properties and application with failure time data. *Thailand Statistician,* 20, pp. 308–324.

Handique, L., Aidi, K., Chakraborty, S., Ibrahim, E. and Ali, M. M., (2023). Analysis and Model Validation of Right Censored Survival Data with Complementary Geometric-Topp-Leone-G family of distributions. *International Journal of Statistical Sciences,* 23, pp. 13–26.

Handique, L., Chakraborty, S. Morshedy, M. L., Afify, A. Z. and Eliwa, M. S., (2024). Modelling Veterinary Medical Data Utilizing a new generalized Marshall-Olkin Transmuted Generator of distributions with Statistical Properties. *Thailand Statistician*, 22, pp. 219–236.

Ibahim, E., Handique, L. and Chakraborty, S., (2024). Truncated Cauchy Power Kumaraswamy generalized family of distributions: Theory and Applications. *Stat., Optim. Inf. Comput.*, 12, pp. 364–380.

Jayakumar, K., Mathew, T., (2008). On a generalization to Marshall-Olkin scheme and its application to Burr type XII distribution. *Statistical Papers*, 49, pp. 421–439.

Marshall, A., Olkin, I., (1997). A new method for adding a parameter to a family of distributions with applications to the exponential and Weibull families. *Biometrika*, 84, pp. 641–652.

Moors, J. J. A., (1988). A quantile alternative for kurtosis. *The Statistician*, 37, pp. 25–32.

Song, K. S., (2001). Rényi information log likelihood and an intrinsic distribution measure. *Journal of Planning and Statistical Inference*, 93, pp. 51–69.

Tahir, M. H., Zubai, M., Cordeiro, G. M., Alzaatreh, A. and Mansoor, M., (2016) The Poisson-X family of distributions. *Journal of Statistical Computation and Simulation*, 86, pp. 2901–2921.

Thiago, A. N., Chakraborty, S., Handique, L. and Frank, G. S., (2019). The Extended generalized Gompertz Distribution: Theory and Applications. *Journal of Data science*, 17, pp. 299–330.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 135–154, https://doi.org/10.59139/stattrans-2025-008
Received – 26.10.2023; accepted – 11.09.2024

Impact of human capital on the innovation performance of EU economies

Iwona Skrodzka[1]

Abstract

The purpose of the paper is to empirically determine the impact of human capital on the innovation performance of EU economies. Currently, most researchers consider human capital a significant factor of economic growth based on knowledge and innovation. Depending on the amount and quality of the available resources, human capital can play various parts in an economy, e.g. that of a user of existing knowledge and technology (general human capital), an implementer of new solutions, or a creator of previously undiscovered knowledge (specialised human capital). However, there is a gap in the literature regarding empirical research into the influence of human capital on the innovativeness of economies. This is related to the difficulties associated with the measurement of the two categories, as well as the limited number of methods to study the relationships between unobservable variables. The research described in the paper fills this gap. In order to study the relationship between human capital (general and specialised) and the innovation performance of economies, the partial least squares structural equation modelling (PLS-SEM) was used. The research spanned the years 2014-2020. Four PLS-SEM models were estimated based on cross-sectional data for the EU economies. The results showed that human capital significantly boosts the innovation performance of EU economies. Both general human capital and specific human capital had a significant positive impact on the innovation performance of these countries in the analysed years. The results can have a practical application and serve as an instrument of innovation policies or as a tool helpful in creating conditions for innovation systems.

Key words: human capital, innovativeness, innovation performance, structural equation modeling, PLS-SEM.

1. Introduction

Human capital, understood as the knowledge, skills, competences and other attributes embodied in individuals that are relevant to economic activity (OECD, 1998), has nowadays become a crucial factor behind knowledge- and innovation-based

[1] Faculty of Economics and Finance, University of Bialystok, Białystok, Poland. E-mail: i.skrodzka@uwb.edu.pl, ORCID: https://orcid.org/0000-0002-3261-8687.

growth. The significance of human capital is corroborated by numerous studies (see Azariadis and Drazen, 1990; Mankiw et al., 1992; Benhabib and Spiegel, 1994; Barro, 2001). Many of them emphasize direct relationships between human capital and economic growth. There are, however, reasons to believe that these relationships are more complex than is often assumed (Aleknavičiūtė et al., 2016). Depending on the size and quality of resources, human capital can play various parts in an economy, e.g. that of a user of existing knowledge and technology, an implementer of new solutions, or a creator of previously undiscovered knowledge.

The article analyses the problem of human capital in terms of its impact on the innovativeness of EU countries. Innovativeness is defined as the ability to create and implement innovations. Moreover, two categories to describe innovation are distinguished:

- innovation capacity, i.e. the extent to which an economy is capable of creating and commercialize new ideas,
- innovation performance, i.e. the outcome stemming from a combination of society's creativity and the financial assets of a given economic and institutional environment.

The purpose of the paper is to empirically identify the impact of human capital on the innovation performance of EU economies. Two kinds of human capital are distinguished: general human capital, i.e. overall base of knowledge, skills, competences, and qualifications indispensable in processes associated with diffusion of knowledge and innovation; and specialized human capital, i.e. specialized knowledge, skills, competences and qualifications used for creating new knowledge and developing innovative solutions.

The paper consists of five parts. Section 2 presents selected empirical studies featuring analyses of the relationships between human capital and the innovativeness of European economies. Section 3 describes the research method – partial least squares structural equation modelling. Section 4 discusses the results of modeling. Section 5 sums up the conducted research.

2. Literature review

Empirical verification of the hypothesis that human capital significantly influences the innovativeness of economies presents numerous difficulties. First, the definitions of both of these categories vary in the literature. Second, neither of them is directly observable. Third, there is no universally accepted method to measure them. Fourth, few econometric methods make it possible to examine the influence of one unobservable variable on another. Below presented are examples of empirical research regarding European economies.

R. Aleknavičiūtė, V. Skvarciany and S. Survilaitė (2016) analyzed the impact of human capital on innovation in 26 EU countries. The study covered the years 2002-2012. Ten indicators were used to measure human capital and one indicator to measure innovation.

The studied countries were divided into two clusters: highly innovative economies (Austria, Belgium, Czechia, Cyprus, Dania, Estonia, Finland, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Sweden, United Kingdom) and economies with low innovation levels (Bulgaria, France, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia, Spain). Correlation analysis was the research method used. The following conclusions were reached (Aleknavičiūtė et al., 2016):

1. Among the countries with low innovation, 9 human capital indicators were found to have significant correlations with the level of innovativeness, whereas one – participation of young people in education – was insignificantly correlated. Lifelong learning and high level of computer skills were the most strongly correlated indicators.
2. In the group of highly innovative countries, 6 indicators proved to be significantly correlated with innovation, while 4 (lifelong learning, secondary and higher education, high level of computer skills, and the level of satisfaction with one's education) had insignificant correlations. 'Results achieved by school students in Mathematics' was the indicator which was the most closely correlated with innovation performance.
3. In all the analyzed countries, 8 human capital indicators showed significant correlations with the level of innovation in the economies, with one of them (population with secondary or higher education) being negatively correlated. Two indicators (participation of young people in education and high level of computer skills) were insignificantly correlated. Indicators of the quality of human capital were the most strongly associated with the level of innovation.

One of the advantages of the above-discussed research is the fact that it takes into consideration the qualitative aspect of human capital. As far as its limitations are concerned, innovation is addressed one-dimensionally. Apart from this, analysis of dependencies on the basis of correlation coefficients poses interpretation problems, because it is difficult to unequivocally determine the direction of each dependency.

The influence of human capital and social capital on the innovation activity of economies was investigated by A. Kaas, E. Parts and H. Kaldaru (2012). The statistical sample consisted of 30 European countries. Data on human and social capital were derived from the year 1999, while data on innovation activity from the period of 2002-2004. Innovation was measured with 4 indicators, human capital with 2 indicators, and social capital with 10 indicators.

The countries were divided into 4 groups:

– large, developed Western European economies: Austria, Belgium, France, Germany, Greece, Italy, Netherlands, Portugal, Switzerland, Spain, Sweden, Turkey, United Kingdom,
– small, developed Western European economies: Denmark, Finland, Ireland, Island, Luxembourg, Malta,
– large, catching-up post-communist economies: Bulgaria, Czechia, Hungary, Poland, Romania,
– small, catching-up post-communist economies: Estonia, Latvia, Lithuania, Slovakia, Slovenia.

The values of variables 'human capital' and 'social capital' were estimated by means of the confirmatory factor analysis. The conclusions of the study were as follows (Kaasa et al., 2012):

1. Small, developed Western European economies were found to be the most innovating, followed by large, developed Western European economies. Western economies were relatively far ahead of small, catching-up economies, whereas large, catching-up economies were in the most difficult situation.
2. An analogous pattern applied to the levels of human capital and social capital.
3. Catching-up economies were characterized by less innovation activity and, at the same time, lower levels of human and social capital.

Among the merits of the study is that it accounts for several different indicators of innovation and that it measures human and social capital using the confirmatory factor analysis. What raises doubts, however, is the large disproportion between the numbers of indicators ascribed to the categories under analysis. Besides, the conclusions regarding dependencies were drawn merely on the basis of comparison between the values of latent constructs and the mean values of innovation indicators.

Different statistical and econometric methods were applied by M. Dakhli and D. Clercq (2004) in their research into the impact of human and social capital on the country's level of innovation. The statistical sample comprised 59 countries: 30 from Europe, 13 from Asia and Australia, and 3 from Africa. Data related to human and social capital were from 1995, while data on innovation from 1998.

Innovation was measured with 3 indicators, human capital with 4, while social capital with 31 indicators deriving from surveys. The first stage involved construction of synthetic measures of human capital and social capital, and their dimensions. Next, a correlation analysis was conducted, which revealed that (Dakhli and Clercq, 2004):

1. Human capital was positively correlated with each of the indicators of the level of innovation in an economy.
2. 'Level of overall confidence' and 'trust in institutions' were positively correlated with at least one indicator of innovation.

3. 'Activity in associations' and 'norms of civic behavior' did not have any correlation with the level of innovation in an economy.

In the next step, three regression models were estimated. The innovation indicators were used as dependent variables, while human capital and selected dimensions of social capital were independent variables. Moreover, each country's population size was taken into consideration. In order to ascertain whether social polarization had an impact on the relationship between social capital and innovation, a control variable – 'income gap' – was included into the models. The analysis yielded the following conclusions (Dakhli and Clercq, 2004):

1. Human capital had a positive impact on each of the specified indicators of the level of innovation in an economy.
2. Level of overall confidence and trust in institutions had a positive impact on at least one of the three indicators of innovativeness, i.e. a high level of overall confidence leads to an increase in the number of patents and amount of expenditure on R&D, while trust in institutions had a positive influence on the volume of high-tech exports.
3. 'Activity in associations' had a positive impact on only one indicator of innovativeness, and namely 'R&D expenditure index'.
4. 'Norms of civic behavior' had a negative influence on the level of high-tech exports.
5. Inclusion of 'income gap' as a control variable resulted in higher parameter estimates. Nevertheless, the control variable proved significant only in the model where 'R&D expenditure' was the dependent variable.

Application of various methods of statistical analysis should be regarded as an asset of the study. However, the paper also seems to have several weaknesses. The level of innovation in an economy was approached in a one-dimensional way in each of the regression models. What is more, no full statistical verification of the estimated models was performed. The authors failed to include information as to, e.g. whether the estimated models met the rigorous standards of the least squares method. There is also an evident disproportion between the number of indicators used for measuring the analyzed types of capital.

3. Research method

3.1. Fundamentals of PLS-SEM modelling

Structural equation models (SEM) include a number of statistical methodologies meant to estimate a network of causal relationships, defined according to a theoretical model, linking two or more latent complex concepts, each measured through a number of observable indicators. Among the methods of estimating SEM models, the

covariance-based method (CB), invented by K. G. Jöreskog, enjoyed the greatest popularity for a long time. Its recognition was so universal that in social sciences the phrases: structural equation modeling (SEM) and covariance-based structural equation modeling (CB-SEM) used to be synonymous for many years (Chin, 1998). Meanwhile, H. Wold developed an alternative approach – the partial least square method (PLS).

An SEM model consists of two submodels: a structural one and a measurement one. A structural model describes the relationships among latent variables, whereas a measurement model – the relationships among the latent variables and the indicators by which they are identified (Wold, 1980). Definition of latent variables by means of indicators can be done either deductively or inductively (Rogowski, 1990). Under the former approach, indicators reflect the defined latent variable. In the case of inductive definition, it is assumed that indicators make up the latent variables, hence the expressions formative indicators.

Estimation of a PLS-SEM model is performed using the PLS method. The algorithm simultaneously estimates inner model parameters – path coefficients – and outer model parameters – outer weights and outer loadings. The procedure also yields estimations of the values of all the latent variables included in the model (see Hair et al., 2022). Verification of a PLS-SEM model is a two-stage process. First, the structural model is assessed. Second, if the validity of the structural model has been confirmed, the structural model is tested. Table 1 lists the properties of the model which should undergo evaluation.

Table 1: Evaluation of PLS-SEM model

<table>
<tr><td colspan="6">Evaluation of the measurement models</td></tr>
<tr><td colspan="3">Reflective measurement model</td><td colspan="3">Formative measurement model</td></tr>
<tr><td rowspan="2">Internal consistency</td><td>Cronbach's alpha</td><td>0.60-0.95</td><td rowspan="2">Convergent validity</td><td rowspan="2">Redundancy analysis</td><td rowspan="2">≥ 0.7 correlation</td></tr>
<tr><td>Composite reliability</td><td>0.60-0.95</td></tr>
<tr><td rowspan="2">Convergent validity</td><td>Loadings</td><td>≥ 0.7</td><td rowspan="2">Collinearity between indicators</td><td rowspan="2">Variance Inflation factor (VIF)</td><td rowspan="2">≥ 0.5</td></tr>
<tr><td>Average variance extracted (AVE)</td><td>≥ 0.5</td></tr>
<tr><td rowspan="3">Discriminant validity</td><td>Cross-loadings</td><td>-</td><td rowspan="3">Significance of outer weights</td><td rowspan="3">p-value</td><td rowspan="3">< 0.05</td></tr>
<tr><td>Fornell-Larcker criterion</td><td>-</td></tr>
<tr><td>Heterotrait-monotrait ratio (HTMT)</td><td>< 0.9</td></tr>
</table>

Table 1: Evaluation of PLS-SEM model (cont.)

Evaluation of the structural model		
Collinearity	Variance Inflation factor (VIF)	≥ 0.5
Predictive power	Coefficients of determinations (R^2)	values of 0.75, 0.50 and 0.25 are considered substantial, moderate and weak
Predictive relevance	Stone-Geisser's Q^2 value	≥ 0
Significance of path coefficients	p-value	< 0.05

Source: own work on the basis of (Hair et al., 2017, p. 106).

3.2. PLS-SEM models with higher order latent variables

Introducing a higher-order latent variable to an SEM model has numerous advantages associated, among other things, with the theoretical usefulness of the model, the level of abstraction, or the integrity and accuracy of the measurement model. Nevertheless, using higher-order latent variables also involves several challenges, e.g. the decision to choose the type of higher-order latent variable measurement model, selection of estimation method, or the more complex process of statistical verification of the model (Wetzels et al., 2009).

The literature offers a variety of approaches to identification and estimation of models with higher-order latent variables. The most frequently cited is the approach proposed by Wold, now known as the repeated indicators approach. In this approach, higher-order latent variables are defined by means of the indicators of all the lower-order latent variables which define them (Sarstedt et al., 2019).

Statistical verification of a PLS-SEM model with higher-order latent variables is relatively complicated. Admittedly, the evaluation criteria used are analogous to those applied in standard PLS-SEM models, but particular attention must be paid to distinguishing the relationships which are part of the measurement model from those which belong to the structural model. The measurement model of a higher-order latent variable is a complex one, which should be taken into consideration at the evaluation stage. It consists of a measurement model of lower-order latent variables and a measurement model of higher-order latent variables (as a whole), represented by the relationships among the higher-order variable and the lower-order variables (Hair et al., 2022).

The PLS-SEM method is not without its limitations. Some researchers note that the non-parametric nature of this modelling technique is a serious flaw. Also, collection of samples of insufficient size and application of PLS-SEM instead of CB-SEM is subject to criticism in the case of studies based on sample sets. Another disadvantage of PLS-

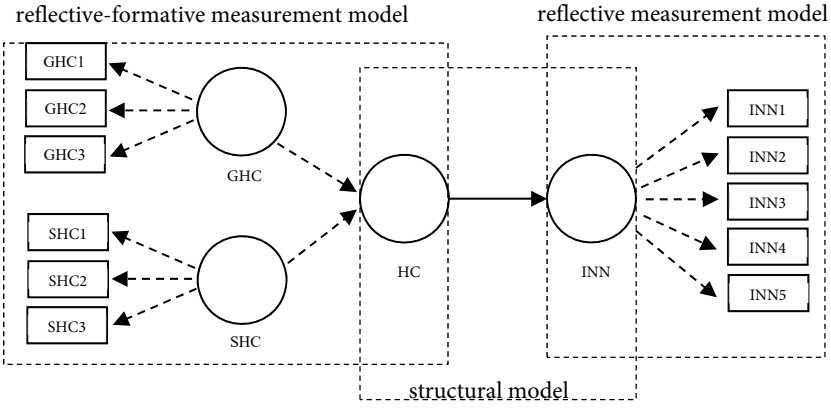
SEM models is that they are linear, whereas the relationships between many economic variables are of non-linear nature.

3.3. PLS-SEM model specification

In line with the stated research objective, the following main hypothesis was adopted: Human capital has a positive influence on the innovation performance of EU economies. Apart from this, two specific hypotheses were verified:

1. General human capital has a positive impact on the innovation performance of EU economies.
2. Specialized human capital has a positive impact on the innovation performance of EU economies.

The PLS-SEM model (a diagram of which is shown in Figure 1) was used to verify the above hypotheses. Latent variable HC was defined by means of two unobservable indicators comprising: general human capital (GHC) and specialized human capital (SHC). The model contained, therefore, a second-order latent variable (HC). In the next step, latent variables GHC and SHC were defined by means of reflective indicators. A deductive approach and reflective indicators were also applied to define latent variable INN. The indicators which defined the latent variables are presented in Table 1.



HC – 2nd order latent variable,
GHC, SHC, INN – 1st order latent variables,
GHC*i*, SHC*i*, INN*j* – indicators, $i = 1, 2, 3$, $j = 1,\ldots, 5$.

Figure 1: Specification of PLS-SEM model.

Source: own work.

Table 2: Indicators of latent variables

Latent variable	Indicator	Description	Source
GHC	GHC1	Population aged 25-64 having completed tertiary education (%).	Eurostat
	GHC2	Employees aged 20-64 having completed tertiary education (%).	Eurostat
	GHC3	Population aged 25-64 participating in education and training (%).	Eurostat
SHC	SHC1	Population aged 25-64 employed in science and technology (%).	Eurostat
	SHC2	Researchers (% of total employment).	Eurostat
	SHC3	Employment in technology and knowledge-intensive sectors (% of total employment).	Eurostat
INN	INN1	SMEs introducing product innovations (%).	EIS
	INN2	SMEs introducing business process innovations (%).	EIS
	INN3	PCT patent applications per billion GDP (PPS).	EIS
	INN4	Scientific publications among the top-10% most cited publications worldwide (% of total scientific publications of the country).	EIS
	INN5	Knowledge-intensive services exports (% of total services exports).	EIS

Source: own work.

The database, constructed with the use of data from the Eurostat, the World Bank and the European Innovation Scoreboard (EIS), consisted of 42 indicators. Seventeen of them regarded the innovation performance of economies, while 25 – human capital. As a result of statistical verification, at various stages of the modelling process, the indicators were removed from the base, e.g. due to gaps in data, insufficient variation, of negative verification of the measurement model. Eventually, 11 indicators were selected for estimation (Table 1). The model was estimated using the SmartPLS software, on the basis of cross-sectional data for four years: 2014, 2016, 2018, and 2020.

4. Results and discussion

The results of the estimation of the models are depicted in Figures 2–5. The estimated models underwent multi-stage statistical verification. First, the properties of the measurement models of the first-order latent variables (GHC, SHC, INN) were tested. Tables 3–6 present the results of these analyses. The indicators fulfilled the criteria of convergent validity, internal consistency reliability, and discriminant validity, and thus were approved.

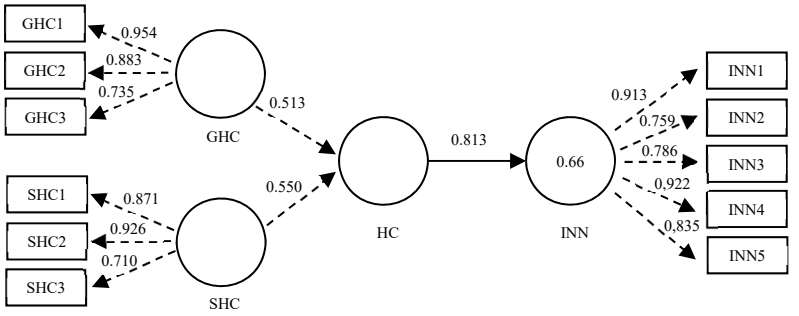


Figure 2: PLS-SEM$_{2014}$ results of estimation

Source: own work.

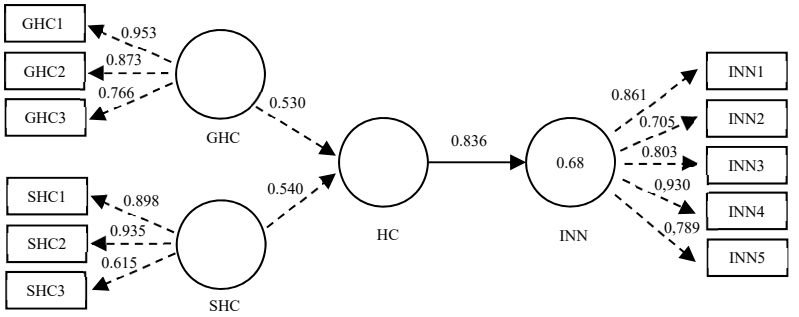


Figure 3: PLS-SEM$_{2016}$ results of estimation

Source: own work.

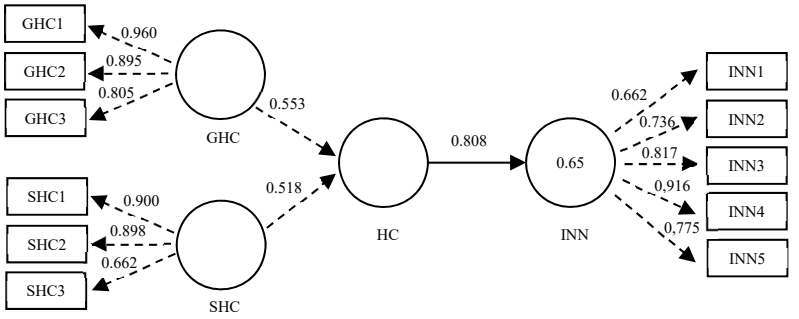


Figure 4: PLS-SEM$_{2018}$ results of estimation

Source: own work.

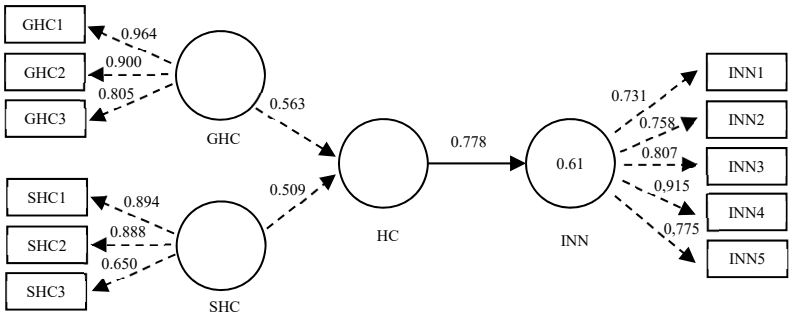


Figure 5: PLS-SEM$_{2020}$ results of estimation

Source: own work.

Table 3: Assessment of reflective measurement model in PLS-SEM$_{2014}$

Latent variable	Indicator	Convergent validity		Internal consistency reliability		Discriminant validity
		Loading	AVE	Composite reliability	Cronbach's alpha	Cross loadings criteria
		>0.7	>0.5	0.6-0.95	0.6-0.95	
GHC	GHC1	0.954	0.74	0.82	0.82	Yes
	GHC2	0.883				
	GHC3	0.735				
SHC	SHC1	0.871	0.71	0.83	0.79	Yes
	SHC2	0.926				
	SHC3	0.710				
INN	INN1	0.913	0.72	0.91	0.90	Yes
	INN2	0.759				
	INN3	0.786				
	INN4	0.922				
	INN5	0.835				

Source: own work.

Table 4: Assessment of reflective measurement model in PLS-SEM$_{2016}$

Latent variable	Indicator	Convergent validity		Internal consistency reliability		Discriminant validity
		Loading	AVE	Composite reliability	Cronbach's alpha	Cross loadings criteria
		>0.7	>0.5	0.6-0.95	0.6-0.95	
GHC	GHC1	0.953	0.75	0.84	0.83	Yes
	GHC2	0.873				
	GHC3	0.766				
SHC	SHC1	0.898	0.69	0.84	0.76	Yes
	SHC2	0.935				
	SHC3	0.615				
INN	INN1	0.861	0.67	0.90	0.88	Yes
	INN2	0.705				
	INN3	0.803				
	INN4	0.930				
	INN5	0.789				

Source: own work.

Table 5: Assessment of reflective measurement model in PLS-SEM$_{2018}$

Latent variable	Indicator	Convergent validity		Internal consistency reliability		Discriminant validity
		Loading	AVE	Composite reliability	Cronbach's alpha	Cross loadings criteria
		>0.7	>0.5	0.6-0.95	0.6-0.95	
GHC	GHC1	0.960	0.79	0.87	0.87	Yes
	GHC2	0.895				
	GHC3	0.805				
SHC	SHC1	0.900	0.69	0.81	0.76	Yes
	SHC2	0.898				
	SHC3	0.662				
INN	INN1	0.662	0.62	0.88	0.85	Yes
	INN2	0.736				
	INN3	0.817				
	INN4	0.916				
	INN5	0.775				

Source: own work.

Table 6: Assessment of reflective measurement model in PLS-SEM$_{2020}$

Latent variable	Indicator	Convergent validity		Internal consistency reliability		Discriminant validity
		Loading	AVE	Composite reliability	Cronbach's alpha	Cross loadings criteria
		>0.7	>0.5	0.6-0.95	0.6-0.95	
GHC	GHC1	0.964	0.80	0.87	0.87	Yes
	GHC2	0.900				
	GHC3	0.805				
SHC	SHC1	0.894	0.67	0.81	0.75	Yes
	SHC2	0.888				
	SHC3	0.650				
INN	INN1	0.731	0.64	0.89	0.86	Yes
	INN2	0.758				
	INN3	0.807				
	INN4	0.915				
	INN5	0.775				

Source: own work.

Next, the second part of the measurement models of the second-order latent variable (HC) was verified. The unobservable indicators of HC were not colinear, whereas the estimates of weights proved to be statistically significant (Table 7). Therefore, the models were approved.

Table 7: Significance testing results of the formative model weights

Relation	Weight	*t* value	*p* value	95% confidence interval	Significance (p<0.05)?
PLS-SEM$_{2014}$					
GHC→HC	0.513	13.79	0.000	(0.43, 0.58)	Yes
SHC→HC	0.550	13.14	0.000	(0.48, 0.65)	Yes
PLS-SEM$_{2016}$					
GHC→HC	0.530	14.01	0.000	(0.44, 0.59)	Yes
SHC→HC	0.540	11.76	0.000	(0.47, 0.65)	Yes
PLS-SEM$_{2018}$					
GHC→HC	0.553	13.29	0.000	(0.48, 0.64)	Yes
SHC→HC	0.518	11.65	0.000	(0.44, 0.62)	Yes
PLS-SEM$_{2020}$					
GHC→HC	0.563	12.75	0.000	(0.48, 0.65)	Yes
SHC→HC	0.509	11.00	0.000	(0.43, 0.61)	Yes

Source: own work.

In the last step, statistical verification of the structural models was conducted. In every case, variable HC showed a statistically significant effect on variable INN (Table 8). The statistical hypothesis that HC did not have significant effect on INN was, therefore, rejected in favor of the alternative hypothesis.

Table 8: Significance testing results of the structural model path coefficients

Model	Path coefficient	*t* value	*p* value	95% confidence interval	Significance (p<0.05)?
PLS-SEM$_{2014}$	0.813	16.86	0.000	(0.72, 0.91)	Yes
PLS-SEM$_{2016}$	0.825	17.45	0.000	(0.74, 0.92)	Yes
PLS-SEM$_{2018}$	0.808	14.50	0.000	(0.70, 0.92)	Yes
PLS-SEM$_{2020}$	0.778	11.85	0.000	(0.65, 0.91)	Yes

Source: own work.

The coefficients of determination had values ranging from 0.61–0.68 (Figures 2–5), which means that the variability of INN was explained by the models to a satisfactory degree. The Q^2 values of the Stone-Geisser test were positive (Table 9), and thus the models proved to have high prognostic accuracy. The structural models were positively assessed. The next stage of the modelling process involved analysis of the obtained results.

Table 9: Q^2 values

Indicators	Q^2			
	PLS-SEM$_{2014}$	PLS-SEM$_{2016}$	PLS-SEM$_{2018}$	PLS-SEM$_{2020}$
INN1	0.37	0.24	0.11	0.14
INN2	0.22	0.13	0.11	0.13
INN3	0.48	0.54	0.54	0.49
INN4	0.48	0.57	0.54	0.51
INN5	0.53	0.46	0.40	0.34
General	0.63	0.64	0.62	0.57

Source: own work.

The estimates of the parameters of structural models demonstrated that general human capital had a strong, positive influence on the innovation performance of EU economies in each of the four analyzed years. The path coefficients assumed values within the range 0.778–0.836. Moreover, both general human capital and specialized human capital had a positive impact on the innovation performance of the economies under study. This is evidenced by the parameters of substitution relationships, which can be derived by substituting latent variable HC with the relationships of its

measurement model (Table 10). The strength of the influence exerted by both kinds of capital on innovation performance was comparable, although it should be noted that in the years 2014 and 2016, specialized human capital had a slightly stronger impact, while in 2018 and 2020 the influence of general human capital was more pronounced.

Table 10: Significance testing results of the substitution relation parameters

Relation	Parameter	*t* value	*p* value	95% confidence interval	Significance (p<0.05)?
PLS-SEM$_{2014}$					
GHC→INN	0.417	9.66	0.000	(0.33, 0.48)	Yes
SHC→INN	0.447	13.33	0.000	(0.39, 0.52)	Yes
PLS-SEM$_{2016}$					
GHC→INN	0.437	9.42	0.000	(0.34, 0.52)	Yes
SHC→INN	0.445	12.95	0.000	(0.39, 0.53)	Yes
PLS-SEM$_{2018}$					
GHC→INN	0.447	9.06	0.000	(0.35, 0.54)	Yes
SHC→INN	0.418	11.77	0.000	(0.36, 0.50)	Yes
PLS-SEM$_{2020}$					
GHC→INN	0.438	8.80	0.000	(0.35, 0.54)	Yes
SHC→INN	0.396	9.11	0.000	(0.31, 0.49)	Yes

Source: own work.

PLS-SEM modelling also yielded estimates of the values of the latent variables included in the model. They were treated as values of synthetic measures and used for ranking and classification of the studied countries. Four typological groups were created: Group I – very high level of analyzed category; Group II – high/medium level; Group III – low level; and Group IV – very low level. Interval boundaries were calculated using the mean and standard deviation of the synthetic measures.

The classification of EU countries according to the level of human capital in 2014 was as follows (the order of countries within groups corresponds to the ranking status):
- Group I: Finland, Denmark, Sweden, Luxembourg, Ireland,
- Group II: Netherlands, Belgium, France, Austria, Estonia, Slovenia,
- Group III: Germany, Spain, Lithuania, Cyprus, Czechia, Latvia, Malta, Portugal, Hungary, Poland, Greece, Bulgaria,
- Group IV: Slovakia, Italy, Croatia, Romania.

The division of the studied countries into typological groups in terms of their innovation performance in 2014 is presented below:
- Group I: Finland, Sweden, Netherlands, Ireland, Germany, Belgium, Denmark,
- Group II: Luxembourg, Austria, France, Cyprus, Portugal, Italy,

– Group III: Greece, Slovenia, Czechia, Spain, Estonia, Malta, Hungary, Lithuania, Croatia,
– Group IV: Slovakia, Latvia, Bulgaria, Poland, Romania

In 2020 several changes occurred in both classifications, as compared to 2014. In the human capital clustering, Lithuania rose from group III to group II, whereas Slovakia moved up from group IV to group II. Bulgaria, meanwhile, dropped from group III to group IV. In the innovation performance clustering, Ireland fell from group I to group II, Portugal – from group II to group III, whereas Greece advanced from group III to group II.

The present empirical study confirmed that human capital is an important factor behind enhancing the innovation performance of EU economies. Similar conclusions can be drawn from theoretical and empirical research by other authors. In particular, selected endogenous models emphasize the indirect effect of human capital on increased productivity due to improvement of capacity for creating domestic innovations and absorption of new technologies (see Nelson and Phelps; 1966, Romer, 1990; Grossman and Helpman, 1991; Aghion and Howitt, 1992; Jones, 2003). Empirical investigations performed for various groups of countries indicate that human capital exerts a positive influence on the level of innovation in economies and increases their capacity to transfer knowledge and technology (see Benhabib and Spiegel, 2005; Vandenbussche et al., 2006, Ang et al., 2011; Danquah and Ouattara, 2014, Balcerzak and Pietrzak, 2016).

5. Conclusions

Empirical research on the relationship between human capital and innovativeness of economies is a very complex issue. This is related to the difficulties associated with measurement of the two categories, as well as the limited number of methods to study the relationships between unobservable variables. Nevertheless, various authors have attempted to identify the strength and direction of the impact of human capital on different aspects of innovativeness. This paper also makes such an attempt.

The research focused on EU economies during the years 2014–2020. PLS-SEM models were developed and estimated, containing the variables: human capital, general human capital, specialized human capital, and the innovation performance of the economy. The results of the modeling revealed a positive impact of human capital on the innovation performance of the analyzed economies. This indicates that economies with higher levels of human capital are also more innovative. Moreover, the model showed that the impact of general human capital and specialized human capital on innovation performance was comparable.

Based on the obtained results, the following conclusions can be drawn. The diversification of human capital is crucial for the innovativeness of an economy. General human capital provides flexibility and broad adaptability to new technologies and market changes, while specialized human capital enables the creation of advanced technological innovations. Optimal conditions for innovation arise when both types of human capital are well-developed and complement each other. Although both types of human capital have their specific functions, their combination is crucial for maximizing the innovativeness of an economy. General human capital creates the foundation on which specialized human capital can develop, meaning that countries must invest in both forms simultaneously.

The results of the conducted study can have a practical application and serve as an instrument of innovation policies or as a tool helpful in creating conditions for innovation systems.

Future research can be improved by considering other factors of innovation, e.g. financial factors. Then, it would be possible to verify which type of factors, tangible or intangible, have a stronger impact on innovation. Models accounting for relationships between various aspects of an economy's innovation capacity and its innovation performance provide an interesting direction for future research.

References

Aghion, P., Howitt, P., (1992). A model of growth through creative destruction. *Econometrica*, 60(2), pp. 323–351. doi: 10.2307/2951599.

Aleknavičiūtė, R., Skvarciany, V. and Survilaitė, S., (2016). The role of human capital for national innovation capability in EU countries. *Economics and Culture*, 13(1), pp. 114–125. doi: 10.1515/jec-2016-0014.

Ang, J. B., Madsen, J. B. and Rabiul Islam, Md., (2011). The effects of human capital composition on technological convergence. *Journal of Macroeconomics*, 33(3), pp. 465–476. doi: 10.1016/j.jmacro.2011.03.001.

Azariadis, C., Drazen, A., (1990). Threshold externalities in economic development. *The Quarterly Journal of Economics*, 105(2), pp. 501–526. doi: 10.2307/2937797.

Balcerzak, A. P., Pietrzak, M. B., (2016). Structural equation modeling in evaluation of technological potential of European Union Countries in the years 2008–2012. In M. Papież and S. Śmiech (Eds.). The 10th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings (pp. 9–18). Cracow: *Foundation of the Cracow University of Economics*.

Barro, Robert, J., (2001). Human capital and growth. *The American Economic Review*, 91(2), pp. 12–17, doi: 10.1257/aer.91.2.12.

Benhabib, J., Spiegel, M. M., (1994). The role of human capital in economic development evidence from aggregate cross-country data. *Journal of Monetary Economics*, 34(2), pp. 143–173. doi: 10.1016/0304-3932(94)90047-7.

Benhabib, J., Spiegel, M. M., (2005). Human capital and technology diffusion. In P. Aghion and S. Durlauf (Eds.). Handbook of economic growth (pp. 935–966). *Elsevier*. doi: 10.1016/S1574-0684(05)01013-0.

Chin, W. W., (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.). Modern methods for business research (pp. 295–336). Mahwah: *Lawrence Erlbaum*.

Dakhli, M., Clercq, D. D., (2004). Human capital, social capital, and innovation: a multi-country study. *Entrepreneurship & Regional Development*, 16(2), pp. 107–128. doi: 10.1080/0898562041000167783 5.

Danquah, M., Ouattara, B., (2014). Productivity growth, human capital and distance to frontier in Sub-Saharan Africa. *Journal of Economic Development*. pp. 39(4), pp. 27–48.

Grossman, G., Helpman, E., (1991). Trade, knowledge spillovers, and growth. *European Economic Review*, 35(2-3), pp. 517–526. doi: 10.1016/0014-2921(91)90153-A.

Hair, J. F., Hult, G. T. M., Ringle, C. M. and Sarstedt, M., (2022). A primer on partial least squares structural equation modeling (PLS-SEM). Thousand Oaks, CA: *Sage*.

Hair, J. F., Sarstedt, M., Ringle, C. M. and Gudergan, S. P., (2018). Advanced Issues in Partial Least Squares Structural Equation Modeling. Thousand Oaks, CA: *Sage*.

Kaasa, A., Parts, E. and Kaldaru, H., (2012). The role of human and social capital for innovation in catching-up economies. In E. G. Carayannis, U. Varblane and T. Roolaht (Eds.). Innovation systems in small catching-up economies (pp. 259–276). Berlin, Heidelberg: *Springer*. doi: 10.1007/978-1-4614-1548-0_14.

Jones, C. I., (2003). Human capital, ideas, and economic growth. In L. Paganetto and E. S. Phelps (Eds.). Finance, research, education and growth (pp. 51–74). London: *Palgrave Macmillan*. doi: 10.1057/9781403920232_4.

Mankiw, N.G., Weil, D. and Romer, D., (1992). A contribution to the empirics of economic growth. The Quarterly Journal of Economics, 107, pp. 407–437. doi: 10.2307/2118477.

Nelson, R. R., Phelps, E. S., (1966). Investment in humans, technological diffusion, and economic growth. *The American Economic Review*, 56(1/2), pp. 69–75.

OECD, (1998). Human capital investment: an international comparison. Paris: *OECD* Publishing. doi: 10.1787/9789264162891-en.

Rogowski, J., (1990). Soft models. Theory and application in economic research. Białystok: *Wydawnictwo Filii UW w Białymstoku*.

Romer, P. M., (1990). Endogenous technological change. *Journal of Political Economics*, 98(5), pp. S71–S102.

Sarstedt, M., Hair, J. F., Cheah, J.-H., Becker, J.-M. and Ringle, C. M., (2019). How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australasian Marketing Journal*, 27(3), pp. 197–211. doi: 10.1016/j.ausmj.2019.05.003.

Skrodzka, I., (2021). Human capital and the innovative performance of Central and Eastern European countries - PLS-SEM modelling. Warszawa: *CeDeWu*.

Vandenbussche, J., Aghion, P. and Meghir, C., (2006). Growth, distance to frontier and composition of human capital. *Journal of Economic Growth*, 11(2), pp. 97–127. doi: 10.1007/s10887-006-9002-y.

Weresa, M. A., (2014). Innovation policy. Warszawa: *PWN*.

Wetzels, M., Odekerken-Schröder, G. and van Oppen, C., (2009). Using PLS path modeling for assessing hierarchical construct models: guidelines and empirical illustration. *MIS Quarterly*, 33(1), pp. 177–195. doi: 10.2307/20650284.

Wold, H., (1980). Soft modelling: Intermediate between traditional model building and data analysis. *Banach Center Publications*, 6(1), pp. 333–346.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 155–181, https://doi.org/10.59139/stattrans-2025-009
Received – 13.02.2024; accepted – 04.12.2024

Improving detectability of the indicator saturation approach through winsorization: an empirical study in the cryptocurrency market

Suleiman Dahir Mohamed[1], **Mohd Tahir Ismail**[2], **Majid Khan Bin Majahar Ali**[3]

Abstract

Despite the introduction of several adjustments, mitigating data anomalies in financial datasets has proven challenging, particularly in the context of cryptocurrencies with extreme values and increased volatility. The progress in properly addressing these anomalies prior to testing remains restricted, highlighting the unique and complex nature of financial data in this domain. Thus, in this paper we propose a hybrid approach called the Win-IS strategy. It is meant to address the influence of extreme outliers in the tail and subsequently identify breaks, trend breaks and outliers in cryptocurrencies. This methodology uses the winsorization (Win) process to enhance the effectiveness of the indicator saturation (IS) approach. The study uses cryptocurrencies like Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), Tether (USDT), and Ripple (XRP). The results of the research indicate that the winsorization strategy improved the detectability of the IS approach, with Win-IS outperforming the IS method in terms of the Bayesian Information Criterion. Furthermore, the Win-IS technique uncovered additional breaks, trend breaks and outliers that were previously unknown and repeated in some cases as detected by the IS strategy. The effect of winsorization is dependent on the chosen percentile and dataset attributes. Through detailed examination and comparison, the findings of this research contribute to the improvement of other detection approaches, providing a valuable perspective for researchers and practitioners in the field. Additionally, this hybrid approach can improve decision-making, risk management and model creation, benefiting investors, legislators and scholars.

Key words: breaks, outliers, winsorization, indicator saturation, cryptocurrency.
JEL Classification: C22, C58, C61, G23, G32.

[1] School of Mathematical Sciences, Universiti Sains Malaysia, Minden, Pulau Pinang, 11800, Malaysia. E-mail: suleimandm2017@gmail.com. ORCID: http://orcid.org/0000-0002-3368-204X.

[2] Corresponding Author. School of Mathematical Sciences, Universiti Sains Malaysia, Minden, Pulau Pinang, 11800, Malaysia. E-mail: m.tahir@usm.my. ORCID: http://orcid.org/0000-0003-2747-054X.

[3] School of Mathematical Sciences, Universiti Sains Malaysia, Minden, Pulau Pinang, 11800, Malaysia. E-mail: majidkhanmajaharali@usm.my. ORCID: http://orcid.org/0000-0002-5558-5929.

1. Introduction

A structural break is an abrupt change in a time series of data. The structural break and outliers are an important aspect to consider in economics and statistics since they are unexpected. A structural break denotes a change in the behavior of a variable over time, such as a rise in the money stock, or a shift in a previously observed link between observable variables, such as inflation and unemployment, or the balance of trade and the exchange rate (Brooks, 2019). Outliers are data points that deviate from the norm (Hawkins, 1980). Extreme values can have a significant impact on the performance of statistical tests. As a result, correctly identifying changepoints in time series data becomes challenging when working with big samples including a large number of extreme values (outliers), which can either coincide with or disguise major shifts (breaks). Aside from the masking effect, recognizing and correctly identifying breaks and outliers concurrently is another significant challenge. According to Mulry et al. (2014), correctly detecting breaks and outliers is crucial for making educated investment decisions, managing risk, and maintaining the accuracy of financial analysis. Particularly when dealing with financial data, such as cryptocurrency, which is known to undergo major changes as a result of external factors such as wars, natural disasters, etc. (Chatzikonstanti, 2017). Satoshi Nakamoto, the alias for an anonymous computer programmer or group of programmers, invented the first cryptocurrency on January 3, 2009, when the Bitcoin software was made public. Later, numerous additional coins appeared. Cryptocurrencies have emerged as a disruptive digital asset class that has the potential to disrupt traditional financial systems. As these digital assets gain popularity, the need for reliable and effective solutions to monitor and manage data fluctuations grows.

So far, a variety of statistical methodologies have been used to identify breaks and anomalies within cryptocurrency datasets, including single change detection approaches such as Chow (1960) and Quandt (1960), two change detection approaches such as Papell and Prodan (2003), and multiple change detection approaches such as BP of Bai and Perron (1998, 2003), as well as the Iterative Cumulative Sum of Squares (ICSS) approach developed by Iclan and Tiao (1994). In addition, researchers used these strategies separately for break or outlier detection. Chatzikonstanti (2017) used the wavelet approach to handle outliers and the CUSUM approach to find breaks. Mandaci and Cagli (2022) used Bai and Perron to determine the frequency of cryptocurrency breaks. Yen et al. (2022) used the BP approach to analyze ten cryptocurrencies and found structural breaks in return, price, and squared return. Sahoo (2021) employed the Narayan and Popp (2010) endogenous two structural breakdowns unit root test to identify breaks in bitcoin returns. Canh et al. (2019) applied the Wald test and discovered structural fractures in all well-known

cryptocurrencies. Thies and Molnár (2018) used Bayesian change point (BCP) analysis to examine the presence of many segments in the Bitcoin return distribution and demonstrated evidence of structural breaks in the first and second moments of the return distribution. Dutta and Bouri (2022) used Ane et al. (2008) technique and found no indication that any of the top cryptocurrencies have outliers except the Bitcoin return series. Kaseke et al. (2022) used the Pruned Exact Linear Time (PELT) method to identify breakpoints in the cryptocurrency market, which tests for changepoints in the mean, variance, and both mean/variance of the series. Abakah et al. (2020) used Bai and Perron approach and its extension to the fractional case and discovered the existence of breaks in the cryptocurrency market. Aharon (2023) found breaks in cryptocurrency by employing modified ICSS technique. Jiang and Yoon (2023) used BP and ICSS to detect cryptocurrency breaks.

However, many of these traditional techniques may struggle in situations when outliers exist (Fearnhead and Rigaill, 2019). Rodrigues and Rubia (2011) demonstrate that outliers can conceal the presence of structural breaks. Thus, the challenge is to determine which magnitude can be classified as a break or an outlier. The topic of distinguishing between changepoints and outliers has gotten very little consideration. As a result, when evaluating data for structural changes, outliers can frequently obscure major trends, yielding incomplete or misleading conclusions. Traditional approaches may not properly discriminate between genuine changes and those disguised by extreme values.

To solve this problem, Hendry (1999) suggested the indicator saturation strategy known as the IS approach. So far, this technology has been able to detect various data patterns such as breaks and outliers in the data simultaneously. However, because the IS technique can detect many data patterns at the same time, a masking effect may occur, as previously discussed. Specifically, when the IS technique is used in very high frequency datasets such as rapidly fluctuating markets like cryptocurrencies. Therefore, this study uses the IS technique to first identify and record the dates of the breaks, trend breaks, and outliers simultaneously in five distinct cryptocurrency log returns. Second, by ignoring the outcome of the first objective, lessens the impact of extreme observations in each data set using Winsorization technique. Thirdly, employs the IS approach to simultaneously re-identify and record the dates of outliers, trend breaks, and breaks in the Winsorized log returns of each cryptocurrency. Fourth, detects presence of masking effect by comparing and categorizing results into repeating and emerging changes.

The study is driven by the fact that outliers can influence data analysis, masking actual structural changes that are crucial for accurate interpretation. By tackling outliers first, we can uncover hidden breaks, resulting in a better comprehension of the data and more informed decision-making. When applying the IS technique to high-frequency

data, a caution is to be exercised since outliers and rapid changes in the data may generate a masking effect, and high-frequency data with rapid ups and downs may confound the IS approach's automatic detection system. This paper presents a methodological framework for integrating the winsorization strategy into the indicator saturation approach. The hybrid technique might be referred to as the Win-IS approach. The hybrid technique identifies Hidden Patterns that are disguised by outliers, improving the accuracy of break identification. This method may improve risk management and assist investors, traders, and financial analysts in making sound judgments in the ever-changing world of digital assets and other financial markets. The paper is organized as follows. Section 2 contains the body of the current literature; Section 3 describes the approach used; Section 4 shows the results and the discussion; and Section 5 offers the conclusion.

2. Literature review

Literature documented various approaches for detecting data breaks and outliers. Most known break detection methods are based on regression. Chow (1960) pioneered structural break testing for regression models, developing the F-test for a single break, assuming that the break date is previously known under the null of no break. Quandt (1960) altered the Chow framework to consider the F-statistic with the highest value among all potential break dates to loosen the requirement that the candidate break date be known. Later research revealed the assumption of prior knowledge of the break dates, which expanded on previous experiments to allow for multiple breaks, particularly the Bai and Perron tests (Bai and Perron, 1998, 2003). The Bai and Perron approach is limited to trimming 15% of the data and a maximum of 5 breaks. Ohara (1999) also used a method based on Zivot and Andrews (2002) sequential t-tests to analyze the case with m breaks with ambiguous break dates. Papell and Prodan (2003) developed a test based on restricted structural change that explicitly enables two offsetting structural modifications. Detection of breaks in the case of variance has also been investigated using Iclan and Tiao (1994) iterative cumulative sum of squares (ICSS) approach.

Two popular ways to deal with outliers are trimming and winsorization (Moir, 1988). Winsorizing, also known as Winsorization, is a statistical transformation technique that limits extreme values in statistical data to lessen the impact of potentially inaccurate outliers. It is named after the engineer-turned-biostatistician Charles P. Winsor (1895–1951). According to Tukey (1962) when Winsor discovered an outlier in a sample, he did not just discard it; instead, he altered its value. Winsor proposed utilizing the size of the next greatest (or smallest) observation to estimate the magnitude of an extreme, poorly known, or unknown observation. According to Xao et al. (2014)

Winsorization is another reliable method for handling non-normal distributions to prevent information loss and preserve the original sample size. In another classification, the most common outlier treatments in finance are winsorizing, trimming, and dropping (Adams et al., 2019). To lessen the influence of the outlying points, robust solutions based on Winsorization are frequently used (Cheng & Young, 2023). Winsorization method provides adjustments for the observed influential value and winsorization processes can be one-sided or two-sided (Mulry et al., 2014). However, determining the cutoff sites is a vital part of these approaches (Cheng and Young 2023). The more the data is winsorized, the bigger the bias in the coefficient estimates for variables (Lien et al., 2005). Adams et al. (2019) extended the winsorization approach into multivariate case. Hamadani and Ganai (2023) cleaned and processed their data using the winsorization approach. Li et al. (2021) combined the change detection method and the Winsorization method into the prediction model based on the autoregressive moving average model.

On the other hand, Hendry (1999) suggested a strategy called indicator saturation (IS). The indicator saturation methodology employs an automatic multi-path search strategy that can handle more candidate variables N than observations T, separates variables into blocks, and records significant ones using impulse-indicator saturation (Castle et al., 2011). According to Pretis et al. (2017), indicator saturation offers an alternate technique based on an expanded general-to-specific methodology based on model selection. Starting with a full set of indicators and discarding all except the most significant ones, structural breaks can be identified without specifying a minimum break length, maximum break number, or imposed co-breaking. Hendry et al. (2008) demonstrated that different numbers of splits and uneven splits have no effect on the retention rate. According to Castle (2022), numerous indicator saturation estimators (ISEs) are available to model a wide range of non-stationarity phenomena. However, each ISE is created to address a particular issue. For example, the step indicator saturation (SIS) of Castle et al. (2015) for location shifts, the trend indicator saturation (TIS) of Pretis et al. (2015) for trend breaks, and the impulse indicator saturation (IIS) of Hendry et al. (2008) and Johansen and Nielsen (2016) for outliers. Pretis et al. (2018) developed an algorithm for the indicator saturation approach in general to specific modelling (Gets), which provides a straightforward computation of this approach. This method has been applied in a variety of fields in the literature. Marczak and Proietti (2016) used IIS and SIS in the framework of structural time series models. Ghouse (2021) employed IIS to discover structural breaks in Pakistan Islamic banks data. Pretis et al. (2015) used TIS and SIS to assess climate models. Castle et al. (2021) utilized TIS and SIS in identifying shifts in trends within a long-term UK production function. Ghouse et al. (2022) utilized the IIS technique to identify the structural breaks in the returns and volatility of commercial banks in Pakistan. Muhammadullah (2022)

applied IIS to detect outliers in the cross-sectional analysis estimated through the application of regularization techniques with COVID-19 data. Panday (2015) utilized IIS to investigate the influence of monetary policy on the exchange rate of Nepal. Castle et al. (2012) employed U.S. real interest rates to assess the effectiveness of IIS in comparison to the BP approach and found that IIS successfully reproduces the results of the BP. Ismail and Nasir (2018) conducted a comparison between IIS and BP in ASEAN sharia-compliant indices and found almost identical results. However, IIS can identify significant breaks and outliers that occur at the start and end of a sample, while BP necessitates a designated percentage of the sample for analysis. Mohamed et al. (2023) conducted an empirical study to compare BP and IS in detecting cryptocurrency breaks and outliers and found that the IS approach produce more.

Structural break tests allow us to assess when and whether there is a major change in data. The indicator saturation approach uses regression model and identifies non stationarity shifts at any point and location. According to Choi (2009), empirical results can be misleading when researchers disregard abnormal observations, particularly when it comes to dependent variables. The study assumes that the winsorization approach helps the IS approach in lessening the extreme values. Brownen (2019) and Afanasyev et al. (2019) investigated how the winsorization technique influences the performance of regression models and discovered three factors: the level of data inaccuracies in the tails, the characteristics of enterprises affected by the process, and the usage of scaling. Moir (1988) found that when the distribution is non-normal, winsorization is suggested as an alternative to trimming. Rivest (1994) also recommended the use of winsorization for skewed distributions. Dixon (1960) claims that maintaining symmetric winsorizaiton and making suitable changes will result in improved estimators. Sharma and Chatterjee (2021) discovered that Winsorization is a versatile strategy for compensating for data outliers. Yuliyani and Indahwati (2017) used winsorization in linear mixed models to detect violations of the normalcy assumption in national exam data. In this study, just 1% of the observations are winsorized in order to prevent winsorization from affecting all observations. Adams et al. (2019) found that the estimation of trimmed or winsorized least squares can still be influenced (possibly dramatically) by a single lingering outlier

Finally, the study combines two existing methodologies. The two approaches are Hendry's indicator saturation technique (1999) and Charles P. Winsor's winsorization strategy (1895-1951). There is a huge knowledge gap regarding the combination of these two approaches to increase the detectability of the indicator saturation strategy. This study addresses this gap by offering a hybrid strategy that combines the winsorization and indicator saturation approaches. The IS strategy is responsible for recognizing outliers, breaks, and trend breaks all at once, whereas winsroization aids the IS technique in reducing and mitigating extreme outliers.

3. Methodology

3.1. Datasets

The data used in this study consists of five different cryptocurrencies: Tether (USDT), Litecoin (LTC), Ripple (XRP), Ethereum (ETH), and Bitcoin (BTC). The data was obtained from https://finance.yahoo.com/. All prices are through June 30, 2023. The total number of observations for BTC, which began on November 22, 2014, LTC, which began on September 22, 2014, and XRP, ETH, and USDT, which began on November 13, 2017, are 3143, 3204, and 2056, respectively. The study employed price log-returns calculated using the following formula:

$$\mathrm{r_t} = \ln\left(\frac{\mathrm{P_t}}{\mathrm{P_{t-1}}}\right) = \ln(\mathrm{P_t}) - \ln(\mathrm{P_{t-1}}) \tag{1}$$

In this case, r_t represents returns, P_t is the lag price at time t, and P_{t-1} is the lag price from time $t-1$.

3.2. Theoretical framework

3.2.1. Winsorization Approach

The Winsorization approach is a robust test that targets reducing the impact of outliers using certain percentile values and dealing with non-normality. This method preserves the original sample size by replacing the tail of the data rather than removing it. The default percentiles of the winsorization approach are the 5th and 95th percentiles. Our strategy assigns the extreme values to a very tight percentile of returns. So, the study uses a 99% winsorization that would assign all returns below the 1st percentile to the 1st percentile value and data above the 99th percentile to the 99th percentile value. Mathematically speaking, given a sample of T observations, Winsorizing entails replacing the k highest values with the $k-1$ value and the l lowest values with the $l-1$ value, then calculating the desired statistic on the T values. For example, consider the situation of BTC with $T = 3142$ observed ordered values and $k = l = 32$, where the winsorized values are the $k + l = 32 + 32 = 64$ substituted values.

The winsorized vector of returns is achieved by

$$w(r_t) = \begin{cases} -v & \text{For } r_t \le -v \\ r_t & \text{For } |r_t| < v \\ v & \text{For } r_t \ge v \end{cases}$$

where $w(r_t)$ represents the winsorized vector, r_t represents the original returns of each cryptocurrency, v is the $k-1$ observation (nearest 99th percentile) and $-v$ is the $l-1$ observation (nearest 1st percentile).

To winsorize the returns of every cryptocurrency, the following steps have been taken. The process is based on Bitcoin, however, Table 1 shows the outcomes of applying the same process to other currencies:

1. Order each cryptocurrency data in ascending order.
2. Decide length of k and l.
3. For the case of BTC, the length of l can be found by the 1st percentile × sample size $(0.01 \times 3142 \approx 32)$, while the length of k is found by 99th percentile × sample size $(0.99 \times 3142 - 3142 \approx 3110)$.
4. In this case we have symmetric values. So, $k = l = 32$ matching 32 observations less than 33rd observation and another 32 observations greater than 3111th observation.
5. Pick the $-v$ value which is 33rd observation and v value which is 3111th observation.
6. Replace Values Greater than the 99th percentile ($k\ values$). So, all values greater than the 99th percentile are replaced with the v.
7. Replace Values Less than the 1st percentile. So, all values less than the 1st percentile are replaced with the $-v$.

Table 1: The framework of winsorized observations

Series	Sample size (T)	Winsorized observations	minVal($-v$)	maxVal($-v$)
BTCW	3142	$k = 32, l = 32, k + l = 64$	-0.11092	0.104009
LCTW	3203	$k = 32, l = 32, k + l = 64$	-0.14447	0.169612
ETHW	2055	$k = 20, l = 20, k + l = 40$	-0.14658	0.127828
USDTW	2055	$k = 20, l = 20, k + l = 40$	-0.01277	0.013738
XRPW	2055	$k = 20, l = 20, k + l = 40$	-0.15583	0.200854

Note: symbol BTCW stands for winsorized Bitcoin return.

3.2.2. Indicator Saturation Approach

Hendry (1999) introduced the IS approach. Pretis et al. (2018) state that the IS technique was created in order to identify and model outliers as well as structural breaks in the mean. Several IS estimators (ISEs) can be utilized to model distinct aspects of wide-sense non-stationarity (Castle & Hendry, 2022). Step-indicator saturation (SIS) for location shifts, impulse indicator saturation (IIS) for outliers, and trend indicator saturation (TIS) for trend breaks are a few examples. IIS is a reliable and effective statistical method, according to Hendry (1999) and Johansen et al. (2009). Both Castle et al. (2015) and Pretis et al. (2015) broadened the definition of IIS to encompass SIS and TIS.

The IS method is based on a regression model that uses a general-to-specific modeling strategy to produce indicator variables for every observation. A different approach that uses a general-to-specific procedure based on model selection is provided

by IS, claim Pretis et al. (2018). In other words, a regression model is overloaded with indicators, which are then chosen using the general-to-specific at a predefined level of significance. All indicators that are not significant are then removed without imposing co-breaking or defining a minimum or maximum break segment. By starting with a general model (the GUM) and lowering variables along search paths while assessing the diagnostics at each stage, the general-to-specific technique (GETS) offers an organized search. Marczak and Proietti (2016) state that IS has been shown to be effective and useful when used with a dynamic regression model. This IS method was developed by Pretis et al. (2018) using the GETS package in the R programming language. Three IS estimators—SIS, TIS, and IIS—are used in this study to identify breaks, trend breaks, and outliers. The following is the formulation of these estimators:

$$\text{IIS } y_t = \mu + \sum_{j=1}^{n} \delta_j 1_{\{t=j\}} + \varepsilon_t \tag{2}$$

$$\text{SIS } y_t = \mu + \sum_{j=2}^{n} \delta_j 1_{\{t\geq j\}} + \varepsilon_t \tag{3}$$

$$\text{TIS } y_t = \mu + \sum_{j=1}^{n} \delta_j 1_{\{t>j\}}(t-j) + \varepsilon_t \tag{4}$$

A break, trend break, or outlier's size is denoted by δ, errors are represented by ε, the BTC return over time is represented by y_t, and the constant term is denoted by μ. We regressed with the constant and used y_t as a dependent variable in order to apply the IS technique. According to suggestion by Ismail and Nasir (2018) we set alpha value that is based on the sample size, $\alpha = 1/T$. With an alpha value determined by sample size, the three equations were executed concurrently. The tight alpha value allows us to limit the number of significant indicators (dates); however, larger alpha values can be allowed if the interest is to increase number of changepoints.

During the application, the IS approach creates dummy variables automatically when the algorithm is executed. The IS approach will first create dummy variables representing each estimator that is equal to the number of observations in the returns. The BTC dataset generates 9423 indicators when three IS estimators (IIS, SIS, and TIS) are performed concurrently, divided into 105 blocks of 30 indicators each. Table 2 shows details of the dummy variables and their blocks.

Table 2: Dummy variables

Returns	Sample size	Alpha	Dummy Variables	
			Indicators	Blocks
BTC	3142	0.0003	9423	105
LCT	3203	0.0003	9606	107
ETH	2055	0.0005	6162	69
USDT	2055	0.0005	6162	69
XRP	2055	0.0005	6162	69

3.2.3. Incorporating Winsorization into IS approach

The incorporation begins by winsorizing the returns of each digital coin as stated. Then, each winsorized return is considered as the dependent variable. Since the IS approach is based on regression model, we regress a constant to the winsorized returns (dependent variable). Then, the IS approach automatically allows the incorporation of dummy variables into the regression equation to detect either breaks, trend breaks or outliers. Different alpha values can be considered under the null of no breaks or outlier or trend breaks. The algorithm of the IS approach estimators can be executed either separately or jointly. We allow simultaneously detection of breaks by SIS, trend breaks by TIS and outliers by IIS. The general formula of the Win-IS approach is as follows:

$$W(\mathrm{r_t}) = \mu + IS + \varepsilon_t \tag{5}$$

This is a quite general formula but since the IS approach has estimators including IIS, SIS and TIS, this general formula can be reformatted as:

$$W(\mathrm{r_t}) = \mu + \mathrm{IIS} + \mathrm{SIS} + \mathrm{TIS} + \varepsilon_t \tag{6}$$

Pretis et al. (2018) formulated the three types of IS approaches. Equation 6 can be rewritten as:

$$W(\mathrm{r_t}) = \mu + \sum_{\mathrm{j=1}}^{\mathrm{n}} \delta_{\mathrm{j}} 1_{\{\mathrm{t=j}\}} + \sum_{\mathrm{j=2}}^{\mathrm{n}} \delta_{\mathrm{j}} 1_{\{\mathrm{t\geq j}\}} + \sum_{\mathrm{j=1}}^{\mathrm{n}} \delta_{\mathrm{j}} 1_{\{\mathrm{t>j}\}}(\mathrm{t-j}) + \varepsilon_t \tag{7}$$

In this case, IS denotes the indicator saturation approach, r_t represents original returns of each coin, $W(\mathrm{r_t})$ represents winsorized vector of returns of each coin, μ represents constant term, δ represents the size of break, trend break, or outlier, and ε represents errors. Equation 7 represents the developed hybrid approach. The strategy is simply winsorizing the dependent variable and then applying IS approach. However, the Win-IS has sub formulas that can be derived from equation 4 including Win-IIS for outlier identification, Win-SIS for break identification and Win-TIS for trend break identification.

3.3. Empirical Application of Win-IS Approach

The process of empirical application of the Win-IS approach begins by:

1. Applying IS technique to identify and record the location and dates of breaks, trend breaks, and outliers in the original returns of each cryptocurrency.
2. Then winsorize each returns following steps given above.
3. Again, apply the IS approach to identify and record the location and dates of breaks, trend breaks, and outliers in the winsorized returns of each cryptocurrency.
4. Compare the performance of the IS approach in the two returns.
5. Report improvements achieved.

4. Results and Discussions

4.1. Descriptive Statistics

Table 3 displays descriptive statistics for each cryptocurrency's winsorized and original returns. The table is split into two panels. Panel A for the original returns, whereas Panel B for the winsorized returns. The results in Table 3 can be classified as central tendency measures, variability measures, and distribution tests. Except for USDT, the original and winsorized returns for each coin have positive mean as expected. This suggests that holders of these coins profited during the examined period, whilst USDT holders lost, signifying its tendency to underperform and generate losses on average. The standard deviation for both the original and winsorized results of each coin is quite high. These high standard deviations indicate significant risk and that returns are very varied or spread around the mean. However, winsorization lowered the risk marginally. According to the data range for each coin, the winsorized returns had a smaller data range than the original returns, indicating that extreme values were pushed closer to the mean. For example, in BTC, the original series ranged from -0.465 to 0.2251, but when winsorized, the minimum value increased to -0.11 and the maximum value reduced to 0.104. This broader adjustment implies a compression of the data range, bringing both extremes closer to zero. The original and winsorized returns for BTC and ETH have negative skewness, indicating a left-skewed distribution. However, LTC, USDT, and XRP have a positively skewed distribution. In contrast, all coins have positive kurtosis greater than +3 in both returns, indicating that the distributions are heavy-tailed and non-normal. Furthermore, the heavier tail of original returns suggests that they are riskier. Following winsorization, each kurtosis decreased, indicating that the tails are becoming less heavy. The Jarque-Bera (JB) test statistic confirms the kurtosis values. Original returns have significant and exceedingly high JB, indicating a large departure from the normally distribution. In comparison, the JB test for winsorized returns is substantially lower, but still significant. This means that, while both the original and winsorized returns are non-normal, the adjustment technique has reduced the divergence from normality by some amount. However, the JB test does not take into consideration variables other than skewness and kurtosis. The JB test still yields a low p-value since the non-normality may be caused by other non-normal features, outliers, or structural breaks in the data.

Table 3: Descriptive statistics

Panel A: Original Returns

Series	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis	JB
BTC	-0.465	0.2251	0.001419	0.038	-0.7895	14.2458	16883.03
LCT	-0.515	0.512	0.001011	0.055	0.103561	15.851	22044.84
ETH	-0.551	0.235	0.000880	0.0497	-0.923868	13.145	9104.12
USDT	-0.053	0.0567	-0.000005	0.004283	0.745575	53.31	216890.1
XRP	-0.551	0.6068	0.000412	0.0615	0.850	20.35	26032.13

Panel B: Winsorised Returns

Series	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis	JB
BTC	-0.1109	0.104009	0.001569	0.0347	-0.174	4.9258	501.45
LCT	-0.14447	0.169612	0.001014	0.048	0.222771	5.239	695.69
ETH	-0.146580	0.127828	0.001096	0.046	-0.243741	4.428	195.03
USDT	-0.012770	0.013738	-0.000016	0.003285	0.10443	8.9266	3011.28
XRP	-0.155830	0.200854	0.000173	0.0514	0.496705	6.2204	972.52

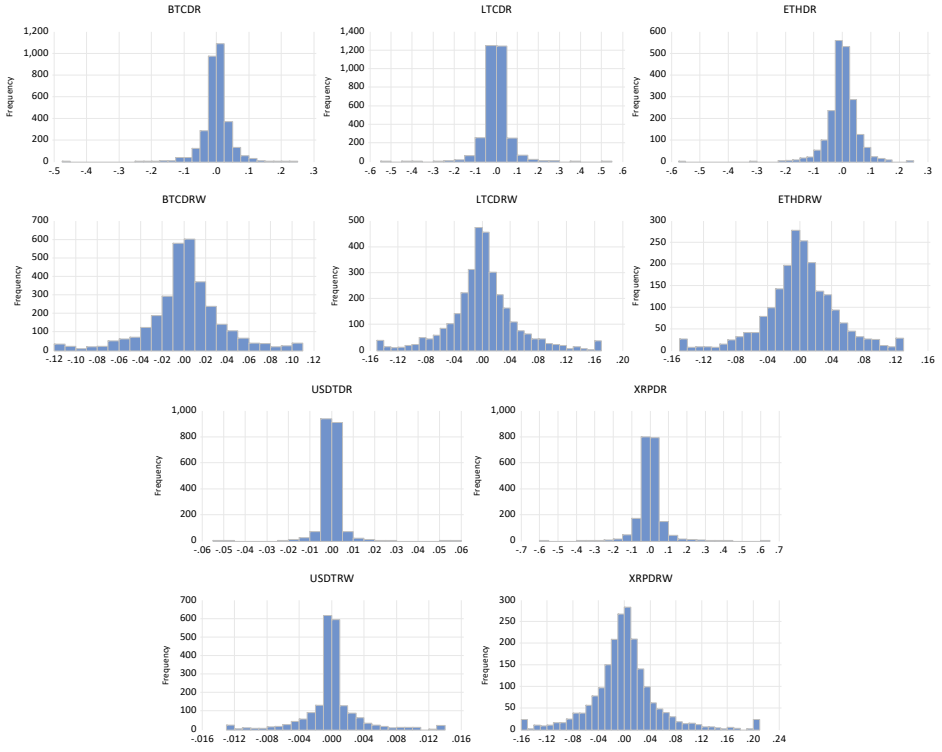


Figure 1: Histograms of Winsorized and unwinsorized returns

Figure 1 shows how the tails of each two histograms presenting winsorized and original returns have changed. This suggests that the distribution of the returns data for

each cryptocurrency has been impacted by winsorizing. The winsorizing strategy effectively capped or shaped the extreme results to a less extreme value, reducing the impact of outliers. The original data histograms show a wider distribution with more noticeable tails. The inherent uncertainty and possible non-normality in financial log returns are reflected in this broader spread. These features can make it more difficult to identify structural breaks by adding noise, even if they are crucial for capturing real-world financial dynamics. Detecting structural breaks in the dataset entails locating the points at which the data's statistical characteristics alter. Maintaining a certain amount of non-normality with returns is crucial for capturing dynamics and changes over time in the real world. We reduced the impact of extreme outliers, which otherwise have the potential to distort the results and make it difficult to identify structural fractures, by winsorizing just 1% of the data. Without completely enforcing normalcy, this adjustment results in a distribution that is more symmetrical and closer to normal. The integrity of the data's dynamic features is preserved by such a slight modification, which is essential when looking for structural breaks in time series. Thus, the winsorization acts as a tactical middle ground, improving the data's regression analysis applicability while maintaining important non-stationary components.

4.2. Comparison of IS and Win-IS Performance

4.2.1. Breaks, Trend breaks, and Outliers Detected by IS and Win-IS

Descriptive statistics in Table 3 reveal that even when severe outlier are mitigated by the winsorization strategy, their influence persists since winsorized returns exhibit non-normality. Outliers have a significant effect on summary statistics such as mean and standard deviation. With such high frequency and non-normal data returns, we investigated existing breaks, trend breaks, and outliers that continue to cause this non-normality. This section discusses the performance of IS and Win-IS approaches in detecting breaks, trend breaks, and outliers. Table 4 provides an overview of the performance of the two approaches, IS and Win-IS. Tables 5, 6, and 7 report the significant indicators recognized and retained as outliers, breaks, or/and trend breaks, respectively. Significant indicators are those retained by the method after rejecting the null hypothesis of no break, outlier, or trend break. Each table is divided into two panels, panel A gives the IS approach results and panel B the Win-IS approach results to compare the dates and total of the identified outliers, breaks, and trend breaks in each cryptocurrency. The brackets indicated by positive (+) or negative (-) sign, represent whether the shock on the identified date is up or down. Tables 5, 6, and 7 also provide additional information, such as the dates of the breaks, trend breaks, and outliers found with the two techniques. The findings of Mohamed et al. (2024) are in line with the some of the IS method results shown in these tables.

Table 4: Overall Performance of IS and Win-IS approaches

Returns	IS approach Original Returns				Win-IS approach Winsorized Returns			
	IIS (Outliers)	SIS (Breaks)	TIS (Trend Breaks)	BIC	Win-IIS (Outliers)	Win-SIS (Breaks)	Win-TIS (Trend Breaks)	BIC
BTC	25	12	5	-3.88	0	16	3	-3.89
LCT	28	11	2	-3.24	0	29	3	-3.26
ETH	13	10	8	-3.33	0	12	0	-3.34
USDT	28	8	5	-8.79	32	17	23	-8.91
XRP	26	11	9	-3.19	17	20	15	-3.24

Table 4 compares the overall performance of both IS and Win-IS approaches with 3 estimators for each. On all coins except USDT, the Win-IIS reduced number of outliers to zero in the cases of BTC, LTC, ETH and XRP while increased slightly from 28 to 32 in USDT. Win-SIS revealed more breaks in each coin than SIS. The Win-TIS technique identified fewer trend breaks than TIS tests for BTC, LTC, and ETH, but more trend breaks in USDT and XRP. Overall, the developed Win-IS reduced the number of outliers and trend breaks while revealed hidden breaks. Hence, the Win-IS approach enabled detecting hidden important breaks, trend breaks and outliers (see Table 8). Digital markets perform differently, and BTC, LTC, and ETH appear to behave similarly, as do USDT and XRP.

The comparison of the two approaches IS vs Win-IS is based on Bayesian Information Criteria (BIC). The selection is based on the lowest BIC criterion values. Table 4 shows that the Win-IS technique had the lowest BIC values in all comparisons, indicating that it outperformed the IS approach. As a result, including the winsorization strategy into the IS approach improved its effectiveness. This also emphasizes how significant the newly detected breaks and trend breaks were and how the existence of outliers could obscure them if not handled carefully (masking effect).

Table 5: IIS and Win-IIS Results (Outliers)

Series	Alpha	Panel A: IIS results (Outlier)	Total
BTC	0.0003	1/13/2015(-),1/14/2015(-), 1/15/2015(+), 8/18/2015(-), 1/15/2016(-), 1/11/2017(-), 7/17/2017(+), 7/20/2017(+), 9/14/2017(-), 9/15/2017(+), 12/06/2017(+), 12/07/2017(+), 1/16/2018(-), 2/05/2018(-), 4/02/2019(+), 6/27/2019(-), 7/16/2019(-), 10/25/2019(+), 3/12/2020(-), 3/19/2020(+), 1/21/2021(-), 2/08/2021(+), 5/12/2021(-),5/19/2021(-), 6/13/2022(-)	25
ETH	0.0005	12/22/2017(-), 9/05/2018(-), 10/11/2018(-), 9/24/2019(-), 3/08/2020(-), 3/12/2020(-), 3/13/2020(+), 3/19/2020(+), 1/03/2021(+), 1/21/2021(-), 5/19/2021(-), 5/24/2021(+), 6/21/2021(-)	13

Table 5: IIS and Win-IIS Results (Outliers) (cont.)

Series	Alpha	Panel A: IIS results (Outlier)	Total
LTC	0.0003	1/03/2015(-), 1/14/2015(-), 5/22/2015(+), 6/16/2015(+), 7/10/2015(-), 6/22/2016(-), 12/23/2016(+), 3/30/2017(+), 4/05/2017(+), 5/03/2017(+), 5/23/2017(+), 9/14/2017(-), 12/08/2017(+), 12/09/2017(+), 12/11/2017(+), 12/12/2017(+), 1/16/2018(-), 2/14/2018(+), 2/08/2019(+), 4/02/2019(+), 3/12/2020(-), 1/11/2021(-) , 5/12/2021(-), 5/19/2021(-), 5/24/2021(+), 6/21/2021(-), 9/07/2021(-), 6/30/2023(+)	28
USDT	0.0005	11/30/2017(+), 12/07/2017(+), 12/08/2017(-), 12/12/2017(+), 12/13/2017(-), 12/14/2017(-), 12/24/2017(-), 12/30/2017(+), 1/16/2018(+), 1/17/2018(-), 1/19/2018(-), 2/08/2018(+), 3/24/2018(+), 11/14/2018(-), 11/15/2018(+), 11/23/2018(-), 12/08/2018(+), 6/28/2019(+), 3/12/2020(+), 3/13/2020(-), 3/17/2020(-), 3/19/2020(+), 3/27/2020(+),3/28/2020(-), 5/06/2020(+), 5/07/2020(-), 7/03/2020(-), 8/14/2020(-)	28
XRP	0.0005	12/12/2017(+), 12/13/2017(+), 12/14/2017(+), 12/21/2017(+), 12/29/2017(+), 1/03/2018(+), 1/08/2018(-), 1/16/2018(-), 8/17/2018(+), 9/20/2018(+),9/21/2018(+), 5/14/2019(+),3/12/2020(-),11/21/2020(+), 11/23/2020(+), 12/23/2020(-), 12/24/2020(+), 1/07/2021(+), 1/30/2021(+), 2/01/2021(-), 4/10/2021(+), 4/26/2021(+), 5/19/2021(-), 5/24/2021(+), 5/11/2022(-), 3/21/2023(+)	26
Series	**Alpha**	**Panel B: WIN-IIS results (Outlier)**	**Total**
BTC	0.0003	No	0
ETH	0.0005	No	0
LTC	0.0003	No	0
USDT	0.0005	11/30/2017(+), 12/08/2017(-), 12/12/2017(+), 12/13/2017(-), 12/14/2017(-), 1/14/2018(+), 1/16/2018(+), 1/19/2018(-), 2/05/2018(-), 2/08/2018(+), 2/09/2018(-), 3/19/2018(-), 3/24/2018(+), 4/25/2018(-), 11/23/2018(-), 11/28/2018(+), 3/29/2019(-), 4/25/2019(-), 5/19/2019(+), 6/28/2019(+), 7/16/2019(-), 8/06/2019(-), 11/25/2019(-), 12/18/2019(+), 3/13/2020(-), 3/27/2020(+), 3/28/2020(-), 5/06/2020(+), 5/07/2020(-), 7/02/2020(+), 7/03/2020(-), 8/14/2020(-)	32
XRP	0.0005	12/12/2017(+), 12/13/2017(+), 12/14/2017(+), 12/21/2017(+), 12/29/2017(+), 1/03/2018(+), 8/17/2018(+), 11/21/2020(+), 11/23/2020(+), 12/24/2020(+), 4/10/2021(+),4/13/2021(+), 4/26/2021(+), 5/24/2021(+), 2/07/2022(+), 9/22/2022(+), 3/21/2023(+)	17

Table 6: SIS and Win-SIS Results (Breaks)

Series	Alpha	Panel A: SIS results (Breaks)	Total
BTC	0.0003	11/02/2015(+), 11/04/2015(-), 6/21/2016(-), 6/23/2016(+), 1/05/2017(-), 1/07/2017(+), 12/07/2017(-), 12/17/2017(-), 11/19/2018(-), 11/21/2018(+), 11/08/2022(-), 11/10/2022(+)	12
ETH	0.0005	12/11/2017(+), 12/13/2017(-), 2/06/2018(+), 11/19/2018(-), 11/21/2018(+), 5/21/2021(-), 6/10/2022(-) ,6/14/2022(+), 11/08/2022(-), 11/10/2022(+)	10
LTC	0.0003	1/24/2015(+), 1/26/2015(-), 7/05/2015(+), 5/03/2017(+), 5/08/2017(-), 5/25/2017(-), 5/27/2017(+), 6/16/2017(+), 6/18/2017(-), 5/21/2021(-), 5/25/2021(+)	11

Table 6: SIS and Win-SIS Results (Breaks) (cont.)

Series	Alpha	Panel A: SIS results (Breaks)	Total
USDT	0.0005	12/24/2017(-), 2/05/2018(-), 2/06/2018(+), 2/07/2018(-), 2/10/2018(+), 11/19/2018(-), 11/21/2018(+), 11/24/2018(-)	8
XRP	0.0005	1/08/2018(-), 1/17/2018(+), 1/19/2018(-), 2/11/2018(-), 11/24/2020(+), 4/07/2021(-), 4/08/2021(+), 4/14/2021(-), 5/25/2021(+), 6/21/2021(-), 6/23/2021(+)	11
Series	**Alpha**	**Panel B: WIN-SIS results (Breaks)**	**Total**
BTC	0.0003	1/13/2015(-), 1/15/2015(+), 11/02/2015(+), 11/04/2015(-), 6/21/2016(-), 6/23/2016(+), 1/05/2017(-), 1/07/2017(+), 3/16/2017(-), 3/19/2017(+), 11/19/2018(-), 11/21/2018(+), 12/24/2020(+), 1/09/2021(-), 11/08/2022(-), 11/10/2022(+)	16
ETH	0.0005	1/29/2018(-), 2/06/2018(+), 11/19/2018(-), 11/21/2018(+), 12/17/2018(+), 12/25/2018(-), 9/02/2020(-), 9/06/2020(+), 6/10/2022(-), 6/19/2022(+), 11/08/2022(-), 11/10/2022(+)	12
LTC	0.0003	1/24/2015(+), 1/26/2015(-), 7/10/2015(-), 7/13/2015(-), 7/17/2015(+), 3/30/2017(+), 4/06/2017(-), 4/20/2017(+), 5/10/2017(-), 5/25/2017(-), 5/27/2017(+), 6/16/2017(+), 6/18/2017(-),7/02/2017(+), 7/05/2017(-), 8/27/2017(+), 9/02/2017(-), 12/08/2017(+), 12/13/2017(-),4/02/2019(+), 4/04/2019(-), 12/16/2020(+), 12/20/2020(-), 5/21/2021(+), 5/24/2021(+), 11/08/2021(+), 11/10/2021(+), 11/08/2022(-), 11/10/2022(+)	29
USDT	0.0005	12/09/2017(-), 12/21/2017(+), 12/24/2017(-), 12/31/2017(-), 1/18/2018(+), 2/07/2018(-), 3/25/2018(+), 11/15/2018(+), 11/19/2018(-), 11/21/2018(+), 11/23/2018(-), 12/09/2018(-), 12/30/2018(-), 1/01/2019(+), 3/20/2020(-), 8/12/2020(+), 8/15/2020(-)	17
XRP	0.0005	1/08/2018(-), 1/17/2018(+), 1/19/2018(-), 2/11/2018(-), 4/18/2018(+), 4/21/2018(-), 9/18/2018(+), 9/22/2018(-), 5/16/2019(-), 11/25/2020(-), 12/25/2020(+), 1/06/2021(+), 1/08/2021(-), 2/01/2021(-), 4/07/2021(-), 5/21/2021(-), 5/25/2021(+), 8/15/2021(-), 11/08/2022(-), 11/10/2022(+)	20

Table 7: TIS and Win-TIS Results (Trend Breaks)

Series	Alpha	Panel A: TIS results (Trend Breaks)	Total
BTC	0.0003	7/13/2017(-), 7/15/2017(+), 7/17/2017(-), 12/06/2017(+), 12/23/2017(-)	5
ETH	0.0005	1/14/2018(-), 1/16/2018(+), 1/20/2018(-), 1/21/2018(+), 1/27/2018(-), 2/06/2018(+), 5/21/2021(+), 5/25/2021(-)	8
LTC	0.0003	7/12/2015(-), 7/13/2015 (+)	2
USDT	0.0005	12/20/2017(+), 12/24/2017(-), 1/18/2018(-), 1/30/2018(+), 2/03/2018(-)	5
XRP	0.0005	2/05/2018(+), 2/09/2018(-), 11/23/2020(-), 11/26/2020(+), 11/27/2020(-), 4/03/2021(+), 4/05/2021(-), 5/20/2021(-), 5/25/2021(+)	9
		Panel B: WIN-TIS results (Trend Breaks)	
BTC	0.0003	12/16/2017(-), 12/22/2017(+), 12/23/2017(-)	3
ETH	0.0005	No	0
LTC	0.0003	6/24/2015(+), 7/10/2015(+), 7/12/2015(-)	3

Table 7: TIS and Win-TIS Results (Trend Breaks) (cont.)

Series	Alpha	Panel B: WIN-TIS results (Trend Breaks)	
USDT	0.0005	12/02/2017(+), 12/14/2017(-), 12/20/2017(+), 1/03/2018(-), 1/04/2018(+), 1/06/2018(-), 1/30/2018(+), 2/10/2018(-), 3/18/2018(+), 3/19/2018(-), 3/25/2018(+), 11/12/2018(-), 11/15/2018(+), 12/05/2018(+), 12/09/2018(-), 11/21/2019(+), 11/23/2019(-), 11/26/2019(+), 3/08/2020(-), 3/09/2020(+) , 3/13/2020(-), 3/17/2020(+), 3/20/2020(-)	23
XRP	0.0005	2/05/2018(+), 2/09/2018(-), 5/12/2019(+), 5/14/2019(-), 11/26/2020(+), 11/27/2020(-), 12/15/2020(+), 12/16/2020(-), 12/25/2020(+), 1/28/2021(+), 1/30/2021(-), 4/03/2021(+), 4/08/2021(-), 8/08/2021(+), 8/15/2021(-)	15

4.2.2. Improvements of the Detectability IS Approach

In Tables 5,6, and 7 we presented the dates of the detected breaks, trend breaks, and outliers in the five cryptocurrencies. The results show that new breaks, trend breaks, and outliers were revealed after extreme observations are lessened by the winsorization approach. As discussed, this become evidence that some outliers mask some breaks or trend breaks. Table 8 show the overall number of new breaks, trend breaks, and outliers emerged together with those repeated.

Table 8: Win-IS performance and discovery

Series	Win-IIS		Win-SIS		Win-TIS		Total
	Repeated Outliers	New Outliers	Repeated Breaks	New Breaks	Repeated Trend Breaks	New Trend Breaks	
BTCDW	0	0	10	6	1	2	19
ETHDW	0	0	6	6	0	0	12
LTCDW	0	0	6	23	1	2	32
USDTDW	20	12	3	14	2	21	72
XRPDW	14	3	5	15	6	9	52
Total	**34**	**15**	**30**	**64**	**10**	**34**	**187**

Table 8 shows that, in contrast to those detected by IIS alone, Win-IIS did not find any new or recurring outliers in BTC, ETH, or LTC. Furthermore, Win-IIS identified three new outliers in XRP and 12 new outliers in USDT, whereas 14 and 20 outliers, respectively, repeated. Win-SIS discovered 94 breaks across the five markets, 30 of which were repeated as SIS detected, and revealed 64 new breaks. Win-TIS, on the other hand, revealed 44 trend breaks across five markets, 10 of which were previously spotted by TIS and 34 of which were new. Table 8 displays the distribution of 94 breaks and 44 trend breaks among the five markets. Additional details about the overall outcomes shown in Table 8 are provided in Tables 9–12. These details include the type of date (original value or winsored value), the type of estimator captured (Win-IIS, Win-SIS, and Win-TIS), and the status of each date (repeated, new, or changed to another).

By comparing and classifying the data into recurrent and emerging changes, these tables also demonstrate the presence of the masking effect.

Table 9: BTCDW and ETHDW: Emerging and Repeated Patterns Detected by Win-IS

No.	Win-IS Performance in BTCDW				Win-IS Performance in ETHDW			
	Date	Type	Estimator	Status	Date	Type	Estimator	Status
1.	1/13/2015	Winsored	Win-SIS	IIS→Win-SIS	1/29/2018	Origin	Win-SIS	New
2.	1/15/2015	Winsored	Win-SIS	IIS→Win-SIS	2/6/2018	Origin	Win-SIS	Repeated
3.	11/2/2015	Winsored	Win-SIS	Repeated	11/19/2018	Winsored	Win-SIS	Repeated
4.	11/4/2015	Origin	Win-SIS	Repeated	11/21/2018	Origin	Win-SIS	Repeated
5.	6/21/2016	Origin	Win-SIS	Repeated	12/17/2018	Origin	Win-SIS	New
6.	6/23/2016	Origin	Win-SIS	Repeated	12/25/2018	Origin	Win-SIS	New
7.	1/5/2017	Origin	Win-SIS	Repeated	9/2/2020	Origin	Win-SIS	New
8.	1/7/2017	Origin	Win-SIS	Repeated	9/6/2020	Origin	Win-SIS	New
9.	3/16/2017	Origin	Win-SIS	New	6/10/2022	Origin	Win-SIS	Repeated
10.	3/19/2017	Origin	Win-SIS	New	6/19/2022	Origin	Win-SIS	New
11.	11/19/2018	Winsored	Win-SIS	Repeated	11/8/2022	Winsored	Win-SIS	Repeated
12.	11/21/2018	Origin	Win-SIS	Repeated	11/10/2022	Winsored	Win-SIS	Repeated
13.	12/24/2020	Origin	Win-SIS	New				
14.	1/9/2021	Origin	Win-SIS	New				
15.	11/8/2022	Origin	Win-SIS	Repeated				
16.	11/10/2022	Origin	Win-SIS	Repeated				
17.	12/16/2017	Origin	Win-TIS	New				
18.	12/22/2017	Winsored	Win-TIS	New				
19.	12/23/2017	Origin	Win-TIS	Repeated				

Table 10: LTCDW: Emerging and Repeated Patterns Detected by Win-IS

Win-IS Performance in LTCDW									
No.	Date	Type	Estimator	Status	No.	Date	Type	Estimator	Status
1.	1/24/2015	Winsor	Win-SIS	Repeated	21.	4/4/2019	Origin	Win-SIS	New
2.	1/26/2015	Origin	Win-SIS	Repeated	22.	12/16/2020	Origin	Win-SIS	New
3.	7/10/2015	Winsor	Win-SIS	IIS→Win-SIS	23.	12/20/2020	Origin	Win-SIS	New
4.	7/13/2015	Origin	Win-SIS	TIS→Win-SIS	24.	5/21/2021	Winsor	Win-SIS	New
5.	7/17/2015	Origin	Win-SIS	New	25.	5/24/2021	Winsor	Win-SIS	IIS→Win-SIS
6.	3/30/2017	Winsor	Win-SIS	IIS→Win-SIS	26.	11/8/2021	Origin	Win-SIS	New
7.	4/6/2017	Origin	Win-SIS	New	27.	11/10/2021	Origin	Win-SIS	New
8.	4/20/2017	Origin	Win-SIS	New	28.	11/8/2022	Winsor	Win-SIS	New
9.	5/10/2017	Origin	Win-SIS	New	29.	11/10/2022	Winsor	Win-SIS	New
10.	5/25/2017	Origin	Win-SIS	Repeated	30.	6/24/2015	Origin	Win-TIS	New
11.	5/27/2017	Origin	Win-SIS	Repeated	31.	7/10/2015	Winsor	Win-TIS	IIS→Win-TIS
12.	6/16/2017	Origin	Win-SIS	Repeated	32.	7/12/2015	Winsor	Win-TIS	Repeated
13.	6/18/2017	Origin	Win-SIS	Repeated					
14.	7/2/2017	Origin	Win-SIS	New					
15.	7/5/2017	Origin	Win-SIS	New					
16.	8/27/2017	Origin	Win-SIS	New					
17.	9/2/2017	Origin	Win-SIS	New					
18.	12/8/2017	Winsor	Win-SIS	IIS→Win-SIS					
19.	12/13/2017	Origin	Win-SIS	New					
20.	4/2/2019	Winsor	Win-SIS	IIS→Win-SIS					

Table 11: USDTDW: Emerging and Repeated Patterns Detected by Win-IS

Win-IS Performance in USDTW									
No.	**Date**	**Type**	**Estimator**	**Status**	**No.**	**Date**	**Type**	**Estimator**	**Status**
1.	11/30/2017	Winsor	Win-IIS	Repeated	37.	1/18/2018	Origin	Win-SIS	New
2.	12/8/2017	Origin	Win-IIS	Repeated	38.	2/7/2018	Origin	Win-SIS	New
3.	12/12/2017	Winsor	Win-IIS	Repeated	39.	3/25/2018	Origin	Win-SIS	New
4.	12/13/2017	Winsor	Win-IIS	Repeated	40.	11/15/2018	Winsor	Win-SIS	IIS→Win-SIS
5.	12/14/2017	Winsor	Win-IIS	Repeated	41.	11/19/2018	Origin	Win-SIS	Repeated
6.	1/14/2018	Winsor	Win-IIS	New	42.	11/21/2018	Winsor	Win-SIS	Repeated
7.	1/16/2018	Winsor	Win-IIS	Repeated	43.	11/23/2018	Winsor	Win-SIS	IIS→Win-SIS
8.	1/19/2018	Winsor	Win-IIS	Repeated	44.	12/9/2018	Origin	Win-SIS	New
9.	2/5/2018	Winsor	Win-IIS	SIS→Win-IIS	45.	12/30/2018	Origin	Win-SIS	New
10.	2/8/2018	Winsor	Win-IIS	Repeated	46.	1/1/2019	Origin	Win-SIS	New
11.	2/9/2018	Winsor	Win-IIS	Repeated	47.	3/20/2020	Origin	Win-SIS	New
12.	3/19/2018	Winsor	Win-IIS	Repeated	48.	8/12/2020	Origin	Win-SIS	New
13.	3/24/2018	Winsor	Win-IIS	Repeated	49.	8/15/2020	Origin	Win-SIS	New
14.	4/25/2018	Winsor	Win-IIS	New	50.	12/2/2017	Origin	Win-TIS	New
15.	11/23/2018	Winsor	Win-IIS	Repeated	51.	12/14/2017	Winsor	Win-TIS	IIS→Win-TIS
16.	11/28/2018	Winsor	Win-IIS	New	52.	12/20/2017	Origin	Win-TIS	Repeated
17.	3/29/2019	Origin	Win-IIS	New	53.	1/3/2018	Origin	Win-TIS	New
18.	4/25/2019	Origin	Win-IIS	New	54.	1/4/2018	Origin	Win-TIS	New
19.	5/19/2019	Winsor	Win-IIS	New	55.	1/6/2018	Origin	Win-TIS	New
20.	6/28/2019	Winsor	Win-IIS	Repeated	56.	1/30/2018	Origin	Win-TIS	Repeated
21.	7/16/2019	Origin	Win-IIS	New	57.	2/10/2018	Origin	Win-TIS	SIS→Win-TIS
22.	8/6/2019	Origin	Win-IIS	New	58.	3/18/2018	Origin	Win-TIS	New
23.	11/25/2019	Winsor	Win-IIS	New	59.	3/19/2018	Winsor	Win-TIS	New
24.	12/18/2019	Origin	Win-IIS	New	60.	3/25/2018	Origin	Win-TIS	New
25.	3/13/2020	Winsor	Win-IIS	Repeated	61.	11/12/2018	Origin	Win-TIS	New
26.	3/27/2020	Winsor	Win-IIS	Repeated	62.	11/15/2018	Winsor	Win-TIS	IIS→Win-TIS
27.	3/28/2020	Winsor	Win-IIS	Repeated	63.	12/5/2018	Origin	Win-TIS	New
28.	5/6/2020	Origin	Win-IIS	Repeated	64.	12/9/2018	Origin	Win-TIS	New
29.	5/7/2020	Winsor	Win-IIS	Repeated	65.	11/21/2019	Origin	Win-TIS	New
30.	7/2/2020	Winsor	Win-IIS	New	66.	11/23/2019	Origin	Win-TIS	New
31.	7/3/2020	Winsor	Win-IIS	Repeated	67.	11/26/2019	Origin	Win-TIS	New
32.	8/14/2020	Winsor	Win-IIS	Repeated	68.	3/8/2020	Origin	Win-TIS	New
33.	12/9/2017	Origin	Win-SIS	New	69.	3/9/2020	Origin	Win-TIS	New
34.	12/21/2017	Origin	Win-SIS	New	70.	3/13/2020	Winsor	Win-TIS	IIS→Win-TIS
35.	12/24/2017	Winsor	Win-SIS	Repeated	71.	3/17/2020	Winsor	Win-TIS	IIS→Win-TIS
36.	12/31/2017	Origin	Win-SIS	New	72.	3/20/2020	Origin	Win-TIS	New

Repeated breaks, trend breaks, and outliers as shown in Tables 9-12 signify their importance and persistence despite certain observations being weighted down. The appearance of new outliers, breaks, and trend breaks suggests that they were important and if extreme values are not winsored, they will be buried. The Win-IS technique also enabled to redetect some of the treated data, as shown in Tables 9–12. Table 9 demonstrates, for instance, that four of the winsored extreme values in BTC appear as breaks and one as a trend break, whereas three winsored observations were classified as breaks in ETH. Tables 10–12 demonstrate that for the remaining coins, a portion of the winsorized observations is identified as either outliers, trend breaks, or breaks. Some of

the winsored observations in these tables shift to either break or trend break, indicating that they remain noteworthy. It also shows that Win-IS can reveal previously unseen data points with possibly unique characteristics or behaviors. This emphasizes the need of using sophisticated techniques, such as Win-IS, to improve the detection sensitivity of IS approaches and acquire a more thorough understanding of the changing dynamics inside market data.

Table 12: XRPDW: Emerging and Repeated Patterns Detected by Win-IS approach

Win-IS Performance in XRPTW									
No.	**Date**	**Type**	**Estimator**	**Status**	**No.**	**Date**	**Type**	**Estimator**	**Status**
1.	12/12/2017	Winsor	Win-IIS	Repeated	27.	11/25/2020	Origin	Win-SIS	New
2.	12/13/2017	Winsor	Win-IIS	Repeated	28.	12/25/2020	Origin	Win-SIS	New
3.	12/14/2017	Winsor	Win-IIS	Repeated	29.	1/6/2021	Origin	Win-SIS	New
4.	12/21/2017	Winsor	Win-IIS	Repeated	30.	1/8/2021	Origin	Win-SIS	New
5.	12/29/2017	Winsor	Win-IIS	Repeated	31.	2/1/2021	Winsor	Win-SIS	IIS→Win-SIS
6.	1/3/2018	Winsor	Win-IIS	Repeated	32.	4/7/2021	Winsor	Win-SIS	Repeated
7.	8/17/2018	Winsor	Win-IIS	Repeated	33.	5/21/2021	Winsor	Win-SIS	New
8.	11/21/2020	Winsor	Win-IIS	Repeated	34.	5/25/2021	Origin	Win-SIS	TIS→Win-SIS
9.	11/23/2020	Winsor	Win-IIS	Repeated	35.	8/15/2021	Origin	Win-SIS	New
10.	12/24/2020	Winsor	Win-IIS	Repeated	36.	11/8/2022	Origin	Win-SIS	New
11.	4/10/2021	Winsor	Win-IIS	Repeated	37.	11/10/2022	Origin	Win-SIS	New
12.	4/13/2021	Origin	Win-IIS	New	38.	2/5/2018	Winsor	Win-TIS	Repeated
13.	4/26/2021	Winsor	Win-IIS	Repeated	39.	2/9/2018	Origin	Win-TIS	Repeated
14.	5/24/2021	Winsor	Win-IIS	Repeated	40.	5/12/2019	Origin	Win-TIS	New
15.	2/7/2022	Origin	Win-IIS	New	41.	5/14/2019	Winsor	Win-TIS	Repeated
16.	9/22/2022	Origin	Win-IIS	New	42.	11/26/2020	Winsor	Win-TIS	Repeated
17.	3/21/2023	Winsor	Win-IIS	Repeated	43.	11/27/2020	Origin	Win-TIS	Repeated
18.	1/8/2018	Winsor	Win-SIS	Repeated	44.	12/15/2020	Origin	Win-TIS	New
19.	1/17/2018	Origin	Win-SIS	Repeated	45.	12/16/2020	Origin	Win-TIS	New
20.	1/19/2018	Origin	Win-SIS	Repeated	46.	12/25/2020	Origin	Win-TIS	New
21.	2/11/2018	Origin	Win-SIS	Repeated	47.	1/28/2021	Origin	Win-TIS	New
22.	4/18/2018	Origin	Win-SIS	New	48.	1/30/2021	Winsor	Win-TIS	IIS→Win-TIS
23.	4/21/2018	Origin	Win-SIS	New	49.	4/3/2021	Origin	Win-TIS	Repeated
24.	9/18/2018	Origin	Win-SIS	New	50.	4/8/2021	Origin	Win-TIS	New
25.	9/22/2018	Origin	Win-SIS	New	51.	8/8/2021	Origin	Win-TIS	New
26.	5/16/2019	Origin	Win-SIS	New	52.	8/15/2021	Origin	Win-TIS	New

Overall, the paper firstly detected breaks, trend breaks, and outliers using the IS approach in each coin concurrently. Secondly, we show that extreme values in the data may hide some significant changes due to the simultaneous execution of the three estimators of IS and due to the highly fluctuated data, by undertaking a strategy to lessen only 1% of the extreme observations in each using the winsorization approach. Thirdly, as discussed in the results, the Win-IS approach enabled to reveal new breaks, trend breaks, and outliers after extreme value were treated. The BIC value led to the decision that Win-IS results outperform the IS results. The p-values of the hypothesis also reveal significance of the new results revealed. Win-IS estimators also repeated

some of the observations as identified by IS estimators. Therefore, we emphasize that excessive values can make it difficult for statistical change tests to accurately detect where the breaks and trend breaks occur. This highlights the necessity of developing precise hybrid methodologies that might assist the existing tests in obtaining reliable results.

Finally, results from Tables 5-7 implies that the market returns encountered both upward and downward movements on different occasions which suggests that the coin market encountered phases of instability and unpredictability. Most outliers and breaks detected fall in the years 2017, 2018, 2020, and 2021. 2018 witnessed a total of 100 breaks, trend breaks, and outliers. This was followed by 88 in 2017, 59 in 2021, and 57 in 2020. These shifts can be caused by a variety of factors, including economic events, political developments, and investor sentiment. Specifically, in 2017, BTC achieved an all-time high of $20,000 and saw a rise in interest; in 2018, it marked crypto winter; and in 2020, it saw the Covid-19 epidemic.

5. Conclusions

This article improves the detectability of the IS approach by combining it with the winsorization strategy and hence proposes a technique known as Win-IS. The performance of Win-IS is then empirically compared to IS in five cryptocurrency markets. The study identified and dated outliers, breaks, and trend breaks in each market using both IS and Win-IS estimators. The Win-IS strategy outperformed the IS technique, as demonstrated by BIC scores. Furthermore, the Win-IS technique reduced severe outliers in four coins while revealing new outliers, breaks, and trend breaks, some of which were duplicated from the IS results. The repeated outliers, breaks, and trend breaks show their importance in this market because they remained constant in both winsored and original returns. The new findings demonstrate that if extreme values are not addressed, they will not be discovered. This highlights the importance of thoroughly evaluating the data before using any detection strategy, as some outliers disguise potential breaks. Subsequent research efforts may focus on adapting and expanding this hybrid methodology, as well as its relevance to other financial markets. Other methods can be compared to ours as well. The study concentrated on five digital currencies and only winsorized their first and 99th percentiles.

Although the suggested technique was first used for cryptocurrency datasets, it could have broad relevance in fields including technology research, financial markets, and economic forecasts. Additionally, the work tackles the problem of severe outliers by enhancing detectability of IS technique using Winsorization, which successfully handled tail outliers. Without making any assumptions beforehand, the enhanced IS technique is reliable in identifying outliers, trend breaks, and structural breaks,

guaranteeing thorough analysis across datasets. The study also emphasizes how the Win-IS and IS technique may concurrently capture outliers, trend breaks, and breaks. Lastly, the technique is resilient under fat-tailed distributions, even if the underlying data-generating mechanism assumes near-normal behavior. Future research might go deeper into these areas to improve the technique's resilience and usefulness.

References

Abakah, E. J. A., Gil-Alana, L. A., Madigu, G. and Romero-Rojo, F., (2020). Volatility persistence in cryptocurrency markets under structural breaks. *International Review of Economics & Finance*, 69, pp. 680–691. https://doi.org/10.1016/j.iref.2020.06.035.

Adams, J., Hayunga, D., Mansi, S., Reeb, D. and Verardi, V., (2019). Identifying and treating outliers in finance. *Financial Management*, 48(2), pp. 345–384. https://doi.org/ 10.1111/fima.12269.

Afanasyev, D. O., Fedorova, E. A., (2019). On the impact of outlier filtering on the electricity price forecasting accuracy. *Applied Energy*, 236, pp. 96–210. https://doi.org/10.1016/j.apenergy.2018.11.076.

Aharon, D. Y., Butt, H. A., Jaffri, A. and Nichols, B., (2023). Asymmetric volatility in the cryptocurrency market: New evidence from models with structural breaks. *International Review of Financial Analysis*, 87, pp. 102651. https://doi.org/10.1016/j.irfa.2023.102651.

Ané, T., Ureche-Rangau, L., Gambet, J. B. and Bouverot, J., (2008). Robust outlier detection for Asia–Pacific stock index returns. *Journal of International Financial Markets, Institutions and Money*, 18(4), pp. 326–343. https://doi.org/10.1016/j.intfin.2007. 03.001.

Bai, J., Perron, P., (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1), pp. 47–78. http://doi/10.2307/2998540.

Bai, J., Perron, P., (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1), pp. 1–22. https://doi.org/10.1002/jae.659.

Brooks, C., (2019). Introductory Econometrics for Finance, Fourth Edition. *Cambridge University Press*. https://doi.org/10.1017/9781108524872.

Brownen-Trinh, R., (2019). Effects of winsorization: The cases of forecasting non-GAAP and GAAP earnings. *Journal of Business Finance & Accounting*, 46(1–2), pp. 105–135. https://doi.org/10.1111/jbfa.12365.

Canh, N. P., Wongchoti, U., Thanh, S. D. and Thong, N. T., (2019). Systematic risk in cryptocurrency market: Evidence from DCC-MGARCH model. *Finance Research Letters*, 29, pp. 90–100. https://doi.org/10.1016/j.frl.2019.03.011.

Castle, J. L., Fawcett, N. W. and Hendry, D. F., (2011). Forecasting breaks and forecasting during breaks. *The Oxford Handbook of Economic Forecasting.* https://doi.org/ 10.1093/oxfordhb/9780195398649.013.0012.

Castle, J. L., Doornik, J. A. and Hendry, D. F., (2012). Model selection when there are multiple breaks. *Journal of Econometrics,* 169(2), pp. 239–246. https://doi.org/ 10.1016/j.jeconom.2012.01.026.

Castle, J. L., Doornik, J. A., Hendry, D. F. and Pretis, F., (2015). Detecting Location Shifts During Model Selection by Step-Indicator Saturation. *Econometrics*, 3(2), pp. 240–264. DOI: 10. 3390/econometrics3020240.

Castle, J. L., Doornik, J. A. and Hendry, D. F., (2021). Forecasting facing economic shifts, climate change and evolving pandemics. *Econometrics*, 10(1), 2. https://doi.org/ 10.3390/econometrics10010002.

Castle, J.L., Hendry, D.F., (2022). Econometrics for modelling climate change. *Oxford Research Encyclopedia of Economics and Finance*. https://doi.org/10.1093/acrefore/ 9780190625979.013.675.

Chatzikonstanti, V., (2017). Breaks and outliers when modelling the volatility of the US stock market. *Applied Economics*, *49*(46), pp. 4704-4717.

Cheng, K., Young, D. S., (2023). An Approach for Specifying Trimming and Winsorization Cutoffs. *Journal of Agricultural, Biological and Environmental Statistics*, 28, pp. 299–323. https://doi.org/10.1007/s13253-023-00527-4.

Choi, S. W., (2009). The effect of outliers on regression analysis: regime type and foreign direct investment. *Quarterly Journal of Political Science*, *4*(2), pp. 153–165.

Chow, G.C., (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28(3), pp. 591–605. https://doi.org/10.2307/1910133.

Dixon, W. J., (1960). Simplified estimation from censored normal samples. *The Annals of Mathematical Statistics*, pp. 385–391.

Dutta, A., Bouri, E., (2022). Outliers and time-varying jumps in the cryptocurrency markets. *Journal of Risk and Financial Management*, 15(3), pp.128. https://doi.org/ 10.3390/jrfm15030128.

Fearnhead, P., Rigaill, G., (2019). Changepoint detection in the presence of outliers. *Journal of the American Statistical Association,* 114(525), pp. 169-183. https://doi.org/10.1080/01621459.2017.1385466.

Ghouse, G., Aslam, A. and Bhatti, M. I., (2021). Role of Islamic banking during COVID-19 on political and financial events: Application of impulse indicator saturation. *Sustainability*, 13(21), pp. 11619. https://doi.org/10.3390/su132111619.

Ghouse, G., Bhatti, M. I. and Shahid, M. H., (2022). Impact of COVID-19, political, and financial events on the performance of commercial banking sector. *Journal of Risk and Financial Management,* 15(4), pp. 186. https://doi.org/10.3390/jrfm15040186.

Hamadani, A., Ganai, N. A., (2023). Artificial intelligence algorithm comparison and ranking for weight prediction in sheep. *Scientific Reports*, 13(1), pp. 13242. https://doi.org/10.1038/s41598-023-40528-4.

Hawkins, D. M., (1980). Identification of outliers. London: Chapman and Hall. https://doi.org/10.1007/978-94-015-3994-4.

Hendry, D. F., (1999). An econometric analysis of US food expenditure, 1931–1989. In J. R. Magnus, & M. S. Morgan (Eds.). *Methodology and tacit knowledge: two experiments in econometrics,* pp. 341–361. Chichester: John Wiley and Sons.

Hendry, D. F., Johansen, S., Santos, C., (2008). Automatic Selection of Indicators in a Fully Saturated Regression. *Computational Statistics*, 23(2), pp. 317–335. doi:10.1007/ s00180-007-0054-z.

Inclán, C., Tiao, G. C., (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427), pp. 913–923. https://doi.org/10.1080/01621459.1994.10476824.

Ismail, M. T., Nasir, I. N. M., (2018). *Structural breaks and outliers in ASEAN Shariah compliant indices: The impulse indicator saturation approach*. In AIP Conference Proceedings 1 2013.

Jiang, Z., Mensi, W. and Yoon, S. M., (2023). Risks in Major Cryptocurrency Markets: Modeling the Dual Long Memory Property and Structural Breaks. *Sustainability*, 15(3), pp. 2193. https://doi.org/10.3390/su15032193.

Johansen, S., Nielsen, B., (2016). Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models. *Scandinavian Journal of Statistics*, 43(2), pp. 321–348. doi: 10.1111/sjos.12174.

Kaseke, F., Ramroop, S. and Mwambi, H., (2022). A comparative analysis of the volatility nature of cryptocurrency and JSE market. *Investig. Manag. Financ. Innov*, 19, pp. 23–39. http://dx.doi.org/10.21511/imfi.19(4).2022.03.

Li, L., Xie, X. and Yang, J., (2021). A predictive model incorporating the change detection and Winsorization methods for alerting hypoglycemia and hyperglycemia. *Medical & Biological Engineering & Computing*, 59, pp. 2311–2324. https://doi.org/ 10.1007/s11517-021-02433-8.

Lien, D., Balakrishnan, N., (2005). On regression analysis with data cleaning via trimming, winsorization, and dichotomization. *Communications in Statistics–Simulation and Computation*, 34(4), pp. 839–849. https://doi.org/10.1080/03610910500307695.

Mandaci, P. E., Cagli, E. C., (2022). Herding intensity and volatility in cryptocurrency markets during the COVID-19, *Finance Research Letters*, 46, pp. 102382. https://doi.org/10.1016/j.frl.2021.102382.

Marczak, M., Proietti, T., (2016). Outlier detection in structural time series models: The indicator saturation approach. *International Journal of Forecasting*, 32(1), pp. 180–202. https://doi.org/10.1016/j.ijforecast.2015.04.005.

Mohamed, S. D., Ismail, M. T. and Ali, M. K. M, (2023). Detecting Structural Breaks in Cryptocurrency Market, *Jurnal Ekonomi Malaysia*, 57(2), pp. 107–22. *Jurnal Ekonomi Malaysia.* http://dx.doi.org/10.17576/JEM-2023-5702-08.

Mohamed, S. D., Ismail, M. T. and Ali, M. K. B. M. (2024). Cryptocurrency Returns Over a Decade: Breaks, Trend Breaks and Outliers. *Scientific Annals of Economics and Business*, *71*(1), pp. 1–20.

Moir, R., (1998). A Monte Carlo analysis of the Fisher randomization technique: reviving randomization for experimental economists. *Experimental Economics*, 1, pp. 87–100. https://doi.org/10.1023/A:1009961917752.

Muhammadullah, S., Urooj, A., Mengal, M. H., Khan, S. A. and Khalaj, F., (2022). Cross-Sectional Analysis of Impulse Indicator Saturation Method for Outlier Detection Estimated via Regularization Techniques with Application of COVID-19 Data. *Computational and Mathematical Methods in Medicine*. https://doi.org/10.1155/ 2022/2588534.

Mulry, M. H., Oliver, B. E. and Kaputa, S. J., (2014). Detecting and treating verified influential values in a monthly retail trade survey. *Journal of Official Statistics*, 30(4), pp. 721–747. https://doi.org/10.2478/jos-2014-0045.

Narayan, P. K., Popp, S., (2010). A new unit root test with two structural breaks in level and slope at unknown time. *Journal of Applied Statistics*, 37(9), pp. 1425–1438. https://doi.org/10.1080/02664760903039883.

Ohara, H.I.,(1999). A unit root test with multiple trend breaks: A theory and an application to US and Japanese macroeconomic time-series. *Japanese Economic Review*, 50(3), pp. 266–290. https://doi.org/10.1111/1468-5876.00119.

Panday, A., (2015). Impact of monetary policy on exchange market pressure: The case of Nepal. *Journal of Asian Economics*, 37, pp. 59–71. https://doi.org/10.1016/j.asieco.2015.02.001.

Papell, D. H., Prodan, R., (2003). The Uncertain Unit Root in U.S. Real GDP: Evidence with Restricted and Unrestricted Structural Change. *Journal of Money, Credit and Banking*, 36(3). 10.1353/mcb.2004.0059.

Pretis, F., Mann, M. L. and Kaufmann, R. K., (2015). Testing Competing Models of the Temperature Hiatus: Assessing the Effects of Conditioning Variables and Temporal Uncertainties through Sample-Wide Break Detection. *Climatic Change*, 131(4), pp. 705–718. doi: 10.1007/s10584-015-1391-5.

Pretis, F., Schneider, L., Smerdon, J. E. and Hendry, D. F., (2017). Detecting volcanic eruptions in temperature reconstructions by designed break-indicator saturation. *Environmental Economics and Sustainability*, pp. 7–37. https://doi.org/10.1002/9781119328223.ch2.

Pretis, F., Reade, J. J., Sucarrat, G., (2018). Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. *Journal of Statistical Software,* 86 (3), pp. 1–44. 10.18637/jss.v086.i03.

Quandt, R. E., (1960). Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 55(290), pp. 324–330, DOI: 10.1080/01621459.1960.10482067.

Rivest, L. P., (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika*, 81(2), pp. 373–383. https://doi.org/10.1093/biomet/81.2.373.

Rodrigues, P. M. M., Rubia, A., (2011). The effects of additive outliers and measurement errors when testing for structural breaks in variance. *Oxford Bulletin of Economics and Statistics*, 73, pp. 449–68.

Sahoo, P. K., (2021). COVID-19 pandemic and cryptocurrency markets: an empirical analysis from a linear and nonlinear causal relationship. *Studies in Economics and Finance*, 38(2), pp. 454–468. https://doi.org/10.1108/SEF-09-2020-0385.

Sharma, S., Chatterjee, S., (2021). Winsorization for robust Bayesian neural networks. *Entropy,* 23(11), pp. 1546. https://doi.org/10.3390/e23111546.

Thies, S., Molnár, P., (2018). Bayesian change point analysis of Bitcoin returns. *Finance Research Letters*, 27, pp. 223–227. https://doi.org/10.1016/j.frl.2018.03.018.

Tukey, J. W., (1962). The future of data analysis. *The annals of mathematical statistics*, *33*(1), pp. 1–67.

Xao, O. G., Yahaya, S. S. S., Abdullah, S., Yusof, Z. M., (2014, July). H-statistic with winsorized modified one-step M-estimator for two independent groups design. *AIP Conference Proceedings,* Vol. 1605, No. 1, pp. 928–931. American Institute of Physics.

Yen, T. C., Beng, K. Y., Haur, N. K., Huat, N. K., (2022). Structural change analysis of active cryptocurrency market. *Asian Academy of Management Journal of Accounting and Finance*, 18(2), pp. 63–85. https://doi.org/10.21315/aamjaf2022.18.2.4.

Yuliyani, L., Kurnia, A., Indahwati, I., (2017, March). Winsorization on linear mixed model (Case study: National exam of senior high school in West Java). In *AIP Conference Proceedings*, Vol. 1827, No. 1. AIP Publishing.

Zivot, E., Andrews, D. W. K., (2002). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business and Economic Statistics* 20(1), pp. 25–44. https://doi.org/10.1198/073500102753410372.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 183–203, https://doi.org/10.59139/stattrans-2025-010
Received – 08.07.2023; accepted – 15.07.2024

Reliability properties and applications of proportional reversed hazards in reversed relevation transform

M. Dileepkumar[1], **R. Anand**[2], **P. G Sankaran**[3]

Abstract

The concept of reversed relevation transform was introduced by Di Crescenzo and Toomaj (2015). In this article, we study important reliability properties of the reversed relevation transform under the proportional reversed hazards assumption. The results of research on information measures are presented. Various ageing concepts and stochastic orders are discussed. A new flexible generalisation of the Fréchet distribution is introduced using the proposed transformation, and reliability properties and applications are discussed.

Key words: reversed relevation transform, proportional reversed hazards model, information measures, ageing properties, stochastic orders, quantile function.

1. Introduction

Let X denote lifetime of a component with cumulative distribution function (CDF) $F_X(\cdot)$. Suppose we randomly inspect the status of the component and let Y denote the random inspection time with CDF $F_Y(\cdot)$. Then the distribution function of the random variable $X[Y]$, which denotes the total time of X given that it is less than the random inspection time Y $(i.e.\ X|X \leq Y)$ is given by

$$F_{X[Y]}(x) = F_Y(x) + F_X(x) \int_x^\infty \frac{1}{F_X(t)}\, dF_Y(t), \quad x \geq 0. \tag{1.1}$$

Di Crescenzo and Toomaj (2015) called (1.1) the reversed relevation transform of X and Y. This can be viewed as a dual concept of the well-known relevation transform introduced and studied by Krakowski (1973). When X and Y are identically distributed $(i.e.\ F_Y(x) = F_X(x))$, then (1.1) becomes

$$F_{X[Y]}(x) = F_X(x)(1 - \log F_X(x)). \tag{1.2}$$

Di Crescenzo and Toomaj (2015) studied various properties of a sequence of random variables formed by the repeated application of the reversed relevation transform. Kayal (2016) introduced a generalization of the cumulative entropy (Di Crescenzo and Longobardi

[1]Department of Statistics, University of Calicut, Kerala, India. E-mail: kumarmdileep@gmail.com. ORCID: https://orcid.org/0000-0003-4124-5306.
[2]Department of Statistics, University of Calicut, Kerala, India. E-mail: anand94.r@gmail.com. ORCID: https://orcid.org/0000-0001-7008-9483.
[3]Department of Statistics, Cochin University of Science and Technology, Kerala, India. E-mail: sankaran.pg@gmail.com. ORCID: https://orcid.org/0009-0003-0163-8858.

(2009)) using the idea of the reversed relevation transform. Some results connecting n-fold reversed relevation transform and generalized cumulative residual entropy (GCRE) (Psarrakos and Navarro (2013)) were given by Di Crescenzo and Toomaj (2017). A past inaccuracy measure based on the reversed relevation transform was studied by Di Crescenzo et al. (2018).

Let $f_X(x)$ denote the density function and $\bar{F}_X(x) = 1 - F_X(x)$ represent the survival function of a random variable X. Then the hazard rate of X, defined as $h_X(x) = \frac{f_X(x)}{\bar{F}_X(x)}$ gives the instantaneous rate of failure at any given time of the object under study. Another measure of peculiar interest is the reversed hazard rate, which is defined as $\lambda_X(x) = \frac{f_X(x)}{F_X(x)}$. In the context of lifetime studies, the reversed hazard rate has a crucial role when time elapsed since failure is a quantity of interest in order to predict the actual time of failure. Various properties and applications of the reversed hazard rate can be seen in Block et al. (1998), Chandra and Roy (2001), Gupta and Nanda (2001), Finkelstein (2002) and Chechile (2011). In a parallel system of independent and identically distributed components, we can see that the reversed hazard rate of the system lifetime is proportional to the reversed hazard rate of the lifetime of each component. Lehmann (1953) introduced the concept of the proportional reversed hazards model (PRHM) in contrast to the well-known proportional hazards model (PHM), which is commonly used in reliability theory and survival analysis. Let $\lambda_X(\cdot)$ and $\lambda_Y(\cdot)$ be the reversed hazard rates of X and Y respectively. Then Y is said to be the PRHM of X with proportionality constant θ if

$$\lambda_Y(x) = \theta \lambda_X(x), \quad \theta > 0. \tag{1.3}$$

An equivalent form of (1.3) is

$$F_Y(x) = (F_X(x))^{\theta}, \quad \theta > 0. \tag{1.4}$$

PRHM can accommodate non-monotonic hazard rates even though the baseline hazard rate is monotonic. Mudholkar and Srivastava (1993), Mudholkar et al. (1995), Mudholkar and Hutson (1996), Gupta et al. (1998), Gupta and Kundu (1999, 2001, 2002, 2007), Sarhan and Kundu (2009), Mahmoud and Alam (2010), Popović et al. (2022) and several other authors have studied the importance of PRHM model in various lifetime contexts. Moreover, certain characterization results, ageing properties and stochastic orders of the PRHM can be seen in Di Crescenzo (2000), Gupta and Wu (2001), Kundu and Gupta (2004), Gupta and Kundu (2007) and Shojaee and Babanezhad (2023). Under the assumption of PRHM between X and Y, we call the transform (1.1) as the proportional reversed hazards in the reversed relevation transform (PRHRRT). The aim of the present paper is to uncover special properties and applications of PRHRRT in reliability context. Throughout the paper, the terms increasing and decreasing are used in a wide sense, that is, a function g is increasing (decreasing) if $g(x) \leq (\geq) g(y)$ for all $0 < x < y$. Whenever we use a derivative, an expectation, or a conditional random variable, we are tacitly assuming that it exists.

The remainder of this article is organized as follows. In Section 2, the concept of PRHRRT model is introduced and its basic structural properties are studied. Various reliability properties and certain interesting results based on information measures are

discussed in Section 3. Ageing properties and stochastic orders of PRHRRT are studied in Sections 4 and 5 respectively. In Section 6, we introduce a new generalization of the Fréchet distribution using the concept of PRHRRT and present its distributional properties and applications. Finally, Section 7 provides major conclusions of the study.

2. Proportional reversed hazards in reversed relevation transform

Let X and Y be two non-negative random variables with absolutely continuous CDFs $F_X(\cdot)$ and $F_Y(\cdot)$ respectively. Suppose Y is the PRHM of X, as defined in (1.4). Then the reversed relevation random variable $X[Y]$ has the distribution function of the form

$$F_{X[Y]}(x) = (F_X(x))^\theta + F_X(x) \int_x^\infty \frac{1}{F_X(t)} d\,(F_X(t))^\theta, \quad x \geq 0,\ \theta > 0. \tag{2.1}$$

Di Crescenzo and Toomaj (2015) have showed that the reversed relevation transform is commutative under the assumption of PRHM (*i.e.* $X[Y] \stackrel{d}{=} Y[X]$). When $\theta = 1$, (2.1) reduces to (1.2) and hence in the present study we assume that $\theta \neq 1$. We now establish an identity connecting the distribution functions of $X[Y]$ and the baseline random variable X.

Proposition 2.1. Let X and Y be two non-negative random variables with absolutely continuous CDFs $F_X(x)$ and $F_Y(x)$ respectively. Then Y is the PRHM of X if and only if $F_{X[Y]}(x)$ satisfies

$$F_{X[Y]}(x) = \frac{\theta F_X(x) - (F_X(x))^\theta}{\theta - 1}, \quad \theta > 0. \tag{2.2}$$

Proof. Let Y be the PRHM of X. Then from Di Crescenzo and Toomaj (2015) (Proposition 2), the identity (2.2) follows. Now, to prove the converse part, assume that

$$F_Y(x) + F_X(x) \int_x^\infty \frac{1}{F_X(t)} dF_Y(t) = \frac{\theta F_X(x) - (F_X(x))^\theta}{\theta - 1}.$$

Rearranging and taking the first derivative with respect to x on both sides gives

$$\begin{aligned} \frac{-f_Y(x)}{F_X(x)} &= \frac{1}{(\theta-1)^2 (F_X(x))^2} \Big[(\theta-1) F_X(x) \Big(\theta f_X(x) - \theta (F_X(x))^{\theta-1} f_X(x) - \\ &\quad (\theta-1) f_Y(x) \Big) - \Big(\theta F_X(x) - (F_X(x))^\theta - (\theta-1) F_Y(x) \Big) (\theta-1) f_X(x) \Big]. \end{aligned}$$

Upon simplification, we get

$$\frac{f_X(x)\,(F_X(x))^\theta}{(F_X(x))^2} = \frac{f_X(x) F_Y(x)}{(F_X(x))^2} \implies F_Y(x) = (F_X(x))^\theta, \quad \text{for all } x \geq 0,\ \theta > 0.$$

This completes the proof. □

Remark 2.1. The CDF of $X[Y]$ given in (2.2) can be represented in a mixture form as

$$F_{X[Y]}(x) = \phi F_X(x) + (1-\phi)\left(F_X(x)\right)^{\theta} = \phi F_X(x) + (1-\phi)F_Y(x), \tag{2.3}$$

where $\phi = \frac{\theta}{\theta-1}$ and one of the weights is negative depending on the value of ϕ.

Let $f_{X[Y]}(x)$ denote the density function of the random variable $X[Y]$. Then from (2.2), we get

$$f_{X[Y]}(x) = f_X(x)\left(\frac{\theta}{\theta-1}\left(1-\left(F_X(x)\right)^{\theta-1}\right)\right), \tag{2.4}$$

where $f_X(x)$ is the density function of X. An equivalent representation of (2.2) in terms of the survival function of X, Y and $X[Y]$ denoted respectively by $\bar{F}_X(\cdot)$, $\bar{F}_Y(\cdot)$ and $\bar{F}_{X[Y]}(\cdot)$ is as follows:

$$\bar{F}_{X[Y]}(x) = \frac{\theta\bar{F}_X(x) - \bar{F}_Y(x)}{\theta-1}. \tag{2.5}$$

Now, the expected value of $X[Y]$ can be evaluated as follows:

$$\begin{aligned} E(X[Y]) &= \int_0^{\infty} \bar{F}_{X[Y]}(x)\,dx = \frac{1}{\theta-1}\int_0^{\infty}\left(\theta\bar{F}_X(x) - \left(\bar{F}_X(x)\right)^{\theta}\right)dx \\ &= \frac{1}{\theta-1}\int_0^{\infty}\left(F_X(x) + \theta\bar{F}_X(x) - \theta\bar{F}_X(x) - \left(F_X(x)\right)^{\theta}\right)dx \\ &= E(X) + T_X(\theta), \end{aligned} \tag{2.6}$$

where $T_X(\theta) = \frac{1}{\theta-1}\int_0^{\infty}\left(F_X(x) - \left(F_X(x)\right)^{\theta}\right)dx,\ \theta > 0,\ \theta \neq 1$ is the cumulative Tsallis past entropy (CTE), introduced and studied by Calì et al. (2017). From (2.6), the CTE of X can be evaluated as

$$T_X(\theta) = E(X) - E(X[Y]), \tag{2.7}$$

The identity (2.7) can be used for constructing simple non-parametric estimator for $T_X(\theta)$ by using the estimators of $E(X)$ and $E(X[Y])$.

In reliability theory, PHM models plays a vital role in the comparison of the lifetime of two components. The random variables X and Y satisfy PHM if

$$h_Y(x) = \theta h_X(x),\ \ \theta > 0, \tag{2.8}$$

where $h_Y(x) = \frac{f_Y(x)}{\bar{F}_Y(x)}$ and $h_X(x) = \frac{f_X(x)}{\bar{F}_X(x)}$ are the hazard rates of X and Y respectively. An equivalent representation of (2.8) is $\bar{G}(x) = (\bar{F}(x))^{\theta},\ \ \theta > 0$. For more details on PHM, one could refer to Kalbfleisch and Prentice (2002) and Lawless (2003). When Y is the PRHM of X with proportionality constant θ, the CDF of $X[Y]$ has the form (2.2). Now, in the next proposition, for $\theta = 2$, we provide an interesting characterization of PRHRRT.

Proposition 2.2. Let X and Y be two lifetime random variables. Then, Y is the PRHM of X with proportionality constant $\theta = 2$ if and only if $X[Y]$ is the PHM of X with the same proportionality constant.

Proof. Suppose $F_Y(x) = (F_X(x))^2$. Then from (2.2), we have

$$F_{X[Y]}(x) = 2F_X(x) - (F_X(x))^2 \iff \bar{F}_{X[Y]}(x) = 1 - 2F_X(x) + (F_X(x))^2 \iff \bar{F}_{X[Y]}(x) = (\bar{F}_X(x))^2.$$

Thus, $X[Y]$ is the PHM of X with proportionality constant 2, which completes the proof. □

Remark 2.2. Suppose that the family of distributions of X is invariant under PHM (*i.e.* X and the corresponding PHM random variable belongs to the same family of distributions) and Y is the PRHM of X with proportionality constant $\theta = 2$. Then X is invariant under PRHRRT. For example, under the aforementioned setup, X is exponential with mean λ if and only if $X[Y]$ is exponential with mean $\frac{\lambda}{2}$.

The concept of odds ratio is well known in epidemiological research, serving as a measure of the approximate relative risk of an event, like disease or death, with or without a specific factor. Now, if we define X as an individual's lifespan, extending the event to encompass 'failure occurring by time x' for all $x > 0$, the odds function $\phi_X(\cdot)$, of X can be represented as follows:

$$\phi_X(x) = \frac{P(X > x)}{P(X \leq x)} = \frac{\bar{F}_X(x)}{F_X(x)}.$$

Note that the odds function is a decreasing function of x. For more details on properties and applications of odds functions, one could refer to Bennett (1983), Zimmer et al. (1998), Navarro et al. (2008), Khorashadizadeh et al. (2013), and the references therein.

Proposition 2.3. Let $\phi_X(x)$, $\phi_Y(x)$ and $\phi_{X[Y]}(x)$ denote the odds functions of X, Y and $X[Y]$ respectively. Then Y is the PRHRRT of X if and only if

$$\phi_{X[Y]}(x) = \frac{\theta\,\phi_X(x) - \phi_Y(x)\,(F_X(x))^{\theta-1}}{\theta - (F_X(x))^{\theta-1}}. \tag{2.9}$$

Proof. Under the assumption of PRHM, we have

$$\begin{aligned} \phi_{X[Y]}(x) &= \frac{\bar{F}_{X[Y]}(x)}{F_{X[Y]}(x)} = \frac{\theta\bar{F}_X(x) - (1 - (F_X(x))^\theta)}{\theta F_X(x) - (F_X(x))^\theta} \\ \iff \phi_{X[Y]}(x) &= \frac{\theta\,\phi_X(x) - \phi_Y(x)\,(F_X(x))^{\theta-1}}{\theta - (F_X(x))^{\theta-1}}, \end{aligned}$$

where $\phi_Y(x) = \frac{1-(F_X(x))^\theta}{(F_X(x))^\theta} = \frac{\bar{F}_Y(x)}{F_Y(x)}$. □

3. Reliability properties

The hazard rate of the random variable $X[Y]$ under the assumption of PRHM has the form

$$h_{X[Y]}(x) = \frac{f_{X[Y]}(x)}{1 - F_{X[Y]}(x)} = h_X(x)\left(\frac{\theta(1 - F_X(x))\left(1 - (F_X(x))^{\theta-1}\right)}{\theta(1 - F_X(x)) - \left(1 - (F_X(x))^\theta\right)}\right)$$

$$= \quad h_X(x)\left(\frac{1-\frac{F_Y(x)}{F_X(x)}}{1-\frac{\bar{F}_Y(x)}{\theta\bar{F}_X(x)}}\right), \tag{3.1}$$

where $h_X(x)$ is the hazard rate of X.

Let $m_{X[Y]}(x)$ denote the mean residual life of the random variable $X[Y]$, defined by

$$m_{X[Y]}(x) = \frac{1}{\bar{F}_{X[Y]}(x)}\int_x^\infty \bar{F}_{X[Y]}(t)\,dt, \quad x>0. \tag{3.2}$$

On integrating (2.5) over the interval (x,∞), we get

$$\int_x^\infty \bar{F}_{X[Y]}(t)\,dt = \frac{\theta}{\theta-1}\int_x^\infty \bar{F}_X(t)\,dt - \frac{1}{\theta-1}\int_x^\infty \bar{F}_Y(t)\,dt, \quad x>0. \tag{3.3}$$

This gives

$$\begin{aligned} m_{X[Y]}(x)\bar{F}_{X[Y]}(x) &= \frac{\theta}{\theta-1}m_X(x)\bar{F}_X(x) - \frac{1}{\theta-1}m_Y(x)\bar{F}_Y(x) \\ \Longrightarrow\ m_{X[Y]}(x) &= \frac{\theta\, m_X(x)\bar{F}_X(x) - m_Y(x)\bar{F}_Y(x)}{\theta\bar{F}_X(x)-\bar{F}_Y(x)}, \end{aligned} \tag{3.4}$$

where $m_X(x)$ and $m_Y(x)$ are the mean residual life functions of X and Y respectively.

The mean inactivity time of $X[Y]$ has the form

$$\mu_{X[Y]}(x) = \frac{1}{F_{X[Y]}(x)}\int_0^x F_{X[Y]}(t)\,dt = \frac{\theta\,\mu_X(x) - (F_X(x))^{\theta-1}\mu_Y(x)}{\theta-(F_X(x))^{\theta-1}}, \tag{3.5}$$

where $\mu_X(x)$ and $\mu_Y(x)$ are the mean inactivity times of X and Y respectively.

Glaser's function of a random variable X with density function $f_X(x)$ is defined as $\eta_X(x) = -\frac{f_X'(x)}{f_X(x)}$ (Glaser (1980)), where prime denotes the first derivative. It is used as an alternative for the hazard rate in lifetime studies. Under the PRHM assumption between X and Y, the Glaser's function of $X[Y]$ satisfies the identity

$$\eta_{X[Y]}(x) = \eta_X(x)\left(\frac{(F_X(x))^\theta\left((\theta-1)(F_X(x))^2\right) - (F_X(x))^2 f_X'(x)}{F_X(x)f_X(x)f_X'(x)\left(F_X(x)-(F_X(x))^\theta\right)}\right). \tag{3.6}$$

The reversed hazard rate of $X[Y]$ is given by

$$\begin{aligned} \lambda_{X[Y]}(x) = \frac{f_{X[Y]}(x)}{F_{X[Y]}(x)} &= \theta\lambda_X(x)\left(\frac{F_X(x)-F_Y(x)}{\theta F_X(x)-F_Y(x)}\right) \\ &= \lambda_Y(x)\left(\frac{F_X(x)-F_Y(x)}{\theta F_X(x)-F_Y(x)}\right). \end{aligned} \tag{3.7}$$

The identities (2.5), (3.1), (3.4), (3.5), (3.6) and (3.7) are useful for obtaining the aforementioned reliability measures of $X[Y]$ from those of the baseline random variables X and Y. Moreover, we can make use of these identities to establish various ageing and ordering properties of $X[Y]$ without knowing the distribution of $X[Y]$.

3.1. Distorted representation

A distortion function, $q(u)$, is a non-decreasing function from $[0,1]$ to $[0,1]$, such that $q(0) = 0$ and $q(1) = 1$. Suppose that X and Y are two random variables with survival functions $\bar{F}_X(x)$ and $\bar{F}_Y(x)$ respectively. Then Y is said to be the distorted random variable of X if $\bar{F}_Y(x) = q(\bar{F}_X(x))$, where $q(u)$ is a distortion function. Denneberg (1990) introduced the concept of distortion functions, and later it gained wide popularity in the areas of actuarial science, insurance, economics, and risk analysis. The importance of distorted random variables in reliability studies has been pointed out by various researchers, such as Wang (1996), Sordo and Suárez-Llorens (2011), Navarro et al. (2013, 2014, 2016), Sordo et al. (2015) and Navarro (2022).

Proposition 3.1. If Y is the PRHM of X, then $X[Y]$ is a distorted random variable of X with distortion function

$$q(u) = \frac{1}{\theta - 1}(\theta u - (1 - (1-u)^{\theta})). \tag{3.8}$$

Proof. Since Y is the PRHM of X, from (2.5), the survival function $\bar{F}_{X[Y]}(x)$ can be expressed as

$$\bar{F}_{X[Y]}(x) = q(\bar{F}_X(x)), \quad \text{where } q(u) = \frac{1}{\theta - 1}(\theta u - (1 - (1-u)^{\theta})),\ u \in [0,1].$$

We can easily verify that $q(u)$ given in (3.8) is a distortion function. Thus, $X[Y]$ is a distorted random variable of X with distortion function $q(u)$. □

Note that the distortion function given in (3.8) is a convex function. Expressing $X[Y]$ as a distorted random variable of X will be useful in studying the preservation of various ageing properties from X to $X[Y]$ and establishing stochastic order relations between X and $X[Y]$. We consider this in Sections 4 and 5.

4. Ageing properties

In this section, we discuss some of the ageing properties of $X[Y]$ in connection with the baseline random variable X. Let X be a lifetime random variable with CDF $F_X(x)$, density function $f_X(x)$, survival function $\bar{F}_X(x)$, hazard rate $h_X(x)$ and reversed hazard rate $\lambda_X(x)$. We consider the following ageing properties;

(i) X is said to have an increasing (decreasing) hazard rate (*i.e.* IHR (DHR)) if the hazard rate $h_X(x)$ is increasing (decreasing).

(ii) X is said to have an increasing (decreasing) hazard rate average (*i.e.* IHRA (DHRA)) if $\frac{1}{x}\int_0^x h_X(u)\,du$ is increasing (decreasing).

(iii) X is new better (worse) than used (*i.e.* NBU (NWU)) if $\bar{F}_X(x+t) \leq (\geq)\ \bar{F}_X(x)\bar{F}_X(t)$, for all $x,t > 0$.

(iv) X is new better (worse) than used in hazard rate (*i.e.* NBUHR (NWUHR)) if $h_X(0) \leq (\geq) h_X(x)$, for all $x > 0$.

(v) X is said to have an increasing (decreasing) reversed hazard rate (*i.e.* IRHR (DRHR)) if $\lambda_X(x)$ is increasing (decreasing).

(vi) X is said to have an increasing (decreasing) likelihood ratio (*i.e.* ILR (DLR)) if $\log f_X(x)$ is concave (convex).

For more details on ageing properties and their applications, one may refer to Barlow and Proschan (1975), Lai and Xie (2006), Navarro (2022) and Breneman et al. (2022). In the context of coherent systems having independent and identical components, Navarro et al. (2014) showed that the system lifetime S is a distorted random variable of the component lifetime X with distortion function, say $q(u)$. Since $X[Y]$ is a distorted random variable of X, in the next proposition we present conditions for the preservation of reliability classes under the formation of PRHRRT by adopting results from Navarro et al. (2014).

Proposition 4.1. Let X and Y be two lifetime random variables, with CDFs $F_X(x)$ and $F_Y(x)$ respectively. Let $X[Y]$ be the reversed relevation of X and Y. Assume that Y is the PRHM of X. Then we have the following;

(i) For $\theta \geq 2$, X is IHR $\implies$ $X[Y]$ is IHR.

(ii) For $0 < \theta \leq 2$, X is DHR $\implies$ $X[Y]$ is DHR.

(iii) For $\theta > 0$, X is DRHR $\implies$ $X[Y]$ is DRHR.

(iv) For $0 < \theta \leq 2$, X is DLR $\implies$ $X[Y]$ is DLR.

(v) For $\theta > 0$, X is NWU $\implies$ $X[Y]$ is NWU.

(vi) For $\theta > 0$, X is DHRA $\implies$ $X[Y]$ is DHRA.

Proof. Consider the PRHRRT model given in (2.2). From (3.8), we have the distortion function connecting X and $X[Y]$ as $q(u) = \frac{(1-u)^{\theta}+\theta u-1}{\theta-1},\ u \in [0,1]$. By recalling the results from Navarro et al. (2014) in the context of coherent systems, we have, if $\tau(u) = \frac{uq'(u)}{q(u)}$ is decreasing (increasing) in $(0,1)$ then the IHR (DHR) property will be preserved with respect to the distortion function $q(u)$. Thus, for proving (i) and (ii), we have to examine the monotonicity of $\tau(u) = \frac{uq'(u)}{q(u)} = \frac{u\left(\theta-\theta\ (1-u)^{\theta-1}\right)}{(1-u)^{\theta}+\theta\ u-1}$. For this, we have

$$\tau'(u) = \frac{\theta\left(u\left(\left((\theta-1)^2u-2\right)(1-u)^{\theta}-u+2\right)-\left((1-u)^{\theta}-1\right)^2\right)}{(u-1)^2\left((1-u)^{\theta}+\theta u-1\right)^2}. \tag{4.1}$$

The denominator of (4.1) is always non-negative, and by analyzing the numerator, we observe that the right-hand side is strictly positive for $0 < \theta < 2$, strictly negative for $\theta > 2$ and zero for $\theta = 2$. This completes the proof for (i) and (ii).
Again, from Navarro et al. (2014), we have the result that, if $k(u) = \frac{uq'(1-u)}{1-q(1-u)}$ is decreasing in $(0,1)$ then the DRHR property will be preserved from X to $X[Y]$. We have $k(u) = \frac{uq'(1-u)}{1-q(1-u)} = \frac{\theta\left(u^\theta - u\right)}{u^\theta - \theta\, u}$. On differentiating $k(u)$ with respect to u, we get

$$k'(u) = -\frac{(\theta-1)^2\theta u^\theta}{\left(u^\theta - \theta u\right)^2} \leq 0, \quad \text{for all } \theta > 0 \text{ and } u \in (0,1).$$

Thus, $k(u)$ is decreasing in u and hence the proof of (iii) follows.

Let $l(u) = \frac{uq''(u)}{q'(u)} = -\frac{(\theta-1)u(1-u)^{\theta-1}}{(1-u)^\theta + u - 1}$. From (3.8) we have

$$l'(u) = -\frac{(\theta-1)(1-u)^{\theta-2}\left((1-u)^\theta + u(\theta - \theta u + u) - 1\right)}{\left((1-u)^\theta + u - 1\right)^2},$$

which is non-negative when $0 < \theta \leq 2$ for all $u \in (0,1)$. Now, from Navarro et al. (2014) (Proposition 2.2), proof of (iv) follows.

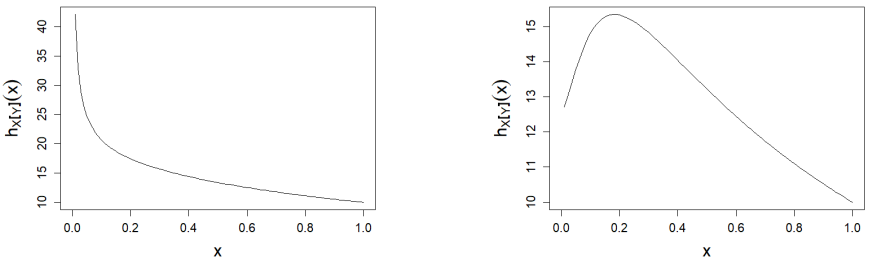
It is easy to verify that the distortion function $q(u)$ is super-multiplicative (*i.e.* $q(u\,v) \geq q(u)\,q(v)$, for all $0 \leq u,\, v \leq 1$). This inequality with Proposition 2.7 of Navarro et al. (2014) completes the proof of (v).

Similarly $q(u)$ satisfies the inequality $q(u^a) \geq (q(u))^a$ for $0 < a < 1$. Now, proof of (vi) follows from Navarro et al. (2014) (Proposition 2.8). □

Example 4.1. Let X be a random variable having a Burr type-XII distribution, with CDF $F_X(x) = 1 - \left(\frac{1}{1+x}\right)^c,\ x > 0,\ c > 0$. Then the hazard rate of X is $h_X(x) = \frac{c}{1+x}$, which is decreasing for all parameter values. Suppose Y is the PRHM of X, then the hazard rate of $X[Y]$ has the form

$$h_{X[Y]}(x) = \frac{c\,\theta\left(\frac{1}{x+1}\right)^{c+1}\left(\left(1-\left(\frac{1}{x+1}\right)^c\right)^\theta + \left(\frac{1}{x+1}\right)^c - 1\right)}{\left(\left(\frac{1}{x+1}\right)^c - 1\right)\left(\theta\left(\frac{1}{x+1}\right)^c + \left(1-\left(\frac{1}{x+1}\right)^c\right)^\theta - 1\right)}.$$

Figure 1(a) illustrate the preservation of DHR property when $0 < \theta \leq 2$. For $\theta > 2$, DHR property will not be preserved as shown in Figure 1(b). Observe that $X[Y]$ has DHR and Upside-down Bathtub (UBT) shaped hazard rates for various parameter combinations while the baseline is always DHR.



(a) $\theta = 0.5$ and $c = 10$ **(b)** $\theta = 5$ and $c = 10$

Figure 1: Plots of $h_{X[Y]}(x)$ for various parameter combinations.

5. Stochastic orders

Stochastic orders are used to compare the characteristics of two lifetime random variables. This section aims to provide different stochastic order relations between X and $X[Y]$. Let X and Y be two continuous lifetime random variables, with CDFs $F_X(x)$ and $F_Y(x)$ respectively. Let $f_X(x)$ and $f_Y(x)$ be the corresponding density functions. Then we have the following:

(i) X is smaller than Y in usual stochastic order, denoted by $X \leq_{st} Y$ if and only if $\bar{F}_X(x) \leq \bar{F}_Y(x)$ for all x.

(ii) X is smaller than Y in hazard rate order, denoted by $X \leq_{hr} Y$ if and only if $\frac{\bar{F}_Y(x)}{\bar{F}_X(x)}$ is increasing in x.

(iii) X is smaller than Y in likelihood ratio order, denoted by $X \leq_{lr} Y$ if and only if $\frac{f_Y(x)}{f_X(x)}$ is increasing in the set of union of their supports.

(iv) X is smaller than Y in increasing convex order, denoted by $X \leq_{icx} Y$ if and only if $\int_x^\infty \bar{F}_X(t)\, dt \leq \int_x^\infty \bar{F}_Y(t)\, dt$ for all x.

(v) X is smaller than Y in convex ordering, denoted by $X \leq_c Y$ if $F_Y^{-1}(F_X(x))$ is convex.

More properties and applications of stochastic orders can be seen in Shaked and Shanthikumar (2007), Belzunce et al. (2016) and Kochar (2022). Di Crescenzo and Toomaj (2015) showed that $X[Y] \leq_{st} X$. In the coming propositions, we establish interesting order properties between $X[Y]$, X and Y under the PRHM assumption between X and Y.

Proposition 5.1. Let Y be the PRHM of X and $X[Y]$ is the corresponding reversed relevation random variable. Then $X[Y] \leq_{lr} \min\{X, Y\}$

Proof. It is enough to show that $X[Y] \leq_{lr} X$ and $X[Y] \leq_{lr} Y$. For this, note that $X[Y]$ and X can be represented as distorted forms of X with respective distortion functions $q_1(u) = \frac{1}{\theta-1}(\theta u - (1-(1-u)^\theta))$ and $q_2(u) = u$. By recalling the results from Navarro et

al. (2013) in the context of stochastic orders between two coherent systems having identical components, we have

$$X[Y] \leq_{lr} (\geq_{lr}) X \text{ if and only if } \frac{q_1'(u)}{q_2'(u)} \text{ is increasing (decreasing) in } u \in (0,1). \tag{5.1}$$

Note that

$$\frac{d}{du}\left(\frac{q_1'(u)}{q_2'(u)}\right) = \frac{d}{du}\left(\frac{\theta-\theta(1-u)^{\theta-1}}{\theta-1}\right) = \theta(1-u)^{\theta-2} > 0, \text{ for all } \theta > 0 \text{ and } u \in (0,1).$$

Thus, $\frac{q_1'(u)}{q_2'(u)}$ is increasing in $u \in (0,1)$ and thus from (5.1), we have $X[Y] \leq_{lr} X$.
In similar lines, we can form $X[Y]$ and Y by distorting Y using the distortion functions $r_1(u) = \frac{\theta\left(1-(1-u)^{\frac{1}{\theta}}\right)-u}{\theta-1}$ and $r_2(u) = u$ respectively. This gives

$$\frac{r_1'(u)}{r_2'(u)} = \frac{(1-u)^{\frac{1}{\theta}-1}-1}{\theta-1}.$$

Note that $\frac{d}{du}\left(\frac{r_1'(u)}{r_2'(u)}\right) = \frac{(1-u)^{\frac{1}{\theta}-2}}{\theta} > 0$, for all $\theta > 0$ and $u \in (0,1)$. The proof thus follows from (5.1). Now, since $X[Y] \leq_{lr} X$ and $X[Y] \leq_{lr} Y$, from Shaked and Shanthikumar (2007), we get $X[Y] \leq_{lr} \min\{X, Y\}$. This completes the proof. □

From Shaked and Shanthikumar (2007) and Proposition (5.1), we have the following implications.

$$X[Y] \leq_{lr} \min\{X,Y\} \implies X[Y] \leq_{hr} \min\{X,Y\} \implies X[Y] \leq_{st} \min\{X,Y\}.$$

Proposition 5.2. Let X_1 and X_2 be two lifetime random variables with distribution functions $F_1(x)$ and $F_2(x)$ respectively. Suppose Y_1 and Y_2 are the PRHM of X_1 and X_2 respectively with the same proportionality constant. Then the following properties hold:

(i) If $X_1 \leq_{st} X_2$, then $X_1[Y_1] \leq_{st} X_2[Y_2]$.

(ii) If $X_1 \leq_{hr} X_2$, then $X_1[Y_1] \leq_{hr} X_2[Y_2]$.

(iii) If $X_1 \leq_{icx} X_2$, then $X_1[Y_1] \leq_{icx} X_2[Y_2]$.

(iv) If $X_1 \leq_{lr} X_2$, then $X_1[Y_1] \leq_{lr} X_2[Y_2]$, for $\theta > 2$.

(v) If $X_1 \leq_{rhr} X_2$, then $X_1[Y_1] \leq_{rhr} X_2[Y_2]$.

Proof. The proof of (i) is intuitive from equation (2.2).

To prove (ii), we need to show that $\frac{u\,q'(u)}{q(u)}$ is decreasing in u. We have

$$\frac{d}{du}\left(\frac{uq'(u)}{q(u)}\right) = \frac{\theta\left(u\left(\left((\theta-1)^2u-2\right)(1-u)^{\theta}-u+2\right)-\left((1-u)^{\theta}-1\right)^2\right)}{(u-1)^2\left((1-u)^{\theta}+\theta u-1\right)^2} \leq 0,$$

for all $\theta > 0$, where $q(u)$ is the distortion function defined in (3.8). Then from Navarro et al. (2013) (Theorem 2.6), result (ii) follows.
Similarly (iii) follows from Navarro et al. (2013) (Theorem 2.6), since $q(u)$ is a convex function in $(0,1)$.

Again, from Navarro et al. (2013) we have the result that if $\frac{u\,q''(u)}{q'(u)}$ is non-negative and decreasing in u, then result (iv) holds. Now,

$$\frac{d}{du}\left(\frac{u\,q''(u)}{q'(u)}\right) = -\frac{(\theta-1)(1-u)^{\theta-2}\left((1-u)^{\theta}+u(\theta-\theta u+u)-1\right)}{\left((1-u)^{\theta}+u-1\right)^2} \leq 0,$$

for all $\theta \geq 2$. Then from Navarro et al. (2013) (Theorem 2.6), result (iv) follows.
Similarly, to prove (v), we use the result from Navarro et al. (2013) that, if $\frac{(1-u)\,q'(u)}{1-q(u)}$ is increasing in u, then result (v) holds. Note that

$$\frac{d}{du}\left(\frac{(1-u)\,q'(u)}{1-q(u)}\right) = \frac{\theta\left((1-u)^{\theta}+u-1\right)}{(1-u)^{\theta}+\theta(u-1)} > 0, \quad \text{for all } \theta > 0.$$

Then from Navarro et al. (2013) (Theorem 2.6), result (v) follows. □

Proposition 5.3. Let X and Y be two lifetime random variables with distribution functions $F_X(x)$ and $F_Y(x)$ respectively. If Y is the PRHM of X, then:

(i) $X[Y] \leq_c X$ for $\theta \geq 2$.

(ii) $X \leq_c X[Y]$ for $0 < \theta \leq 2$.

Proof. Sengupta and Deshpande (1994) showed that, for two non-negative random variables X and Y with hazard rates $h_X(x)$ and $h_Y(x)$ respectively, $X \leq_c Y$ if and only if $\frac{h_X(x)}{h_Y(x)}$ is non-decreasing in x, provided $h_Y(x) \neq 0$. To prove (i), we consider the function $s_1(x)$:

$$s_1(x) = \frac{h_{X[Y]}(x)}{h_X(x)} = \frac{\theta(F_X(x)-1)\left(F_X(x)-(F_X(x))^{\theta}\right)}{F_X(x)\left(\theta F_X(x)-(F_X(x))^{\theta}-\theta+1\right)}.$$

On differentiating with respect to x, we get

$$\frac{d}{dx}(s_1(x)) = \frac{\theta\left(-\left((\theta-1)^2(F_X(x))^2-2(\theta-2)\theta F_X(x)+(\theta-1)^2\right)(F_X(x))^{\theta}+F_X(x)^{2\theta}+(F_X(x))^2\right)f_X(x)}{(F_X(x))^2\left((F_X(x))^{\theta}-\theta F_X(x)+\theta-1\right)^2},$$

which is non-negative for $\theta \geq 2$. Thus, $X[Y] \leq_c X$ for $\theta \geq 2$.

Similarly, to prove (ii) we analyze the monotonicity of the function $s_2(x)$ defined by

$$s_2(x) = \frac{h_X(x)}{h_{X[Y]}(x)} = \frac{F(x)\left(-F(x)^{\theta}+\theta F(x)-\theta-1\right)}{\theta(F(x)-1)\left(F(x)-F(x)^{\theta}\right)}.$$

On differentiating with respect to x, we get

$$\frac{d}{dx}(s_2(x)) = \frac{\left((\theta-1)(F(x)-1)((\theta-1)F(x)-\theta-1)F(x)^{\theta}-F(x)^{2\theta}+F(x)^2\right)F'(x)}{\theta(F(x)-1)^2\left(F(x)-F(x)^{\theta}\right)^2},$$

which is non-negative for $0 < \theta \leq 2$. Thus, $X \leq_c X[Y]$ for $0 < \theta \leq 2$. □

6. Applications

In this section, we propose a generalization of the Fréchet distribution using the idea of PRHRRT. The Fréchet distribution is one of the well-known extreme value model. Extreme value theory is used to estimate the probability of extreme events and to develop strategies to reduce their effects. The classical theory of extremes deals with the distributional properties of the statistics $M_n = \max(X_1, \ldots, X_n)$ and $m_n = \min(X_1, \ldots, X_n)$ of i.i.d random variables $X_1, \ldots, X_n$. Gnedenko(1943) showed that the asymptotic distribution of M_n will be one of the three types of extreme value distributions. Type-I extreme value distribution is the Gumbel distribution, Type-II is the Fréchet or inverse Weibull distribution and Type-III is the reverse Weibull distribution. We have seen in Section 4 that PRHRRT can be used for constructing new lifetime models having more flexible hazard rates. We now assume the Fréchet distribution for the baseline random variable X and study various reliability properties of $X[Y]$. The two parameter Fréchet distribution has CDF

$$F_X(x) = \mathrm{e}^{-\left(\frac{\sigma}{x}\right)^{\alpha}},\ x > 0,\ \sigma > 0,\ \alpha > 0.$$

Then the distribution function of the corresponding PRHRRT random variable $Z = X[Y]$ is obtained as

$$T_Z(x) = \frac{\theta\, \mathrm{e}^{-\left(\frac{\sigma}{x}\right)^{\alpha}} - \left(\mathrm{e}^{-\left(\frac{\sigma}{x}\right)^{\alpha}}\right)^{\theta}}{\theta - 1}, \quad x > 0,\ \sigma,\ \alpha > 0,\ \theta > 0. \tag{6.1}$$

We denote the model (6.1) as the PRHRR-F distribution. The r^{th} raw moment of PRHRR-F denoted by μ'_r is of the form

$$\mu'_r = \frac{\sigma^r \left(\theta - \theta^{\frac{r}{\alpha}}\right) \Gamma\left(1 - \frac{r}{\alpha}\right)}{\theta - 1}, \quad \alpha > r,\ r = 0, 1, 2, \ldots. \tag{6.2}$$

The moment generating function of Z is obtained as

$$M_Z(t) = \sum_{r=0}^{\infty} \mu'_r \frac{t^r}{r!}, \quad \alpha > r.$$

hazard rate of Z has the form

$$h_Z(x) = -\frac{\alpha\, \theta \left(\frac{\sigma}{x}\right)^{\alpha} \left(\left(\mathrm{e}^{-\left(\frac{\sigma}{x}\right)^{\alpha}}\right)^{\theta-1} - 1\right)}{x \mathrm{e}^{\left(\frac{\sigma}{x}\right)^{\alpha}} \left(\theta + \left(\mathrm{e}^{-\left(\frac{\sigma}{x}\right)^{\alpha}}\right)^{\theta} - 1\right) - \theta x}.$$

From Figure 2, we can observe that $h_Z(x)$ incorporates IHR, DHR and upside-down bathtub shapes for various parameter combinations.

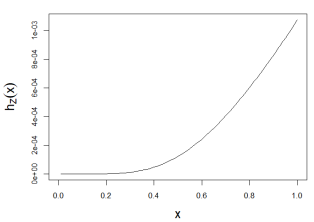
The estimation of unknown parameters of PRHRR-F (σ, α, θ) distribution has been carried out using the method of maximum likelihood. The log-likelihood function of the PRHRR-F for a given sample $x_1,\ldots,x_n$ of size n is

$$\log L(\sigma,\alpha,\theta|x_1,\ldots,x_n)=\sum_{i=1}^{n}\log\left(\frac{\alpha\ \theta\ \mathrm{e}^{-\left(\frac{\sigma}{x_i}\right)^{\alpha}}\left(\frac{\sigma}{x_i}\right)^{\alpha}\left(\left(\mathrm{e}^{-\left(\frac{\sigma}{x_i}\right)^{\alpha}}\right)^{\theta-1}-1\right)}{(1-\theta)x_i}\right).$$

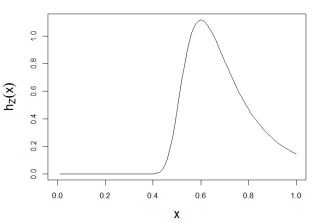
The maximum likelihood estimators (MLE) ($\hat{\lambda}$, $\hat{\alpha}$ and $\hat{\theta}$) can be obtained by solving the equations $\frac{\partial \log L}{\partial \sigma}=0$, $\frac{\partial \log L}{\partial \alpha}=0$ and $\frac{\partial \log L}{\partial \theta}=0$ simultaneously. Since it is difficult to find a solution for this non-linear system of equations analytically, we have employed the Newton-Raphson iterative method to get a solution numerically. We have $\sqrt{n}(\hat{\Theta}-\Theta)$ follows multivariate normal distribution with zero mean and variance-covariance matrix $I^{-1}(\Theta)$, where $\Theta=(\sigma,\alpha,\theta)$ and $I(\Theta)$ denotes the Fisher information matrix. From this, the two-sided $100(1-\alpha)\%$ confidence interval for the parameters can be obtained as

$$\hat{\theta}_i \pm z_{\alpha/2}\sqrt{\frac{I_{ii}^{-1}(\Theta)}{n}}, \tag{6.3}$$

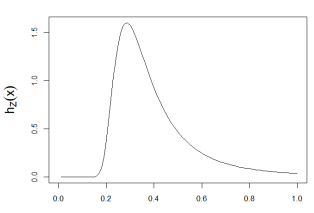
where $z_{\alpha/2}$ is the $\alpha/2^{th}$ percentile of the standard normal distribution and $I_{ii}^{-1}(\Theta)$ is the i^{th} diagonal element of $I^{-1}(\Theta)$, $i=1,\ldots,n$. When $I(\Theta)$ cannot be evaluated analytically, an efficient alternative is the observed Fisher information (OFI) introduced by Cox and Hinkley (1974).



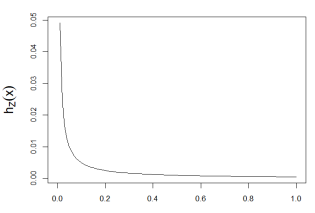
(a) $\sigma=5$, $\alpha=0.5$ and $\theta=2.5$



(b) $\sigma=0.5$, $\alpha=5$ and $\theta=3$

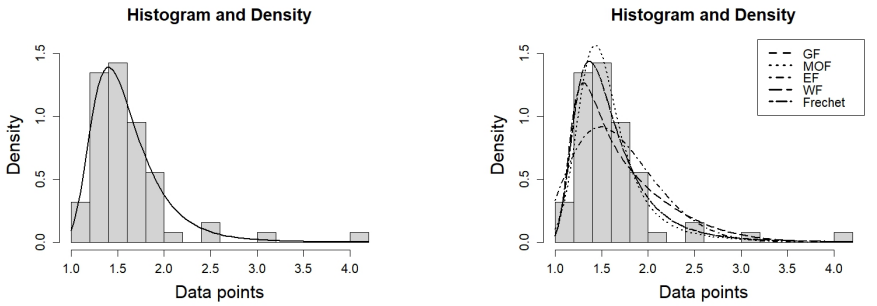


(c) $\sigma=0.2$, $\alpha=3$ and $\theta=4$



(d) $\sigma=0.01$, $\alpha=0.01$ and $\theta=1.5$

Figure 2: Plots of $h_Z(x)$ for various parameter combinations.



(a) PRHRR-F distribution **(b)** Other competing models

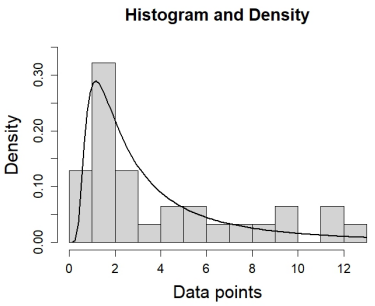
Figure 3: Histogram and Density plots for the first data set.

Table 1: Estimates, K-S statistics and p-values for the first data set.

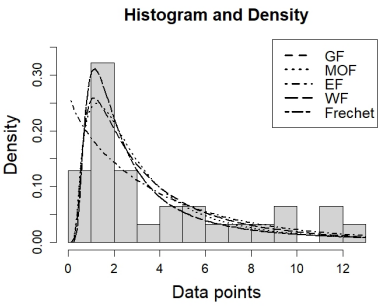
Distributions	Estimates			K-S Statistics	p-value
PRHRR-F (σ,α,θ)	$\hat{\sigma}=1.7527$	$\hat{\alpha}=3.6595$	$\hat{\theta}=0.9995$	0.0683	0.9114
GF (λ,α,β)	$\hat{\lambda}=1.6737$	$\hat{\alpha}=5.4376$	$\hat{\beta}=0.3948$	0.0772	0.8185
MOF (α,β,λ)	$\hat{\alpha}=1.4559$	$\hat{\beta}=5.2227$	$\hat{\lambda}=0.0023$	0.0813	0.7686
EF (α,β,λ)	$\hat{\alpha}=1.7936$	$\hat{\beta}=2.3223$	$\hat{\lambda}=0.3016$	0.1739	0.0389
WF (α,β,λ)	$\hat{\alpha}=1.6248$	$\hat{\beta}=5.9372$	$\hat{\lambda}=0.3750$	0.1659	0.0551
Fréchet (σ,α)		$\hat{\sigma}=1.4108$	$\hat{\alpha}=5.4377$	0.0772	0.8185

To show the applicability of the proposed model in situations other than reliability context, we next consider data that were reported in Hand et al. (1994). The data represents prices of 31 different children's wooden toys on sale in a Suffolk craft shop in April 1991. To show the efficiency of the proposed model over other competing alternatives, we carry out the K-S goodness of fit test. Maximum likelihood estimates and goodness of fit test results of the proposed model and other competing alternatives are listed in Table 2.

From Table 2 it is clear that, for the second data set, the PRHRR-F model outperforms other competing alternatives. The standard errors of $\hat{\sigma}$, $\hat{\alpha}$ and $\hat{\theta}$ are 0.0868, 0.0372 and 2.7654 respectively. The 95% confidence intervals for the model parameters σ, α and θ are $(1.9347,\ 2.2751)$, $(1.0177,\ 1.1636)$ and $(6.6134,\ 17.4540)$ respectively. Figure 4 displays the observed histogram and fitted density functions. Q-Q plot is given in Figure 5(b). These two plots ensures the adequacy of the proposed model for the data.



(a) PRHRR-F distribution

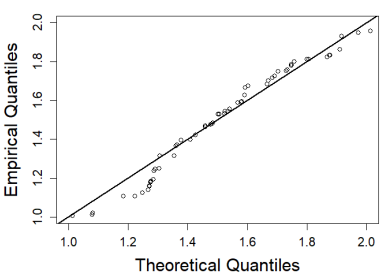


(b) Other competing models

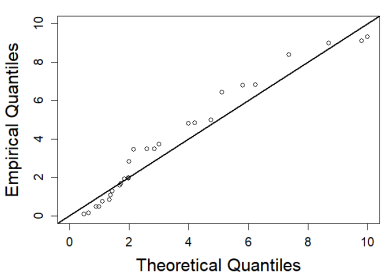
Figure 4: Histogram and Density plots for the second data set.

Table 2: Estimates, K-S statistics and p-values for the second data set.

Distributions	Estimates			K-S Statistics	p-value
PRHRR-F (σ,α,θ)	$\hat{\sigma}=2.1045$	$\hat{\alpha}=1.0906$	$\hat{\theta}=12.0434$	0.0821	0.9851
GF (λ,α,β)	$\hat{\lambda}=1.2321$	$\hat{\alpha}=1.2147$	$\hat{\beta}=1.6709$	0.0980	0.9271
MOF (α,β,λ)	$\hat{\alpha}=1.6728$	$\hat{\beta}=0.8776$	$\hat{\lambda}=6.4507$	0.1014	0.9074
EF (α,β,λ)	$\hat{\alpha}=2.6055\times10^{-14}$	$\hat{\beta}=0.9559$	$\hat{\lambda}=3.9062$	0.1392	0.5848
WF (α,β,λ)	$\hat{\alpha}=2.7451$	$\hat{\beta}=1.0389$	$\hat{\lambda}=0.7502$	0.1000	0.9156
Fréchet (σ,α)		$\hat{\sigma}=1.8802$	$\hat{\alpha}=1.2148$	0.0979	0.9271



(a) First data set



(b) Second data set

Figure 5: Q-Q plots.

7. Conclusions

In this paper, we have presented the proportional reversed hazards in the reversed relevation transform as a special case of the reversed relevation transform. Its reliability properties and results based on entropy measures were discussed in detail. The ageing and stochastic ordering properties of the model were derived. Finally, we introduced the

PRHRR-F $(\sigma,\ \alpha,\ \theta)$ model, studied its important characteristics and illustrated its practical applicability with the help of two real-life data sets.

Acknowledgements

The authors would like to thank the editors and the referees for their constructive comments.

References

Barlow, R. E., Proschan, F., (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. Holt, Rinehart and Winston, New York.

Belzunce, F., Riquelme, C. M. and Mulero, J., (2016). *An Introduction to Stochastic Orders*. Academic Press, London.

Bennett, S., (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2(2), pp. 273–277.

Block, H. W., Savits, T. H. and Singh, H., (1998). The reversed hazard rate function. *Probability in the Engineering and Informational Sciences*, 12(1), pp. 69–90.

Breneman, J. E., Sahay, C. and Lewis, E. E., (2022). *Introduction to Reliability Engineering*. John Wiley & Sons, Hoboken, NJ, USA.

Calì, C., Longobardi M. and Ahmadi, J., (2017). Some properties of cumulative Tsallis entropy. *Physica A: Statistical Mechanics and its Applications*, 486(15), pp. 1012–1021.

Chandra, N. K., Roy, D., (2001). Some results on reversed hazard rate. *Probability in the Engineering and Informational Sciences*, 15(1), pp. 95–102.

Chechile, R. A., (2011). Properties of reverse hazard functions. *Journal of Mathematical Psychology*, 55(3), pp. 203–222.

Cox, D. R., Hinkley, D. V., (1974). *Theoretical Statistics*. Chapman and Hall/CRC, London.

Denneberg, D., (1990). Premium calculation: Why standard deviation should be replaced by absolute deviation. *ASTIN Bulletin: The Journal of the IAA*, 20(2), pp. 181–190.

Di Crescenzo, A., (2000). Some results on the proportional reversed hazards model. *Statistics & Probability Letters*, 50(4), pp. 313–321.

Di Crescenzo, A., Kayal, S. and Toomaj, A., (2018). A past inaccuracy measure based on the reversed relevation transform. *Metrika*, 82(5), pp. 607–631.

Di Crescenzo, A., Longobardi, M., (2009). On cumulative entropies. *Journal of Statistical Planning and Inference*, 139(12), pp. 4072–4087.

Di Crescenzo, A., Toomaj, A., (2015). Extension of the past lifetime and its connection to the cumulative entropy. *Journal of Applied Probability*, 52(4), pp. 1156–1174.

Di Crescenzo, A., Toomaj, A., (2017). Further results on the generalized cumulative entropy. *Kybernetika*, 53(5), pp. 959–982.

Finkelstein, M., (2002). On the reversed hazard rate. *Reliability Engineering & System Safety*, 78(1), pp. 71–75.

Glaser, R. E., (1980). Bathtub and related failure rate characterizations. *Journal of the American Statistical Association*, 75(371), pp. 667–672.

Gnedenko, B., (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3), pp. 423–453.

Gupta, R. C., Gupta, P. L. and Gupta, R. D., (1998). Modeling failure time data by Lehman alternatives. *Communications in Statistics - Theory and Methods*, 27(4), pp. 887–904.

Gupta, R. C. , Wu, H., (2001). Analyzing survival data by proportional reversed hazard model. *International Journal of Reliability and Applications*, 2(1), pp. 1–26.

Gupta, R. D., Kundu, D., (1999). Theory & methods: Generalized exponential distributions. *Australian & New Zealand Journal of Statistics*, 41(2), pp. 173–188.

Gupta, R. D., Kundu, D., (2001). Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 43(1), pp. 117–130.

Gupta, R. D., Kundu, D., (2002). Generalized exponential distributions: Statistical inferences. *Journal of Statistical Theory and Applications*, 1(1), pp. 101–118.

Gupta, R. D., Kundu, D., (2007). Generalized exponential distribution: Existing results and some recent developments. *Journal of Statistical Planning and Inference*, 137(11), pp. 3537–3547.

Gupta, R. D., Nanda, A. K., (2001). Some results on reversed hazard rate ordering. *Communications in Statistics - Theory and Methods*, 30(11), pp. 2447–2457.

Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E., (1994). *Handbook of Small Data Sets*. Boca Raton, FL: Chapman & Hall/ CRC Press, London.

Kalbfleisch, J. D., Prentice, R. L., (2002). *The Statistical Analysis of Failure Time Data, Second Edition*. John Wiley & Sons, New York.

Kayal, S., (2016). On generalized cumulative entropies. *Probability in the Engineering and Informational Sciences*, 30(4), pp. 640–662.

Kerridge, D. F., (1961). Inaccuracy and inference. *Journal of the Royal Statistical Society Series B*, pp. 184–94.

Khorashadizadeh, Rezaei Roknabadi, M. A. H. and Mohtashami Borzadaran, G. R., (2013). Characterization of life distributions using log-odds rate in discrete aging. *Communications in Statistics - Theory and Methods*, 42(1), pp. 76–87.

Kochar, S. C., (2022). *Stochastic Comparisons with Applications: In Order Statistics and Spacings*. Springer Nature, Switzerland.

Krakowski, M., (1973). The relevation transform and a generalization of the gamma distribution function. *Revue Française d'automatique, Informatique, Recherche Opérationnelle. Recherche Opérationnelle*, 7(V2), pp. 107–120.

Kundu, D., Gupta, R. D., (2004). Characterizations of the proportional (reversed) hazard model. *Communications in Statistics – Theory and Methods*, 33(12), pp. 3095–3102.

Lai, C. D., Xie, M., (2006). *Stochastic Ageing and Dependence for Reliability*. Springer Science & Business Media, London.

Lawless, J. F., (2003). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York.

Lehmann, E. L., (1953). The power of rank tests. *The Annals of Mathematical Statistics*, 24(1), pp. 23– 43.

Mahmoud, M. A., Alam, F. M. A., (2010). The generalized linear exponential distribution. *Statistics & Probability Letters*, 80(11-12), pp. 1005–1014.

Mudholkar, G. S., Hutson, A. D., (1996). The exponentiated Weibull family: some properties and a flood data application. *Communications in Statistics – Theory and Methods*, 25(12), pp. 3059–3083.

Mudholkar, G. S., Srivastava, D. K., (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42(2), pp. 299–302.

Mudholkar, G. S., Srivastava, D. K. and Freimer, M., (1995). The exponentiated Weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, 37(4), pp. 436–445.

Nath, P., (1968). Inaccuracy and coding theory. *Metrika*, 13(1), pp. 123–135.

Navarro, J., (2022). *Introduction to System Reliability Theory*. Springer, Berlin.

Navarro, J., Águila, Y. , Sordo, M. A. and Suárez-Llorens, A., (2013). Stochastic ordering properties for systems with dependent identically distributed components. *Applied Stochastic Models in Business and Industry*, 29(3), pp. 264–278.

Navarro, J., Águila,Y., Sordo, M. A. and Suárez-Llorens, A., (2014). Preservation of reliability classes under the formation of coherent systems. *Applied Stochastic Models in Business and Industry*, 30(4), pp. 444–454.

Navarro, J., Del Águila, Y., Sordo, M. A. and Suárez-Llorens, A., (2016). Preservation of stochastic orders under the formation of generalized distorted distributions. applications to coherent systems. *Methodology and Computing in Applied Probability*, 18(2), pp. 529–545.

Navarro, J., del Águila, Y., Sordo, M. A. and Suárez-Llorens, A., (2013). Stochastic ordering properties for systems with dependent identically distributed components. *Applied Stochastic Models in Business and Industry*, 29(3), pp. 264–278.

Navarro, J., Ruiz, J. M. and Del Aguila, Y., (2008). Characterizations and ordering properties based on log-odds functions. *Statistics*, 42(4), pp. 313–328.

Popović, B. V., Genç, A. I. and Domma, F., (2022). Generalized proportional reversed hazard rate distributions with application in medicine. *Statistical Methods & Applications*, 31(3), pp. 459–480.

Psarrakos, G., Navarro, J., (2013). Generalized cumulative residual entropy and record values. *Metrika*, 76(5), pp. 623–640.

Sarhan, A. M., Kundu, D., (2009). Generalized linear failure rate distribution. *Communications in Statistics – Theory and Methods*, 38(5), pp. 642–660.

Sengupta, D., Deshpande, J. V., (1994). Some results on the relative ageing of two life distributions. *Journal of Applied Probability*, 31(4), pp. 991–1003.

Shaked, M., Shanthikumar, J. G., (2007). *Stochastic Orders*. Springer Science & Business Media, New York.

Shannon, C. E., (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), pp. 379–423.

Shojaee, O., Babanezhad, M., (2023). On some stochastic comparisons of arithmetic and geometric mixture models. *Metrika*, 86(5), pp. 499–515.

Smith, R. L., Naylor, J., (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 36(3), pp. 358–369.

Sordo, M. A., Suárez-Llorens, A., (2011). Stochastic comparisons of distorted variability measures. *Insurance: Mathematics and Economics*, 49(1), pp. 11–17.

Sordo, M. A., Suárez-Llorens, A. and Bello, A. J., (2015). Comparison of conditional distributions in portfolios of dependent risks. *Insurance: Mathematics and Economics*, 61, pp. 62–69.

Taneja, H., Kumar, V. and Srivastava, R., (2009). A dynamic measure of inaccuracy between two residual lifetime distributions. *International Mathematical Forum*, 4(25), pp. 1213–1220.

Wang, S., (1996). Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, 26(1), pp. 71–92.

Zimmer, W. J., Wang, Y., and Pathak, P. K., (1998). Log-odds rate and monotone log-odds rate distributions. *Journal of Quality Technology*, 30(4), pp. 376–385.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 205–220, https://doi.org/10.59139/stattrans-2025-011
Received – 05.09.2023; accepted – 23.09.2024

Analytical modelling for COVID-19 data (fatality): a case study of Nigeria for the period of February 2020 – April 2022

Emmanuel Torsen[1], **Umar Muhammad Modibbo**[2], **Mohammed Mijinyawa**[3], **Lema Logamou Seknewna**[4], **Irfan Ali**[5]

Abstract

One of the most significant disruptive events of the 21st century was the COVID-19 epidemic, which was first detected in China in 2019 and quickly spread around the world. While waiting for the development of the vaccine, governments used a variety of strategies to counteract the effects of the pandemic: from simple personal hygiene advice to the introduction of strict lockdowns. In this paper, the confirmed cases of COVID-19 fatalities (count data and having zero inflation) due to COVID-19 in Nigeria modeled using univariate time series models. To describe the attributes of COVID-19 fatalities in Nigeria with zero inflation, the autoregressive integrated moving average (ARIMA), zero-inflated poison autoregressive (ZIPAR), and zero-inflated negative binomial autoregressive (ZINBAR) models were used. Our findings indicate that $ZINBAR(1)$ having the lowest root mean square error (RMSE), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC) outperforms the other two models: Hence, the $ZINBAR(1)$ performs better than the $ZIPAR(1)$ (this is in aggrement with the work of Tawiah et al. (2021)) and the $ARIMA(0,1,1)$. This demonstrates and emphasised the fact that for count time series data, count time series models should be used. Hence, the ZINBAR(1) can be use to predict and forecast COVID-19 in Nigeria.

Key words: ARIMA, Covid-19, Nigeria, Modeling, ZINBAR, ZINPAR.

1. Introduction

The novel coronavirus 2019 (COVID-19) has been a deadly killer since 2019 and throughout the year 2020, until researchers came up with some remedy to reduce the spread.

[1]Department of Statistics, Modibbo Adama University, PMB 2076, Yola, Nigeria. E-mail: torsen@mau.edu.ng. ORCID: https://orcid.org/0000-0002-6517-9668.

[2]Department of Operations Research, Modibbo Adama University, P.M.B. 2076, Yola, Nigeria. E-mail: umarmodibbo@mau.edu.ng. ORCID: https://orcid.org/0000-0002-9242-4948.

[3]Department of Operations Research, Modibbo Adama University, P.M.B. 2076, Yola, Nigeria. E-mail: m.mijinyawa@mau.edu.ng. ORCID: https://orcid.org/0000-0002-7248-3561.

[4]Department of Science and Technology, University of Mayotte (CUFR), France. E-mail: seknewna@gmail.com ORCID: https://orcid.org/0000-0002-2233-463X.

[5]Department of Statistics & Operations Research, Aligarh Muslim University, Aligarh, 202002, India. E-mail: irfii.st@amu.ac.in. ORCID: https://orcid.org/0000-0002-1790-5450.

In December 2019, Wuhan, Hubei Province, China, experienced the outbreak of COVID-19, a brand-new coronavirus illness (Li et al., 2020). China is resolved to effectively stop the spread of the disease, and on January 20, 2020, the National Health Commission of the People's Republic of China unveiled the most comprehensive prevention and control measures against pneumonia (Team, 2020; Kucharski et al., 2020).

For public health, pandemic preparedness, and healthcare systems, this virus poses a significant problem. The extremely contagious SARS-CoV-2 coronavirus causes severe acute respiratory illness (Organization et al., 2020). By coming into direct touch with contaminated surfaces and breathing in respiratory droplets from sick people, COVID-19 is passed from one person to another (Bai et al., 2020). Right now, COVID-19 cannot be prevented or treated by a vaccine or antiviral medication that has received approval (Tang et al., 2020).

Governments have been putting in place various control measures to effectively stop the spread of COVID-19, including strict, mandatory lockdowns and encouraging (and in some cases strictly enforcing) other measures like people keeping a minimum distance between themselves (social distancing), avoiding crowded events, imposing a maximum number of people in any gathering (religious and social), and the use of face masks while in public (Dunford et al., 2020).

There are numerous epidemic models that use mathematics to describe how infectious illnesses spread. Estimating the spread of illnesses and the number of affected people is the primary function of modeling, which enables prudent government to develop a workable strategy. Statistical models enable the evaluation of a number of "what-if" scenarios, which can provide significant insight for decision-makers in public health.

A few mathematical models that attempt to capture the dynamics of COVID-19's and other diseases evolution can be found in the literature (Roosa et al., 2020). In an effort to create and evaluate shortterm projections of the total number of reported cases, these were validated with outbreaks of diseases other than COVID-19.

Recently, a multi-criteria decision-analysis techniques based Fuzzy TOPSIS have been employed in assessing the disruptions caused by the COVID-19 pandemic in different sectors of the world economy. The study combined two MCDA tools (best-worst and fuzzy-TOPSIS) and considered Supply Chain disruptions in particular, ranking the most affected industries (Ali et al., 2023).

Across the globe and indeed in Africa, there have been several statistical models of COVID-19 data aimed at studying, learning and understanding the dynamics of the pandemic. The work of Sam (2020) which compared COVID-19 data from the African region to other regions of the world. Shoko and Njuho (2023) used ARIMA model for predicting the spread of COVID-19 in Southern Africa (South Africa, Zambia and Namibia). A short-term ARIMA model for predicting mortality due to COVID-19 was developed by Chaurasia and Pal (2020). They used the model (ARIMA) to forecast mortality rate. Many other authors studied COVID-19 using different forms of ARIMA models. These includes but are not limited to Alabdulrazzaq et al. (2021), who used the Kuwait COVID-19 data to test the accuracy of the ARIMA model in predictions. The result confirmed the applicability of the ARIMA model. Ribeiro et al. (2020) applied machine learning and ARIMA to forecast

COVID-19 cases in the Brazilian context; Yang et al. (2020) employed a similar ARIMA model for COVID-19 cases prediction in Italy. Lukman et al. (2020), Khan and Gupta (2020), Poleneni et al. (2021), Abdelaziz et al. (2020), Nguyen et al. (2020), Somyanonthanakul et al. (2022), Zhihao et al. (2021), Malki et al. (2021), ArunKumar et al. (2021) and Sah et al. (2022) demonstrated the usage of ARIMA in COVID-19 with different case studies. However, it is observed that ARIMA cannot fit COVID-19 data correctly, because they are count data. Therefore, it gives the basis for further investigating the most suitable model for the pandemic.

In Nigeria, Busari and Samson (2022) studied the dynamics of the COVID-19 pandemic using ARIMA model and other machine learning models, and the results showed that Fine Tree, one of the Machine Learning models, outpreformed ARIMA model. Ibrahim and Oladipo (2020) used only two months dataset to forecast Nigerian COVID-19 spread using ARIMA (1,1,0) model. Similarly, Samson et al. (2020) applied ARIMA models on the COVID-19 confirmed cases and selected the ARIMA (2,1,0) as the best fit model for prediction. Olarenwaju and Harrison (2020) applied ARIMA and Artificial Neural Networks to COVID-19 data of some selected states in Nigeria for prediction. Adesina et al. (2020) demonstrated that ARFIMA model outperformed ARIMA model for modeling and prediction of confirmed cases of COVID-19 in Nigeria. Lukman et al. (2020) evaluated COVID-19 prevalence for 4 countries (South Africa, Nigeria, Ghana and Egypt) in Africa using ARIMA models. The $ARIMA(0,2,3)$, $ARIMA(0,1,1)$, $ARIMA(3,1,0)$ and $ARIMA(0,1,2)$ models were selected as the optimal models for these countries, respectively. Didi et al. (2021) modeled daily COVID-19 confirmed and fatality cases in Nigeria considering ARIMA models. Ortese et al. (2021) explored the ARIMA family of models to assess the infection rate of COVID-19 in Nigeria, and selected ARIMA (0,1,1) as the best model. Aronu et al. (2021) Argawu (2021); Agboola et al. (2021); Suleiman and Sani (2021); Li et al. (2022); Nwafor et al. (2022); Adams and Somto (2022); Inegbedion (2023); Oduntan and Ajayi (2023) displayed the application of ARIMA models for the same COVID-19 data in Nigeria considering different cases. However, the present study used count time series models (ZIBAR and ZIPAR) in comparison with ARIMA model as in the literature. The study showed that for count time series data (Nigeria COVID-19 data), count time series models should be used.

Odekina et al. (2022) modeled the 3rd wave of COVID-19 in Nigeria using Vector Autoregressive (VAR) model, where a steady rise in fatality cases was observed but a small decrease in the number new cases was recorded.

The problem with most of the reviewed literature, the data generated at the instance of COVID-19 pandemic (be it confirmed cases, recovered cases or fatality) are count data. Consequently, such family of Time Series models (such as ARIMA) would not be able to adequately model such data. This is in agreement with Busari and Samson (2022).

The aim of this study is to create a Time Series model that captures the features of COVID-19 data in Nigeria, so that, such a model can be used for prediction and forecasting. Hence, we specifically considered, the number of fatalities due to COVID-19 in Nigeria, which are count data and have zero inflation.

2. Materials and Methods

2.1. Materials

In this paper, the whole country (Nigeria) was considered as a study area. Figure 1 shows the map of Nigeria and the COVID-19 outbreak status across the States of the federation. States (Lagos, Oyo, Rivers, Kaduna) and the Federal Capital Territory (FCT) with dark green coloration had above 10,000 confirmed cases of COVID-19, while Kogi State had between 1 to 100 confirmed cases of COVID-19 during the period under review.

The Gender-Age distribution is presented in Figure 2, the number of confirmed cases and fatalities (death) among males was higher as compared to females in the country. More fatalities were recorded among both males and females of group 45 - 49 and older.

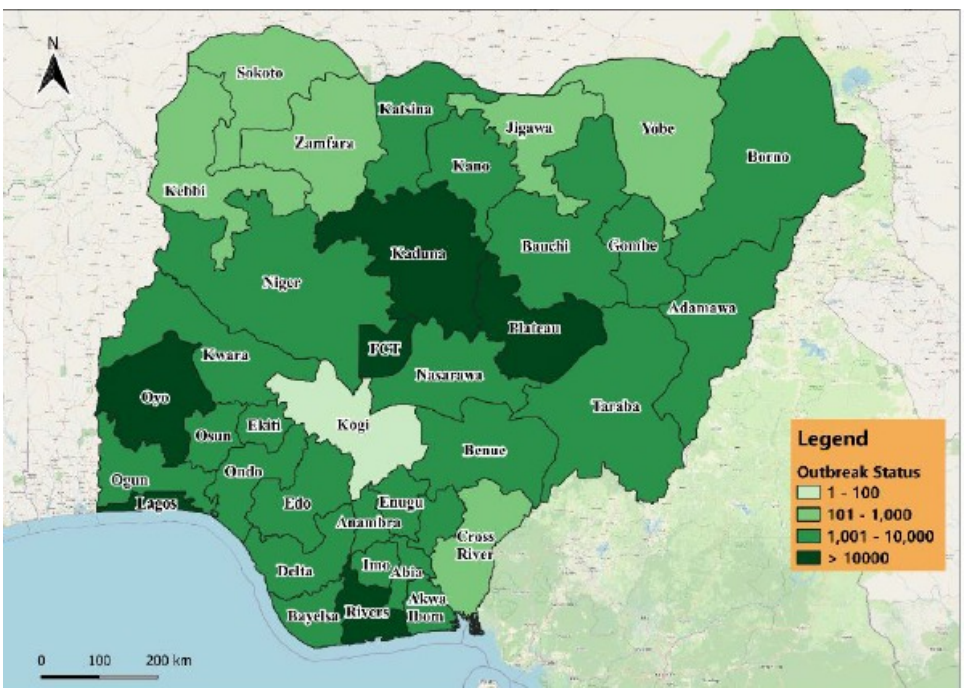


Figure 1: Distribution of cumulative cases of COVID-19 across the six Geo-political zones of the Federation as of epi weeks 17-18, 2022. Source: NCDC COVID-19 situation report

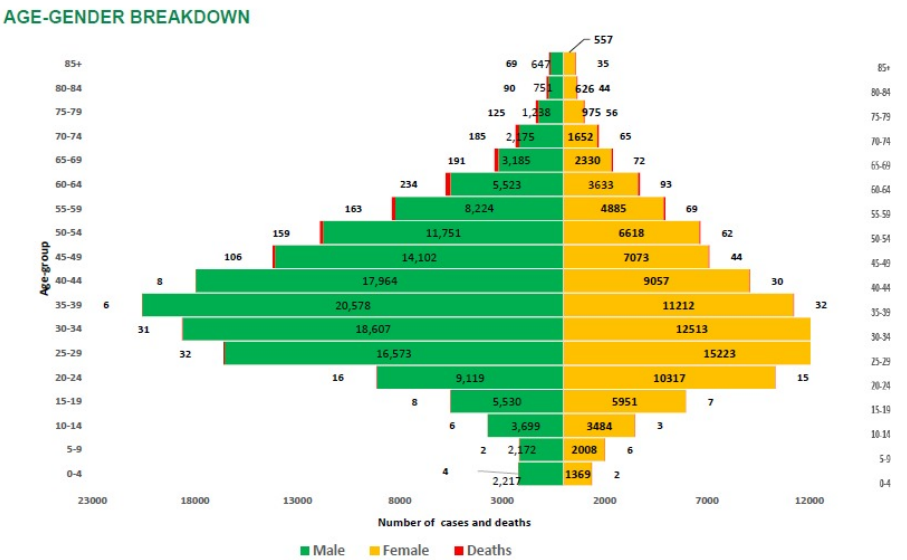


Figure 2: Gender and age distribution of confirmed cases and fatalities of COVID-19 (week 10, 2020 - week 18, 2022). Source: NCDC COVID-19 situation report

2.2. Methods

The fatalities recorded in Nigeria as a result of COVID-19 infections, as observed are discrete count time series data, having zero inflation. The Zero-Inflated Poisson Autoregressive (ZINPAR) and Zero- Inflated Negative Binomial Autoregressive (ZINBAR) models of Yang (2012), and later used by Tawiah et al. (2021) for modeling COVID-19 deaths in Ghana was used as reviewed and discussed in the next section.

Confirmed cases of COVID-19 infections and COVID-19 fatalities in Nigeria were the two series used for the period under review, from 27th February, 2020 to 3rd April, 2022.

2.2.1 Time Series Models

There are several univariate multivariate time series models and their applications in the literature, for example; Abiodun *et al.* (2019), Kumar and Susan (2020), Ogbuagada *et al.* (2022), Yang (2012), Yang *et al.* (2013), Tawiah *et al.* (2021), Chyon *et al.* (2022), Ogbuagada *et al.* (2022), amongst many others.

Let the two series $\{C_t\}_{t=1}^{T}$ and $\{F_t\}_{t=1}^{T}$ denote confirmed cases and fatalities of COVID-19 in Nigeria, respectively. The univariate time series models adapted in this paper are reviewed below.

Autoregressive Integrated Moving Average (ARIMA) Model

The $ARIMA(p,d,q)$ model is a combination of two time series models: Autoregressive $(AR(p))$ model with order p and the Moving Average $(MA(q))$ model of order q, d is the

order of differencing, and I represents integration (Kumar and Susan, 2020).

$$c_t = \alpha + \sum_{i=1}^{p} \phi_i c_{t-i} + v_t + \sum_{j=1}^{q} \theta_j v_{t-j} \tag{1}$$

where $c_{t-1}, c_{t-2}, \ldots, c_{t-p}$ are the past values of confirmed cases of COVID-19 in Nigeria and $v_t, v_{t-1}, v_{t-2}, \ldots, v_{t-q}$ are error terms, and $v_t \sim N(0, \sigma^2)$.

Zero-Inflated Poisson Autoregressive (ZINPAR) Model

Let $\{F_t\}_{t=1}^{T}$ denote COVID-19 fatalities in Nigeria as mentioned earlier. $\{F_t\}_{t=1}^{T}$ is discrete count data and $F_t \sim ZIP(\lambda_t, \theta_t)$ where λ_t is the intensity parameter of the Poisson distribution and θ_t is the Zero-Inflation (ZI) parameter.

The ZIPAR model is given by

$$\mathbb{P}(f_t|f_{t-1} = j) = \begin{cases} \theta_t + (1-\theta_t)exp(-\lambda_t), & if\ \ j = 0 \\ (1-\theta_t)\frac{\lambda_t^{\,j} exp(-\lambda_t)}{j!}, & if\ \ j = 1,2,3,\ldots \end{cases} \tag{2}$$

the λ_t parameter is given as

$$ln\lambda_t = X_{t-1}^{T}\beta \tag{3}$$

and θ_t, the ZI is defined as

$$ln\Big(\frac{\theta_t}{1-\theta_t}\Big) = Z_{t-1}^{T}\gamma \tag{4}$$

where $\beta = (\beta_1, ..., \beta_p)^T$ and $\gamma = (\gamma_1, ..., \gamma_p)^T$ are the regression coefficients for the log-linear part in equation (3) and logistic part in equation (4) respectively.

The conditional mean and variance ZIPAR are given by

$$\mathbb{E}(F_t|f_{t-1}) = \lambda_t(1-\theta_t) \tag{5}$$

and

$$Var(F_t|f_{t-1}) = (1-\theta_t)(1+\lambda_t\theta_t)\lambda_t \tag{6}$$

Zero-Inflated Negative Binomial Autoregressive (ZINBAR) Model

Over-dispersion, which is a situation where by the variance is greater than the mean ($Var(F_t|f_{t-1}) > \mathbb{E}(F_t|f_{t-1})$) can be easily taken care of by ZIPAR and ZINBAR (Tawiah et al., 2021).

The pdf of ZINBAR is given by

$$h\mathbb{P}(f_t|f_{t-1}=j)=\begin{cases}\theta_t+(1-\theta_t)\left(\frac{b_t}{b_t+\lambda_t}\right)^{b_t}, & if\ j=0\\ (1-\theta_t)\frac{\Gamma(b_t+f_t)}{\Gamma(b_t)f_t!}\left(\frac{b_t}{b_t+\lambda_t}\right)^{b_t}\left(\frac{\lambda_t}{b_t+\lambda_t}\right)^{f_t}, & if\ j=1,2,3,\ldots\end{cases} \tag{7}$$

where λ_t and θ_t are as given in equations (3) and (4) respectively.

$$lnb_t = S_{t-1}^T\alpha \tag{8}$$

The dispersion parameter is as defined in equation (8), where $\alpha=(\alpha_1,\ldots,\alpha_p)^T$ are the regression coefficients and $S_{t-1}=(S_{t-1,1},\ldots,S_{t-p,p})^T$ is a vector of past input variables.

Similarly, the expectation and variance of ZINBAR are not different from that of ZIPAR, given in equations (5) and (6) respectively.

2.3. Model Selection Criteria

In decision-making analysis, the same problem can be addressed using a variety of models statistically. However, probability distributions can be applied to the data for best-fitting, and the suitable distribution can be chosen. The optimum model for a specific data set can be determined for this purpose using the Akaike's information criterion (AIC) and Bayesian information criterion (BIC) techniques. According to [**?**], any statistical model that conforms to a specific statistical distribution can have the quality of its fit evaluated using the AIC, whereas the BIC is a criterion for selecting a model from a finite set, with the lowest BIC model being chosen as the top choice. Additionally, the processes might aid in confirming the outcome. The AIC and BIC are calculated as

$$AIC = 2k - 2\log lik \tag{9}$$

and

$$BIC = k\log n - 2\log lik \tag{10}$$

where $\log lik$ is the maximized value of the likelihood function, n is the sample size and k the number of parameters in the model under consideration. The distribution that has the lowest AIC and BIC values is regarded as having the best fit to the given data set.

2.3.1 Measure of model performance

Measures such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), amongst others, are used to assess the performance of models. The RMSE was used in this paper, defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i-\hat{f}_i)^2} \tag{11}$$

where f_i is the actual value of COVID-19 fatalities in Nigeria and $\hat{f}_i$ represents the predicted values of COVID-19 fatalities.

3. Analysis, Results and Discussion

This section discusses the results related to confirmed and fatality cases due to COVID-19, emphasis is placed on fatalities (which is the focused of this study).

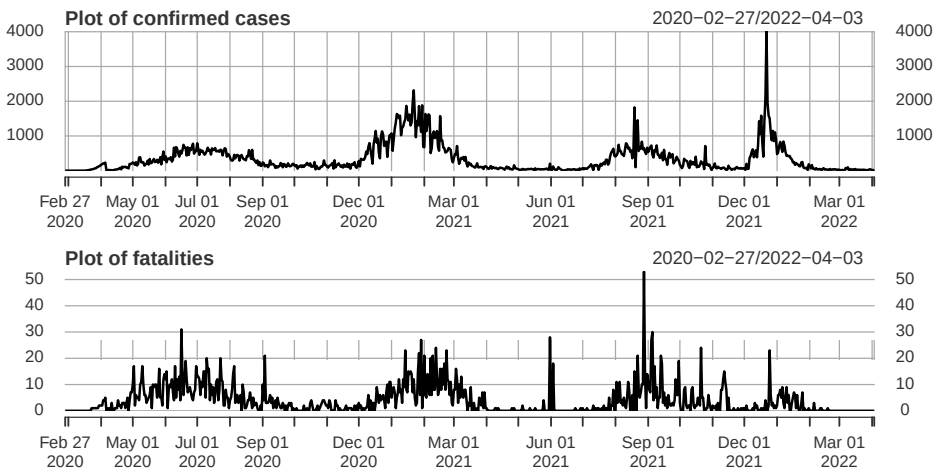


Figure 3: Plots of confirmed cases and fatalities due to COVID-19 in Nigeria

The trend of confirmed cases of COVID-19 and deaths as a result of COVID-19 in Nigeria as can be observed closely in Figures 3, with upwards and downwards movements in both cases (plots of confirmed cases and fatalities). Surges in fatalities can be seen as infection rate increases, a fall in infection rate is closely followed by a decline deaths as a result of COVID-19 in Nigeria in the period under study (February 27th, 2020, when the first COVID-19 case was confirmed in Nigeria (NCDC, 2020) to April 3rd, 2022).

Augmented Dickey-Fuller (ADF) Test of stationarity for the two series (confirmed cases of COVID-19 and fatalities) gave *P-values* of 0.1064 and 0.05515 respectively, which were greater than the 0.05 level of significance. This implied that the two series were not stationary respectively. Data (the two times series) were differenced once before fitting the ARIMA model. Plots of the two differenced series are illustrated in Figures 4 and 5, which clearly shows that the two series are stationary. This is also confirmed by the ADF test, that the two times series are stationary (*P-values* less than 0.05). That is, the P-values are 0.01 respectively for both the series.

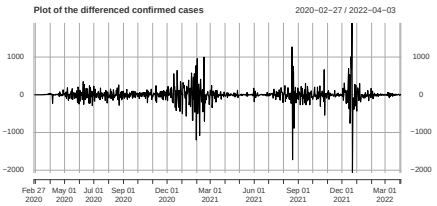


Figure 4: Differenced series of confirmed cases COVID-19

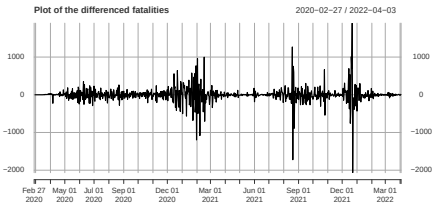


Figure 5: Differenced series of fatality cases due to COVID-19

3.1. Deaths Due to COVID-19 in Nigeria

The histogram in Figure 6 shows that there is a higher proportion of zeros (i.e., no fatalities) on such days in the period under study, that is, from 27^{th} February, 2020 to 3^{rd} April, 2022.

This is an obvious indication that there is zero inflation in the data. Although the number of new cases is still increasing and consequently the number of recorded deaths, it is important to investigate the daily death reports in order to determine if the large percentage of COVID-19 patients in Nigeria have become resistant to the disease outbreak or have reacted favorably to the treatment regimens administered to them at COVID-19 treatment facilities (Tawiah et al., 2021).

Three different models were used in order to have a model that best describe the attributes of fatalities as result of COVID-19 in Nigeria: Autoregressive Integrated Moving Average (ARIMA), Zero-inflated Poison Autoregressive (ZIPAR) model, and Zero-inflated Negative Binomial Autoregressive (ZINBAR) model.

Figures 7, 8, and 9 are the plots of the actual fatalities due to COVID-19 in Nigeria and the predicted fatalities using the ARIMA, ZIPAR, and ZINBAR respectively. Looking at the plots closely, one can observe that the three models considered performs well, but $ZINBAR(1)$ (in Figure 9) with the lowest *RMSE* (**2.920087**) as shown in Table 1, *AIC* (**3451.719**), and *BIC* (**3475.568**) in Table 2 outperformed the other two models. Hence, the Zero Inflated Negative Binomial Autoregressive model performs better than the Zero Inflated Poisson Autoregressive model and the Autoregressive Integrated Moving Average, that is, the $ARIMA(0,1,1)$ model.

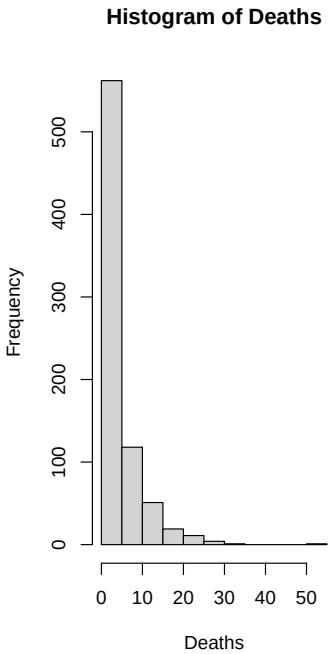


Figure 6: COVID-19 fatalities in Nigeria

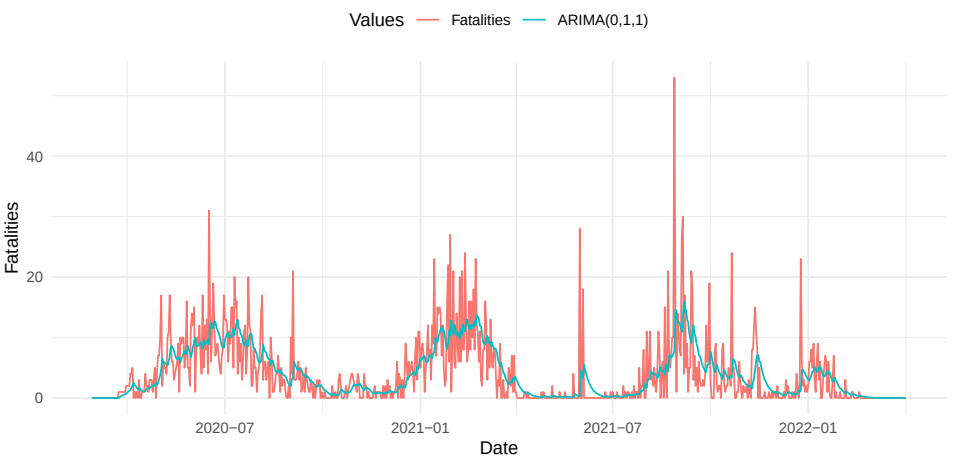


Figure 7: Actual versus predicted of fatalities due to COVID-19 in Nigeria

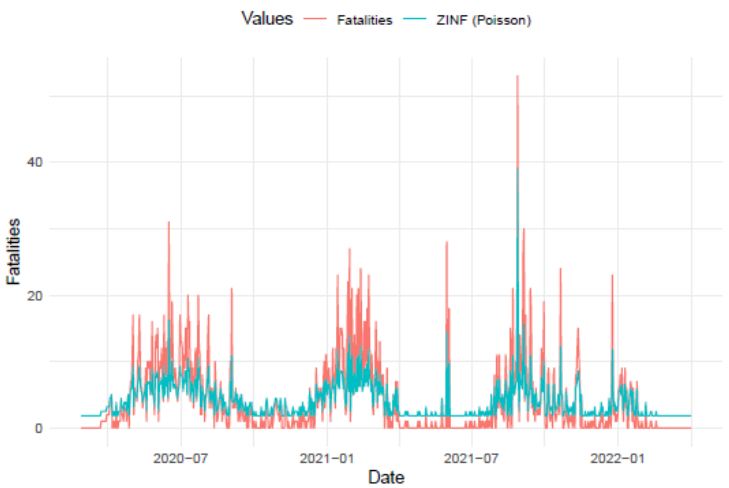


Figure 8: Actual versus predicted of fatalities due to COVID-19 in Nigeria using zero-inflated Poisson model

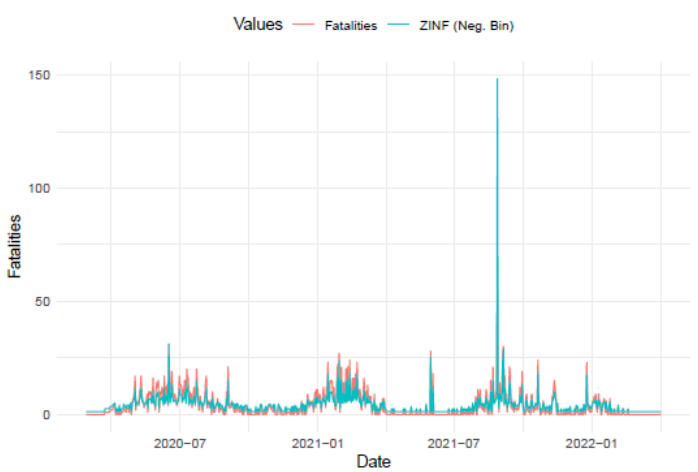


Figure 9: Actual versus predicted number of fatalities due to COVID-19 in Nigeria using zero-inflated negative binomial model

Table 1: The RMSE of the models for COVID-19 deaths cases in Nigeria

Model	RMSE	Rank
ZIPAR	3.98998	2^{nd}
ZINBAR	**2.920087**	1^{st}
ARIMA(0,1,1)	4.418348	3^{rd}

Table 2: The AIC and BIC of the models for COVID-19 deaths cases in Nigeria

Model	AIC	BIC	Rank
ZIPAR	4488.022	4507.101	2^{nd}
ZINBAR	**3451.719**	**3475.568**	1^{st}
ARIMA(0,1,1)	4954.88	4964.42	3^{rd}

4. Conclusion

COVID-19 halted many socio-economic activities in Nigeria, just like it did in other parts of the world. The two series, namely confirmed cases and COVID-19 fatalities (which have many zeros) were modeled, infection cases rose dramatically with death count. But gradually, fatalities were reduced even in the midst of continual daily infections, which was also result of some stringent COVID-19 safety measures adopted by the country through the Nigerian Center for Disease Control (NCDC).

The method used in this study is applicable to COVID-19 data as evident from the literature. In fact, most of the reviewed works employed existing time series models, most of which are of the ARIMA family. However, in this work it is argued that COVID-19 death cases are count data, hence, the use of ARIMA models is inappropriate.

We have demonstrated in this study that using Autoregressive Integrated Moving Average (for instance, in the context of Nigeria *ARIMA*) is not adequate. This was shown by fitting three univariate Time Series models for the over-dispersed zero inflated COVID-19 fatalities series, where the $ZINBAR(1)$ performed better than the other two models, which can be used for prediction and forecasting purposes.

Acknowledgments

The Authors are thankful to the Nigerian Center for Disease Control (NCDC) for making the data for this research available and the reviewers for making valuable comments and criticism that brought the work up to speed.

References

Abdelaziz, M., Ahmed, A., Riad, A., Abderrezak, G. and Djida, A. A., (2020). Forecasting daily confirmed COVID-19 cases in Algeria using ARIMA models. *MedRxiv*, pp. 2020–12.

Abiodun, G. J., Makinde, O. S., Adeola, A. M., Njabo, K. Y., Witbooi, P. J., Djidjou-Demasse, R. and Botai, J. O., (2019). A dynamical and zero-inflated negative binomial regression modelling of malaria incidence in Limpopo province, South Africa. *International Journal of Environmental Research and Public Health*, 16(11)–2000.

Adams, S. O., Somto, G., (2022). Comparative study of the error trend and seasonal exponential smoothing and ARIMA model using COVID-19 death rate in Nigeria. *International Journal of Epidemiology and Health Sciences*, 3(9).

Adesina, O. S., Onanaye, S. A., Okewole, D. and Egere, A. C., (2020). Forecasting of new cases of COVID-19 in Nigeria using autoregressive fractionally integrated moving average models. Asian Res. J. Math, pp. 135–146.

Agboola, S., Niyang, P., Olawepo, O., Ukponu, W., Niyang, S., Ujata, I., Ihueze, A., Ibrahim, R., Shallangwa, J., Adamu, H. et al., (2021). Forecasting the spread and total size of confirmed and discharged cases of COVID-19 in Nigeria using an ARIMA model. *Statistical Journal of the IAOS*, 37(2), pp. 517–522.

Alabdulrazzaq, H., Alenezi, M., Rawajfih, Y., Alghannam, B., Al-Hassan, A. and Al-Anzi, F., (2021). On the accuracy of ARIMA based prediction of COVID-19 spread. *Results phys*, 27, p. 104509.

Ali, I., Charles, V., Modibbo, U. M., Gherman, T. and Gupta, S., (2023). Navigating COVID-19: unraveling supply chain disruptions through best-worst method and fuzzy topsis. *Benchmarking: An International Journal*, 2023.

Argawu, A. S., (2021). Time series models for covid-19 new cases in top seven infected African countries. *Journal of Pharmaceutical Research International*, 33(60B), pp. 983–992.

Aronu, C. O., Ekwueme, G. O., Sol-Akubude, V. I. and Okafor, P. N., (2021). Coronavirus (COVID-19) in Nigeria: survival rate. *Scientific African*, 11, p. e00689.

ArunKumar, K., Kalaga, D. V., Kumar, C. M. S., Chilkoor, G., Kawaji, M. and Brenza, T. M., (2021). Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving average (SARIMA). *Applied soft computing*, 103, p. 107161.

Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.-Y., Chen, L. and Wang, M., (2020). Presumed asymptomatic carrier transmission of COVID-19. *Jama*, 323(14), pp. 1406–1407.

Busari, S., Samson, T. (2022). Modelling and forecasting new cases of COVID-19 in Nigeria: Comparison of regression, ARIMA and machine learning models. *Scientific African*, 18, p. e01404.

Chaurasia, V., Pal, S., (2020). COVID-19 pandemic: Arima and regression model-based worldwide death cases predictions. *SN Computer Science*, 1(5), p. 288.

Chyon, F. A., Suman, M. N. H., Fahim, M. R. I. and Ahmed, M. S., (2022). Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. *Journal of Virological Methods*, 301, p. 114433.

Didi, E. I., Kingdom, N. and Harrison, E. E., (2021). ARIMA modelling and forecasting of COVID-19 daily confirmed/death cases: a case study of Nigeria. *Asian Journal of Probability and Statistics*, 12(3), pp. 59–80.

Dunford, D., Dale, B., Stylianou, N., Lowther, E., Ahmed, M. and de la Torre Arenas, I., (2020). Coronavirus: The world in lockdown in maps and charts. *BBC News*, 9, p. 462.

Ibrahim, R. R., Oladipo, H. O., (2020). Forecasting the spread of COVID-19 in Nigeria using Box-Jenkins modeling procedure. *MedRxiv*, pp. 2020–05.

Inegbedion, H. E., (2023). A time series forecast of COVID-19 infections, recoveries and fatalities in Nigeria. *Sustainability*, 15(9), p. 7324.

Khan, F. M., Gupta, R., (2020). ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *Journal of Safety Science and Resilience*, 1(1), pp. 12–18.

Kucharski, A. J., Russell, T.W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R. M., Sun, F., Jit, M., Munday, J. D. et al., (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The lancet infectious diseases*, 20(5), pp. 553–558.

Kumar, N., Susan, S., (2020). COVID-19 pandemic prediction using time series forecasting models. *In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–7.

Li, C., Sampene, A. K., Agyeman, F. O., Robert, B. and Ayisi, A. L., (2022). Forecasting the severity of COVID-19 pandemic amidst the emerging SARS-COV-2 variants: adoption of ARIMA model. *Computational and Mathematical Methods in Medicine*.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y. et al., (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.

Lukman, A. F., Rauf, R. I., Abiodun, O., Oludoun, O., Ayinde, K. and Ogundokun, R. O., (2020). COVID-19 prevalence estimation: Four most affected African countries. *Infectious Disease Modeling*, 5, pp. 827–838.

Malki, Z., Atlam, E.-S., Ewis, A., Dagnew, G., Alzighaibi, A. R., ELmarhomy, G., Elhosseini, M. A., Hassanien, A. E. and Gad, I., (2021). ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Computing and Applications*, 33, pp. 2929–2948.

Modibbo, U. M., Arshad, M., Abdalghani, O. and Ali, I., (2021). Optimization and estimation in system reliability allocation problem. *Reliability Engineering & System Safety*, 212, p. 107620.

Nguyen, Q. D., Le Phuong, T., Dinh, T. N. Q., Le Thanh, B., Cao, T. A. L. and Phung, T. H. D., (2020). Predicting the pandemic covid-19 using ARIMA model. VNU Journal of Science: Mathematics-Physics, 36(4).

Nwafor, G. O., Iwu, H. C. and Anyasodo, U. N., (2022). Transfer function modeling of COVID-19 pandemic in Nigeria. *Journal of the Nigerian Statistical Association*, Vol. 34.

Odekina, G. O., Adedotun, A. F. and Imaga, O. F. (2022). Modeling and forecasting the third wave of COVID-19 incidence rate in Nigeria using vector autoregressive model approach. *Journal of the Nigerian Society of Physical Sciences*, pp. 117–122.

Oduntan, E. A., Ajayi, O. O., (2023). ARIMA forecast of Nigerian inflation rates with COVID-19 pandemic event in focus. *Theoretical & Applied Economics*, 30(4).

Ogbuagada, S., Okolo, A., Torsen, E. and John, O., (2022). Multivariate time series analysis in modeling Malaria cases in Jimeta metropolis of Adamawa State, Nigeria. *FUDMA Journal of Sciences*, 6(3), pp. 62–69.

Olarenwaju, B. A., Harrison, I. U., (2020). Modeling of COVID-19 cases of selected states in Nigeria using linear and non-linear prediction models. *Journal of Computer Sciences Institute*, 17, pp. 390–395.

Organization, W. H. et al., (2020). Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19). Geneva: World Health Organization; 2020. *Google Scholar*, pp. 1–40.

Ortese, C., Ieren, T. and Tamber, A., (2021). A time series model to forecast COVID-19 infection rate in Nigeria using Box-Jenkins method. *Nigerian Annals of Pure and Applied Sciences*, 4(1), pp. 75–85.

Poleneni, V., Rao, J. K. and Hidayathulla, S. A., (2021). COVID-19 prediction using ARIMA model. *In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 860–865. IEEE.

Ribeiro, M. H. D. M., da Silva, R. G., Mariani, V. C., and dos Santos Coelho, L., (2020). Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. *Chaos, Solitons & Fractals*, 135, p. 109853.

Roosa, K., Lee, Y., Luo, R., Kirpich, A., Rothenberg, R., Hyman, J. M., Yan, P. and Chowell, G., (2020). Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. *Infectious Disease Modeling*, 5, pp. 256–263.

Sah, S., Surendiran, B., Dhanalakshmi, R., Mohanty, S. N., Alenezi, F., Polat, K. et al., (2022). Forecasting COVID-19 pandemic using prophet, ARIMA, and hybrid stacked lstm-gru models in India. *Computational and Mathematical Methods in Medicine*, 2022.

Sam, S. O., (2020). Exploring the statistical significance of Africa COVID-19 data. *International Journal of Statistics and Applied Mathematics*, 5(4), pp. 34–42.

Samson, T. K., Ogunlaran, O. M. and Raimi, M. O., (2020). A predictive model for confirmed cases of COVID-19 in Nigeria. *European Journal of Applied Sciences (EJAS)*, Vol. 8, No. 4, pp 1–10.

Shoko, C., Njuho, P., (2023). ARIMA model in predicting of COVID-19 epidemic for the Southern Africa region. *African Journal of Infectious Diseases*, 17(1), pp. 1–9.

Somyanonthanakul, R., Warin, K., Amasiri, W., Mairiang, K., Mingmalairak, C., Panichkitkosolkul, W., Silanun, K., Theeramunkong, T., Nitikraipot, S. and Suebnukarn, S., (2022). Forecasting COVID-19 cases using time series modeling and association rule mining. *BMC Medical Research Methodology*, 22(1), p. 281.

Suleiman, S., Sani, M., (2021). Application of ARIMA and artificial neural networks models for daily cumulative confirmed COVID-19 prediction in Nigeria. *Equity Journal of Science and Technology*, 7(2), pp. 83–83.

Tang, B., Bragazzi, N. L., Li, Q., Tang, S., Xiao, Y. and Wu, J. (2020). An updated estimation of the risk of transmission of the novel coronavirus (2019-NCOV). *Infectious disease modeling*, 5, pp. 248–255.

Tawiah, K., Iddrisu, W. A. and Asampana, A. K., (2021). Zero-inflated time series modelling of COVID-19 deaths in Ghana. *Journal of Environmental and Public Health*.

Team, E., (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020. *China CDC weekly*, 2(8), p.113.

Yang, M., (2012). Statistical models for count time series with excess zeros *[PhD (doctor of philosophy) Thesis]. University of Iowa*.

Yang, M., Zamba, G. K. and Cavanaugh, J. E., (2013). Markov regression models for count time series with excess zeros: A partial likelihood approach. *Statistical Methodology*, 14, pp. 26–38.

Yang, Q., Wang, J., Ma, H. and Wang, X., (2020). Research on covid-19 based on ARIMA model δ—taking Hubei, China as an example to see the epidemic in Italy. *Journal of infection and public health*, 13(10), pp. 1415–1418.

Zhihao, L., Junpei, W., Xiaoliang, Z. and Huijun, N., (2021). Research on covid-19 epidemic based on on ARIMA model. In Journal of Physics: Conference Series, Vol. 2012, pp. 012-063. *IOP Publishing*, 17.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 221–231, https://doi.org/10.59139/stattrans-2025-012
Received – 31.01.2024; accepted – 30.09.2024

Optimal sample size in a triangular model for sensitive questions

Stanisław Jaworski[1]

Abstract

The estimation of the fraction of a population with a stigmatizing characteristic is the issue that this study attempts to address. In this paper the nonrandomized response model proposed by Tian et al. (2007) is considered. The exact confidence interval (CI) for this fraction is constructed. The optimal sample size for obtaining the CI of a given length is also derived. In order to estimate the proportion of the population with a stigmatizing characteristic, we explore the nonrandomized response model proposed by Tian et al. (2007). The prevalent approach to constructing a CI involves applying the Central Limit Theorem. Unfortunately, such CIs fail to consistently maintain the prescribed confidence level, contradicting the Neyman (1934) definition of CIs. In this paper, we present the construction of an exact CIs for this proportion, ensuring adherence to the designated confidence level. The length of the proposed CI depends on both the given probability of a positive response to a neutral question and the sample size. For these CIs, the probability of a positive response to a neutral question is established in relation to the provided limit on the privacy protection of the interviewee. Additionally, we derive the optimal sample size for obtaining a CI of a given length.

Key words: sensitive questions, nonrandomized response model, exact confidence interval.

1. Introduction

In surveys aiming to estimate the proportion of individuals with a stigmatizing characteristic, respondents often hesitate to provide truthful responses when directly questioned. To address this challenge, various methods of indirect questioning have been developed to safeguard privacy and encourage the disclosure of sensitive information. The initial approach to obscuring answers to sensitive questions was proposed by Warner (1965). This method involves the randomization of responses, with the interviewee determining the randomized answer, and the interviewer remaining unaware of the actual response to the sensitive question. Over time, Warner's model has been extended in different ways by researchers such as Horvitz et al. (1967), Greenberg et al. (1969), Raghavarao (1978), Franklin (1989), and Kuk (1990). Collectively, Warner's model and its extensions fall under the category of randomized response techniques, which necessitate the use of a randomization device.

Tian et al. (2007) and Yu et al. (2008) introduced two innovative techniques for addressing sensitive questions in population surveys: the triangular model and the crosswise model.

[1]The Institute of Economics and Finance, Warsaw University of Life Sciences, Nowoursynowska 159, 02-787 Warsaw, Poland. E-mail: stanislaw_jaworski@sggw.edu.pl. ORCID: https://orcid.org/0000-0002-6169-2886.

Both models involve asking two questions simultaneously – one sensitive and one neutral. A key advantage of these methods is that they do not require a randomization device, unlike earlier approaches. The triangular and crosswise models, along with the parallel model introduced by Tian (2014), belong to the same class of non-randomized models. The issue of determining the optimal sample size for these models has been examined by Liu and Tian (2014) and Yu et al. (2008).

An essential aspect of the sample survey design is determining the number of respondents. Tian et al. (2011) explored sample size determination for the non-randomized triangular model when dealing with sensitive questions in surveys. Their approach involved precision and power analyses for one-sided and two-sided tests, examining the hypothesis $H_0 : \pi = \pi_0$, where π represents the population proportion with the sensitive characteristic, and π_0 is a pre-specified reference value. The sample size determination was guided by controlling the type I and II error rates of the tests. However, the resulting solution depends on both the pre-specified reference value π_0 and the true unknown value of π, making it challenging to apply directly in practical situations.

Qiu et al. (2014) also examined sample size determination for the triangular model, deriving formulas for estimating the parameter π. Unlike Tian et al. (2011), they explicitly incorporated an assurance probability of achieving the pre-specified precision into the formulas. However, these formulas still depend on the unknown value of π and are based on asymptotic confidence intervals, which do not maintain the nominal confidence level.

In this study, we present an alternative approach to determining the optimal sample size for the non-randomized triangular model. This approach was originally introduced by Jaworski and Zieliński (2023) for the non-randomized crosswise model. Their method simultaneously considers both the confidence interval length and the protection of respondent privacy.

In Section 2, we revisit the construction of asymptotic confidence intervals for π and elucidate the process of constructing an exact confidence interval for this parameter. Section 3 introduces the methodology for sample size selection, taking into account the privacy of the interviewee. Section 4 delves into various aspects of the numerical determination of the optimal sample size. Concluding remarks are provided in Section 5.

2. Confidence interval in Triangular Model

Let Y be a binary variable, where $\{Y = 1\}$ indicates the occurrence of drawing a person with a stigmatizing trait, and $\{Y = 0\}$ is the complement to $\{Y = 1\}$. Our focus is on estimating the proportion (denoted by $\pi = P\{Y = 1\}$) of individuals with the stigmatizing trait and constructing a confidence interval for π. The challenge we encounter is that the random variable Y is not reliably observable. Therefore, we observe another variable Z, contingent on respondents' answers to two questions. The relationship between Z and the two questions is specified by the assumed model.

In the triangular model, respondents are simultaneously presented with two independent questions–one neutral and one sensitive. They are instructed to report 0 only if the answers to both questions are not positive (*NO*). Thus, the observable variable in this model

is denoted as Z, where

$$Z = \begin{cases} 0, & \text{if both answers are } NO, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

In the triangular model the probability q of answering *YES* to neutral question is assumed to be known. Therefore

$$Z = \begin{cases} 0, & \text{with probability } (1-\pi)(1-q), \\ 1, & \text{with probability } \pi + (1-\pi)q. \end{cases} \tag{2}$$

Let us denote the probability $\pi + (1-\pi)q$ by ρ. Hence, in the triangular model

$$\pi = \frac{\rho - q}{1-q}. \tag{3}$$

Let $Z_1, Z_2, \ldots, Z_n$ be a sample. Maximum likelihood estimator (MLE) of ρ is $\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Therefore, $\hat{\pi}_q = \dfrac{\hat{\rho} - q}{1-q}$ is a natural estimator of π. However, the MLE of π has the form

$$\hat{\pi} = \max\left\{0, \pi_q\right\}. \tag{4}$$

Yu et al. (2008) proved that the estimators $\hat{\pi}$ and $\hat{\pi}_q$ are asymptotically equivalent. When $n \to \infty$, the central limit theorem implies that $\hat{\pi}_q$ is asymptotically normal. Hence, the following $\delta 100\%$ Wald confidence interval of π can be constructed:

$$\hat{\pi}_q \pm z_{\frac{1-\delta}{2}} \sqrt{v(\hat{\pi}_q)} \tag{5}$$

where z_ν denotes the upper $\nu-$th quantile of the standard normal variable and $v(\hat{\pi}_q) = \hat{\rho}(1-\hat{\rho})/[(n-1)(1-q)^2]$.

It is also possible to construct $\delta 100\%$ Wilson (score) confidence interval of π:

$$\left\{ \pi \in \langle 0,1 \rangle : \ (\hat{\pi}_q - \pi)^2 \leq \frac{z^2_{\frac{1-\delta}{2}} Var(\hat{\rho})}{(1-q)^2} \right\}, \tag{6}$$

where $Var(\hat{\rho}) = \big(\pi + (1-\pi)q\big)(1-\pi)(1-q)/n$.

The Wald and Wilson confidence intervals are known to deviate from the prescribed confidence level, making them imprecise. In contrast, the Clopper-Pearson method (Clopper and Pearson (1934)) can be employed to construct an exact confidence interval for π. Notably, since π is a linear and increasing function of ρ, the resulting exact confidence interval for π is

$$(\pi_L(\hat{\pi}), \pi_R(\hat{\pi})) = \left(\max\left\{0, \frac{\rho_L(\hat{\rho}) - q}{1-q}\right\}, \frac{\rho_R(\hat{\rho}) - q}{1-q} \right) \tag{7}$$

where $(\rho_L(\hat{\rho}), \rho_R(\hat{\rho}))$ is a Clopper-Pearson exact confidence interval of ρ, that is

$$\rho_L(\hat{\rho}) = \begin{cases} 0 & \text{dla } \hat{\rho} = 0, \\ B^{-1}\left(n - n\hat{\rho} + 1, n\hat{\rho}; \frac{1+\delta}{2}\right) & \text{dla } \hat{\rho} > 0, \end{cases} \tag{8}$$

$$\rho_U(\hat{\rho}) = \begin{cases} 1 & \text{for } \hat{\rho} = 1, \\ B^{-1}\left(n - n\hat{\rho}, n\hat{\rho} + 1; \frac{1-\delta}{2}\right) & \text{for } \hat{\rho} < 1, \end{cases} \tag{9}$$

where $B^{-1}(a,b;\cdot)$ denotes the inverse of CDF of the Beta distribution with parameters (a,b). Note, that it is enough to use the $B^{-1}(\cdot,\cdot;\cdot)$ function for setting the exact confidence interval.

3. Optimal sample size

Let us consider the length $l(\hat{\pi};q,n)$ of the exact confidence interval. For the $n\hat{\rho}$ observed *YES* answers to the questionnaire we have

$$l(\hat{\pi};q,n) = \pi_R(\hat{\pi}) - \pi_L(\hat{\pi}), \quad \text{where } \hat{\pi} = \max\left\{0, \frac{\hat{\rho} - q}{1 - q}\right\}. \tag{10}$$

The length of the confidence interval is a random variable concerning $\hat{\pi}$, contingent on q and n. We explore two approaches to minimize the length of the CI:

1. **Minimizing expected length**: Find minimal sample size n such that the expected length of the confidence interval does not exceed a predetermined value.

2. **Almost sure minimizing**: Find minimal sample size n such that there is a high probability that the length of the confidence interval does not exceed a predetermined value.

The solution of these approaches is influenced by the probability of a positive answer to the neutral question. Thus, a rational criterion for the optimal selection of this probability needs to be formulated. Denoting the optimally selected q, dependent on the sample size n, as $q_e(n)$ and $q_d(n)$ in the first and second approaches, respectively. Let Π and Q represent acceptable sets for π and q, respectively. In the absence of prior knowledge about π and reasonable restrictions for q, the sets are $\Pi = (0,1)$ and $Q = \langle 0,1)$.

Optimal q in the first approach. Let $\mathscr{X}$ denote the sample space of $\hat{\pi}$. The problem may be written in the following way:

$$q_e(n) = \arg\min_{q \in Q} \sup_{\pi \in \Pi} E_\pi^{C(\pi)} l(\hat{\pi};q,n), \tag{11}$$

where $E_\pi^{C(\pi)} l(\hat{\pi};q,n) = \sum_{x \in C(\pi)} l(x;q,n) P_\pi\{\hat{\pi} = x\}$ represents the expected length of the CI covering estimated value of π. Here, the set $C(\pi) = \{x \in \mathscr{X} : \ \pi_L(x) < \pi < \pi_R(x)\}$ comprises the values of the variable $\hat{\pi}$ for which the CI covers π.

Optimal q in the second approach. The problem may be written in the following way:

$$q_d(n) = \arg\max_{q\in Q} \inf_{\pi\in\Pi} P_\pi^{C(\pi)} \{l(\hat{\pi};q,n) \le d\}, \tag{12}$$

where $\delta \cdot P_\pi^{C(\pi)} \{l(\hat{\pi},q,n) \le d\} = \sum_{x\in C(\pi)} P_\pi\{\hat{\pi} = x\}\mathbb{1}(l(x,q,n) \le d)$ represents the probability that the length of the CI covering the estimated value of π does not exceed the given value d. The function $\mathbb{1}(p)$ is equal to one if the logical value of p is true and zero otherwise.

In the case of $Q = \langle 0,1)$, the minimal length concerning q is achieved when $q = 0$, equivalent to not asking the neutral question. However, such a questionnaire (without a neutral question) fails to ensure the privacy of respondents. Therefore, it is reasonable to impose a constraint on the probability q, considering the desired level of protection.

Tan et al. (2009) introduced the concept of the degree of privacy protection through the probabilities

$$P_\pi\{Y = 1|Z = 1\} \text{ and } P_\pi\{Y = 1|Z = 0\}. \tag{13}$$

These probabilities are connected with the safety of the interviewee of non-discovering her/his positive answer to the sensitive question. These probabilities should be small enough so that they do not exceed the given value $\gamma \in (0,1)$. The researcher can set this value according to the requirements of the conducted survey. In the triangular model, the aforementioned probabilities are as follows:

$$\begin{aligned} P_\pi\{Y = 1|Z = 1\} &= \frac{\pi}{\pi + (1-\pi)q}, \\ P_\pi\{Y = 1|Z = 0\} &= 0. \end{aligned} \tag{14}$$

The relationship between the probability $P_\pi\{Y = 1|Z = 1\}$ and q is illustrated in Figure 1.

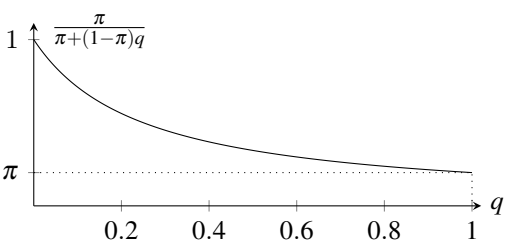


Figure 1: Privacy protection versus q.

We are interested in $q < 1$ such that

$$\frac{\pi}{\pi + (1-\pi)q} \le \gamma \quad \text{for } \pi \in \Pi. \tag{15}$$

Simple algebra yields the following condition for q:

$$q(\pi;\gamma) \le q < 1 \quad \text{for } \pi \in \Pi, \tag{16}$$

where $q(\pi;\gamma) = \frac{\pi(1-\gamma)}{\gamma(1-\pi)}$ increases with respect to π. Since $q(\gamma,\gamma) = 1$, the condition (16) holds if and only if $\gamma > \pi$. This implies that the maximal privacy protection (i.e. the minimal γ to be chosen) is restricted by the percentage of the population that has committed socially stigmatizing characteristic. Consequently, the problem of minimizing the length, assuming $\pi \leq \pi_0$ for a given $\in (0,1)$, is well defined for $q \in \langle q(\pi_0;\gamma),1)$. In the following discussion, we assume that $\Pi = (0,\pi_0\rangle$ and $Q = \langle q(\pi_0;\gamma),1)$, where $\gamma > \pi_0$. The value π_0 reflects our prior knowledge about π, indicating that we know the percentage of people bearing a stigmatizing characteristic is less than π_0. The inequalities (15) and (16) lead us to the conclusion that without this knowledge, determining the appropriate value for γ is not feasible. Note that both Π and Q do not depend on the sample size n. Therefore, the length of the CI can be minimized by selecting an appropriate sample size. Let $d \in (0,1)$ be a given number. Our goal is to determine the sample size that yields a CI with a length not exceeding d. Specifically, we are interested in a CI covering the estimated value of π. We can define two approaches to address this problem.

Optimal sample size in the first approach. Identify minimal n such that

$$E_{\pi}^{C(\pi)} l(\pi, q_e(n), n) \leq d \quad \text{for all } \pi \in \Pi. \tag{17}$$

Optimal sample size in the second approach. Identify minimal n such that

$$P_{\pi}^{C(\pi)} \{l(\pi, q_d(n), n) \leq d\} \geq 1-\lambda \text{ for given } 1-\lambda \text{ and all } \pi \in \Pi. \tag{18}$$

In the first approach, our objective is to ensure that the average length of the CI covering the estimated value of π is less than a given d. In the second approach, our goal is to ensure that the length of at least $(1-\lambda)\%$ of the CIs covering the estimated π is less than the specified d. It is important to note that we have a minimum of $\delta\%$ of intervals covering the unknown parameter π, and for an infinitely large sample size n, the defined value $P_{\pi}^{C(\pi)} \{l(\hat{\pi}, q, n) \leq d\}$ is equal to one.

The approaches to determining the optimal sample size were initially introduced for the non-randomized crosswise model by Jaworski and Zieliński (2023).

4. Numerical consideration

Let us assume that $\pi < 0.5$ and the confidence level is set at $\delta = 0.95$. Moving on to the first approach, an analysis of $E_{\pi}^{C(\pi)} l(\hat{\pi}, q, n)$ reveals (refer to Figure 2) that for each $\pi < 0.5$ and n it increases with q. Consequently, it can be inferred that $q_e(n) = q(\pi_0, \gamma)$ for any $\gamma \in (0, 0.5\rangle$. In the triangular model, $q(\pi_0, \gamma)$ decreases with γ. Hence, our interest lies in identifying the smallest and acceptable value of γ. However, it is crucial to note that the measure of privacy protection revealed by the triangular model cannot be zero. Hence, opting for $\gamma = 0.5$ appears reasonable. In this scenario, the probability that the respondent belongs to a sensitive group is 50%, thereby mitigating the legal risks associated with the respondent's answers in the survey.

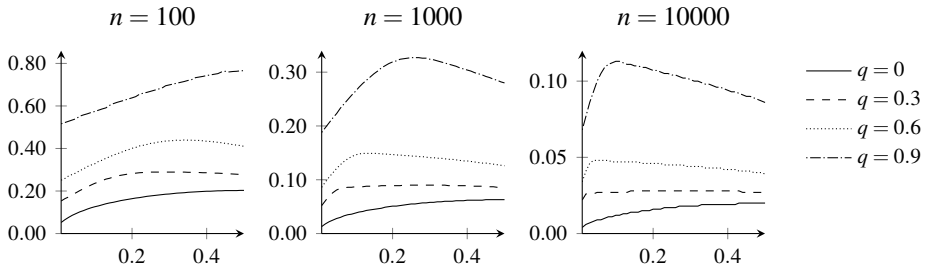


Figure 2: Expected length versus $\pi \in (0, \gamma = 0.5)$ with respect to q under the condition that π is covered by the CI.

The expected length $E_{\pi}^{C(\pi)} l(\hat{\pi}, q, n)$ is not monotonic with π for every q. Let us define

$$\pi_{max}(n; \pi_0) = \underset{\pi \in \Pi}{\operatorname{argsup}} E_{\pi}^{C(\pi)} l(\pi, q(\pi_0, 0.5), n). \tag{19}$$

It is depicted in Figure 3 that if $\pi_0 \leq 0.25$ then $\pi_{max}(n; \pi_0) = \pi_0$ otherwise it is a decreasing function of sample size n (with the accuracy implied by the discreteness of the distribution of the observed variable). This knowledge of π_{max} can be helpful in the optimal sample size numerical finding in the approach. In Table 1 some exemplary of optimal sample sizes are given for confidence level $\delta = 0.95$ and privacy protection level $\gamma = 0.5$. The optimal sample size is increasing with π_0. Larger values of π_0 correspond to higher uncertainty about parameter π. Therefore, the optimal sample sizes are smaller for smaller values of π_0.

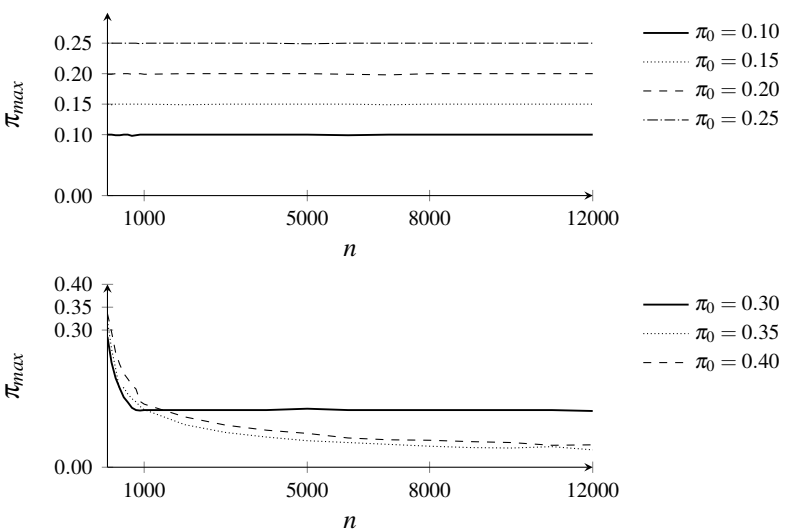


Figure 3: $\pi_{max}(n) = \operatorname{argsup}_{\pi \in \Pi} E_{\pi}^{C(\pi)} l(\pi, q(\pi_0, 0.5), n)$ versus sample size n

Consider the case for $d = 0.06$. The optimal sample size is equal to 822 when $\pi_0 = 0.1$ and is about 8 times greater for $\pi_0 = 0.4$. This means that the costs of conducting a survey are much higher for the latter case. Recall that when we conduct a survey by asking the sensitive question directly with no additional neutral question, the length of the CI for π is equal approximately to 0.06 when sample size $n = 1000$. This remark enable us to conclude that without additional knowledge about the scope of π we will incur much higher research costs with an appropriately secured level of privacy.

Table 1: The smallest n that $\sup_{\pi\in\Pi} E_\pi^{C(\pi)} l(\pi, q(\pi_0,\gamma), n) \le d$.

π_0	$q(\pi_0,\gamma)$	$d = 0.05$	$d = 0.06$
0.1	0.11	1171	822
0.2	0.25	2422	1693
0.3	0.43	4146	2893
0.4	0.67	8120	5646

Note: $q(\pi_0,\gamma) = \frac{\pi_0}{1-\pi_0}$ for $\gamma = 0.5$

Now, let us consider the second approach. The probability $P_\pi^{C(\pi)}\{l(\pi,q,n) \le d\}$ decreases with q, and this dependency is illustrated in Figures 4, 5 and 6. When comparing Figures 4 and 5, we observe that the monotonicity of $P_\pi^{C(\pi)}\{l(\pi,q,n) \le d\}$ concerning π depends on the sample size n. It is noteworthy that the lines in Figures 4, 5 and 6 exhibit some lack of smoothness due to the discreteness of the observed variable, albeit small enough to explore the optimal sample size at $q = q(\pi_0,\gamma)$. In Table 2, we provide some exemples of optimal sample sizes.

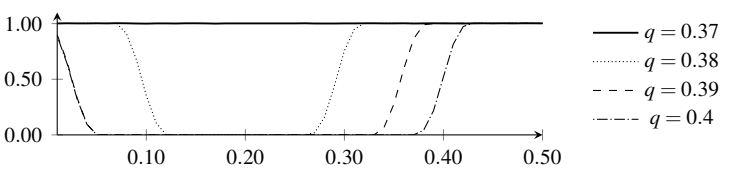


Figure 4: The probability as a function of π, with respect to q under the condition that π is covered by the CI. Here, $n = 4000$.

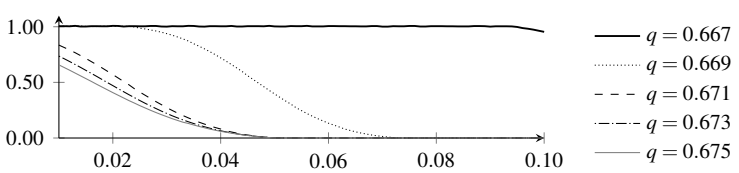


Figure 5: The probability as a function of π, with respect to q under the condition that π is covered by the CI. Here, $n = 1359$.

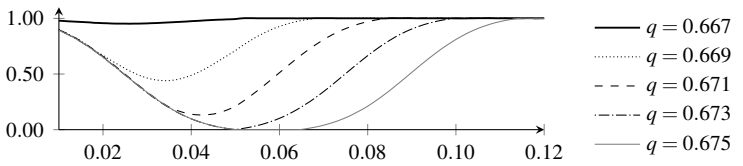


Figure 6: The probability as a function of π, with respect to q under the condition that π is covered by the CI. Here, $n = 12243$.

Please note that the optimal sample size exhibits a modest variation when λ is set to 0.01 compared to 0.05, with the maximum observed difference being 13 (refer to Table 2). However, noteworthy disparities arise in relation to π_0, contingent upon the prior knowledge of the true value of the π parameter. Increased uncertainty regarding π results in a higher optimal sample size requirement. For instance, when $\pi = 0.4$, the optimal sample size is approximately nine times greater than that for $\pi = 0.4$.

Table 2: The smallest n that $\inf_{\pi\in\Pi} P_\pi^{C(\pi)}\{l(\pi, q(\pi_0;\gamma), n) \le d\} \ge 1-\lambda$.

		$d = 0.05$		$d = 0.06$	
π_0	$q(\pi_0,\gamma)$	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.01$	$\lambda = 0.05$
0.1	0.11	1376	1359	973	960
0.2	0.25	2708	2702	1891	1886
0.3	0.43	4774	4774	3324	3324
0.4	0.67	12250	12243	8499	8494

Note: $q(\pi_0,\gamma) = \frac{\pi_0}{1-\pi_0}$ for $\gamma = 0.5$

5. Conclusions

The paper introduces a novel CI for the fraction of sensitive questions in the triangular model. Unlike the widely used asymptotic CI, the new approach maintains the prescribed confidence level, which is consistent with Neyman's (1934) definition of CIs.

Addressing a crucial practical concern, we derived the minimum sample size satisfying two criteria: average length and almost sure length. To obtain these sample sizes, we impose restrictions on privacy protection, specifically the probability of discovering a YES answer to the sensitive question. This probability should be sufficiently small to ensure the interviewee's comfort in answering the questionnaire. Additionally, we limit our analysis to rare phenomena, focusing on sensitive questions with a small (predefined) probability of a positive answer.

It is crucial to emphasize that we refrain from comparing the length of our CI with asymptotic versions. Asymptotic CIs are inherently shorter because they lack the capability to uphold a specified confidence level, leading to a real probability of coverage that is less than the designated confidence level. Consequently, the comparison of lengths is devoid

of meaningful insights. Our CI is characterized by its ease of calculation; even a standard spreadsheet application can efficiently compute the quantiles of the Beta distribution. While asymptotic CIs based on normal approximation served a purpose in times when computers were not readily available, we advocate for the practical application of our CI in contemporary scenarios.

The provided numerical examples demonstrate that incorporating prior knowledge of the true value of π enables a reduction in the minimum sample size necessary to achieve the desired estimation precision. In the absence of this knowledge, the optimal sample size may inflate by more than eight times, posing an unfavorable scenario given the associated research costs.

References

Clopper, C. J., Pearson, E. S., (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika, 26(4)*, pp. 404–413.

Franklin, L. A., (1989). Randomized response sampling from dichotomous populations with continuous randomization. *Survey Methodology*, 15, pp. 225–235.

Greenberg, B. G., Abul-Ela, A.-L. A. and Horvitz, D. G., (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association*, 64, pp. 520–539.

Groenitz, H., (2014). A new privacy-protecting survey design for multichotomous sensitive variables. *Metrika*, 77, pp. 211–224.

Horvitz, D. G., Shah, B. V., Simmons and W. R., (1967). The Unrelated Question Randomized Response Model. *in Proceedings of the Social Statistics Section. American Statistical Association*, pp. 65–72.

Jaworski, S., Zieliński, W., (2023). The Optimal Sample Size in the Crosswise Model for Sensitive Questions. *Applicationes Mathematicae*, 50(1), pp. 2–34.

Kuk, A. Y. C., (1990). Asking Sensitive Question Indirectly. *Biometrika*, 77, pp. 436–438.

Liu, Y., Tian, G.-L., (2014). Sample size determination for the parallel model in a survey with sensitive questions. *Journal of the Korean Statistical Society*, 43(2), pp. 235–249. doi: 10.1016/j.jkss.2013.08.002.

Neyman, J., (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97, pp. 558–625.

Qiu, S.-F., Zou, G. Y. and Tang, M.-L., (2014). Sample size determination for estimating prevalence and a difference between two prevalences of sensitive attributes using the non-randomized triangular design. *Computational Statistics & Data Analysis*, 77, pp. 157–169. doi: 10.1016/j.csda.2014.02.019.

Raghavarao, D., (1978). On an Estimation Problem in Warner's Randomized Response Technique. *Biometrics. [Wiley, International Biometric Society]*, 34(1), pp. 87–90. Available at: http://www.jstor.org/stable/2529591 (Accessed: 1 July 2022).

Tan, M., Tian, G. L. and Tang, M. L., (2009). Sample Surveys With Sensitive Questions: A Non-Randomized Response Approach. *The American Statistician*, 63, pp. 9–16.

Tian, G. L., Yu, J. W., Tang, M. L. and Geng, Z., (2007). A New Nonrandomized Model for Analyzing Sensitive Questions with Binary Outcomes. *Statistics in Medicine*, 26, pp. 4238–4252.

Tian, G. L., Tang, M. L., Liu, Z., Tan, M. and Tang, N. S., (2011). Sample Size Determination for the Non-Randomised Triangular Model for Sensitive Questions in a Survey. *Statistical Methods in Medical Research*, 20, pp. 159–173.

Tian, G. L., (2014). A New Non-Randomized Response Model: The Parallel Model. *Statistica Neerlandica*, 68(4), pp. 293–323.

Warner, S. L., (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60, pp. 63–69.

Yu, J. W., Tian, G. L. and Tang, M. L., (2008). Two New Models for Survey Sampling With Sensitive Characteristic: Design and Analysis. *Metrika*, 67(3), pp. 251–263.

STATISTICS IN TRANSITION new series, March 2025
Vol. 26, No. 1, pp. 233–238

About the Authors

Ali Irfan has received B.Sc., M.Sc., M.Phil., and Ph.D. degrees from Aligarh Muslim University. He is currently a working faculty member with the Department of Statistics and Operations Research, Aligarh Muslim University. He received the Post Graduate Merit Scholarship Award during his M.Sc. (statistics) and the UGC-BSR Scholarship award during his Ph.D. (statistics) program. His research interests include applied statistics, survey sampling, reliability theory, supply chain networks and management, mathematical programming, fuzzy optimization, and multiobjective optimization. He supervised M.Sc., M.Phil., and Ph.D. students in operations research. He completed a research project UGC–Start-Up Grant Project, UGC, New Delhi, India. He published more than 100 research articles in SCI/SCIE and other reputed journals and serves as a Reviewer for several journals. He published some edited books for Taylor France and Springer Nature publishers, and some are in the process of publication. He has currently published one textbook "Optimization-with-LINGO-18-Problems-and-Applications". This book is helpful for academicians, practitioners, students, and researchers in the field of OR. He is a Lifetime Member of various professional societies: Operational Research Society of India, Indian Society for Probability and Statistics, Indian Mathematical Society, and The Indian Science Congress Association. He delivered invited talks in several universities and institutions. He also serves as some journals' Associate Editor and Guest Editor for SCI/SCIE.

Ali Majid Khan Majahar is an Associate Professor and a researcher at the School of Mathematical Sciences, Universiti Sains Malaysia. He has earned his Ph.D. in Mathematics from Universiti Malaysia Sabah. His research spans seaweed cultivation, solar drying systems, and the use of big data for models and simulations to predict moisture loss under various environmental conditions.

Anand R. is a fulltime research scholar in the Department of Statistics at the University of Calicut, Kerala. His research interests include reliability theory, survival analysis, extreme value theory and distribution theory.

Chakraborty Subrata is a Professor (Full) of the Department of Statistics, Dibrugarh University, Assam, India. His research interests are probability distributions (discrete and continuous) and reliability estimation in particular. Professor S.C. has published more than 150 research papers in international/national reputed journals. Professor Chakraborty is a Post-Doctoral research Fellow of the Institute of Mathematical

Sciences, University of Malaya, Kuala Lumpur also Visiting Associate, Department of Mathematics and Statistics, University of Calgary, Canada. Professor S.C. is an active member of many scientific professional bodies.

Chipepa Fastel is a lecturer of Mathematical Statistics at Botswana International University of Science and Technology. His research interests include distributional theory, biostatistics, biometry, survival analysis, categorical data analysis, reliability theory and multivariate statistical analysis. Dr Chipepa has published more than 50 research papers in international journals. He has also published two book chapters. Dr Chipepa is a committee member of the Central Botswana Mathematics and Statistical Sciences Conference and Zimbabwe Statistics Association. He is in the editorial board of two international journal**s.**

Deepawansa Dilshanie Diana, is the Deputy Director (Statistics) at the Department of Census and Statistics in Sri Lanka, where she plays a key role in the planning and execution of household surveys, survey design, and data analysis and report writing to inform policy decisions. Her research interests are poverty measurements including multidimensional approaches, small area estimation, poverty mapping, income inequality and food security. She actively contributes to several technical committees and has been a member of the International Association for Official Statistics (IAOS) and the International Statistical Institute (ISI). Her research focuses on developing robust methodologies to enhance the accuracy of poverty and inequality assessments, supporting evidence-based policymaking.

Dileepkumar M. is an Assistant Professor in the Department of Statistics at the University of Calicut, Kerala. His research interests include reliability theory, survival analysis, and distribution theory. He has authored over 15 research papers published in national and international journals. Additionally, Dr. Dileepkumar M is an active member of various statistical organizations.

Dunusinghe Priyanga a senior lecturer, attached to the Department of Economics, University of Colombo, Sri Lanka. He has obtained his first degree, specializing in economics, from the University of Colombo and pursued his higher studies at Kyushu University, Japan. He teaches quantitative subjects such as econometrics and statistics both at undergraduate and post-graduate levels. His research interest includes labor market, poverty, development economics. He has contributed to a number of journals and worked as a consultant to a number of organizations such as ILO, World Bank, ADB, FAO, WFP, and UNDP. He is also a regular contributor to both printed and electronic media on socio-economic issues.

Handique Laba is currently working as an Assistant Professor in the Department of Statistics, Darrang College, Tezpur, Assam, under Gauhati University, India. He has received his Ph.D. in Statistics from the Dibrugarh University under the supervision of Dr. Subrata Chakraborty. His research interests are probability distribution theory,

probability theory and generalized classes of distributions. He has published more than 30 research papers in international/national reputed journals. He serves as a reviewer and an editorial board member of many reputed journals. Current project is "New extended/generalized families of continuous probability distributions".

Ismail Mohd Tahir is an Associate Professor and a researcher at the School of Mathematical Sciences, Universiti Sains Malaysia. His research interests include financial time series, econometrics, categorical data analysis, and applied statistics. He is currently the Vice President of the Malaysian Mathematical Sciences Society and an active member of other scientific professional bodies.

Jamal Farrukh is currently an Assistant Professor in the Department of Statistics, the Islamia University of Bahawalpur Pakistan. He worked as a lecturer in Government S.A. postgraduate College from 2012 to 2020, and Statistical Officer in the Department Agriculture Government of Punjab from 2007 to 2012. He received Ph.D. degrees in Statistics from the Islamia University of Bahawalpur (IU), Pakistan in 2017 under the supervision of Dr. M. H. Tahir. His research interests are probability theory, distribution theory, generalized classes of distributions, compounding techniques, reliability analysis, machine learning, data science, statistical inference and econometrics. He has more than 130 publications to his credit. He serves as a reviewer and an editorial board member of many international journals

Jaworski Stanisław is an Assistant Professor of the Department of Econometrics and Statistics, Faculty of Applied Informatics and Mathematics, Warsaw University of Life Sciences. His research interests include statistical inference, data analysis, sample surveys with sensitive questions.

Kalinowski Sławomir is the Head of the Department of Rural Economics and Secretary of the Scientific Council at the Institute of Rural and Agricultural Development, Polish Academy of Sciences. He is also a member of the Committee on Labour and Social Policy Sciences of the Polish Academy of Sciences. His research focuses on rural development, socio-economic inequalities, and agricultural policy. He actively participates in public debates on rural and agricultural issues, combining academic work with policy analysis. Additionally, he is involved in editorial activities and expert panels related to rural studies. He is the author of approximately 150 scientific publications on rural issues.

Kumar Parmil is a Professor at the Department of Statistics, Faculty of Mathematical Sciences, University of Jammu, India. His research interests are Bayesian inference, information theory applied statistics and ornithology. Professor Parmil has published more than 80 research papers in international/national journals and conferences. He has also published book chapters and popular science articles. Professor Parmil is an active member of many scientific professional bodies.

Łuczak Aleksandra is an Associate Professor at the Department of Finance and Accounting, Faculty of Economics, Poznań University of Life Sciences, Poland. Her primary research interests focus on multicriteria quantitative methods and their applications in economics and finance. She has a particular interest in taxonomic methods and decision-making methods, especially their applications in solving problems related to local and regional development planning. In recent years, she has also been involved in research on methods for measuring both objective and subjective poverty as a multidimensional phenomenon. She is the author and co-author of over 100 scientific publications. Additionally, she is a member of the Polish Statistical Association and serves on the editorial team of the Journal of Agribusiness and Rural Development.

Modibbo Umar Muhammad is a Senior Lecturer at the Modibbo Adama University, Yola, Nigeria. He received a Ph.D. in Operations Research at the Aligarh Muslim University, Aligarh, India in 2021. He obtained his Master of Technology (M.Tech) and Bachelor of Technology (B.Tech) degrees in Operations Research at the Federal University of Technology, Yola, Nigeria (Now The Modibbo Adama University of Technology, Yola) in 2016 and 2010 respectively. Dr Modibbo is a recipient of the University grant to study M.Tech. Operations Research in 2014, a Nigerian Tertiary Education Trust Fund (TETFund) to study Ph.D. Operations Research in 2018. He is a recipient of a Young Researcher Award and Research Excellence Award from Institute of Scholars (InSc) India, 2020. His research areas include mathematical programming and its applications, soft computing, reliability optimization, fuzzy programming, multi-objective optimization, inventory & supply chain management, and sustainable development goals. He has supervised several PGD., M.Sc., and UG. students in operations research, statistics and information technology. He is a Fellow and President of Operations Research Institute for Decision Sciences & Analytics of Nigeria [ORIDSAN], an executive Member, African Federation of Operations Research Societies [AFROS], International Federation of Operational Research Societies [IFORS] and International Group on Reliability (Gnedenko e-forum). He has published more than 50 research articles in journals of national and international repute with over 1000 Google scholar citations, and attended many conferences and workshops in his domain area serving as guest speaker, scientific committee member, and session chair. He is a reviewer and Guest Editor of many national and international reputed journals.

Mohamed Suleiman Dahir received his B.Sc. in Mathematics from Benadir University, Mogadisho, Somalia, in 2016, and the M.Sc. in Statistics from Universiti Sains Malaysia, Penang, Malaysia, in 2019. Currently, he is actively engaged in pursuing a Ph.D. in Statistics at Universiti Sains Malaysia. His current work is concentrated on forecasting, financial time series modelling, and applied statistics, with a special emphasis on pioneering methods for predicting cryptocurrencies. He also continues to have a strong interest in the detection of anomalies and breaks in statistical analysis.

Mijinyawa Mohammed is a current Ph.D. research fellow in Operations Research at the Modibbo Adama University, Yola, Nigeria. With more than a decade in research teaching and supervision and mentoring for both undergraduate and postgraduate students, Mijinyawa has been a member of Operational Society UK Student Membership, Associate Fellow Institute for Operations research of Nigeria (INFORN now ORIDSAN) published numerous journal papers,

Nkomo Wilbert is a Ph.D. Statistics student at Botswana International University of Science and Technology. Wilbert's research interests encompass distribution theory, reliability theory, statistical inference, survival analysis, and statistical modelling. He has published two research papers in international journals.

Oluyede Broderick is a Full Professor of Mathematics and Statistics, and former Director of the Statistical Consulting Unit (SCU) in the Department of Mathematical Sciences, Georgia Southern University, Statesboro, Georgia, USA. He is currently a Full Professor of Mathematics and Statistics at Botswana International University of Science and Technology (BIUST). He has over thirty years of research and teaching experience at Bowling Green State University, Georgia State University, University of Georgia, Oklahoma State University, Georgia Southern University and BIUST. He has authored and co-authored over two hundred (200) research papers, book chapters and conference proceedings. His research interests include multivariate statistical analysis, distribution theory, reliability theory, survival analysis, categorical data analysis, biostatistics, order restricted inference, stochastic dependence and weighted distributions. He is a member of several scientific bodies, and a member of two journal editorial boards.

Sankaran P. G. is a Senior Professor at the Department of Statistics, Cochin University of Science and Technology, India. He was Chair of the department and Pro-Vice Chancellor and Vice Chancellor of the University. His areas of interest include survival analysis, reliability theory, statistical inference, and distribution theory. Prof. Sankaran published more than 190 research papers in reputed journals. He wrote four books and a few technical reports. He is an active member of several professional bodies of statistics such as the International Statistical Institute and Indian Society for Probability and Statistics. Prof. Sankaran received many awards including BOYSCAST fellowship of the Department of Science and Technology, Government of India.

Seknewna Lema Logamou is a statistician and data scientist, currently a researcher and Senior Data Scientist at the African Institute for Mathematical Sciences Research Center (AIMS RIC). In this role, he works on projects focused on the impact of global warming on air pollution and its consequences on respiratory diseases. In parallel, he acts as a consultant at the National Institute of Statistics of Rwanda (NISR), in partnership with AIMS RIC. His main role is to strengthen the data science capacities

of the Institute's staff and interns, thus contributing to the development of the analytical and technical skills essential to addressing statistical and societal challenges.

Sharma Hemani is research scholar at the Department of Statistics, University of Jammu, India. Her research interests are Bayesian Inference, Information Theory, Reliability and Censored Models. She is a Gold medalist and has been awarded DST INSPIRE Fellowship. She has published several articles in international journals.

Skrodzka Iwona is an Associate Professor at the Division of Quantitative Methods, Faculty of Economics and Finance, University of Bialystok. Her research interests are structural equation modeling CB-SEM and PLS-SEM as well as multivariate statistical analysis. Professor Skrodzka has published more than 40 research papers in international/national journals and conferences. She is a member of two editorial boards: Facta Universitatis, Series: Economics and Organization (University of Niš, Serbia) and Optimum. Economic Studies (University of Bialystok, Poland).

Torsen Emmanuel is a Senior Lecturer at the Modibbo Adama University, Yola, Nigeria. Deeply involved in research teaching and supervision (both undergraduate and postgraduate students. Dr. Torsen has published over 40 papers, conference proceedings and book chapters. He is a reviewer of several international journals in the fields of statistics. His main research interest include time series analysis, financial time series analysis, non-parametric methods, risk measurement, and bio-statistics.

Wywiał Janusz Leszek is a Full Professor of Statistics in the Department of Statistics, Econometrics and Mathematics of the University of Economics in Katowice. His interests are mainly focused on survey sampling, especially problems of optimization of sampling designs and strategies dependent on auxiliary variables, statistical methods in auditing and multivariate analysis. He has published more than 130 research papers in international/national journals and conferences. Moreover, he is the author or co-author of 8 monographs and 8 textbooks. He is member of many scientific bodies.

Xu Zheng is an Assistant Professor at the Department of Mathematics and Statistics, Wright State University. His research interests include biostatistics, bioinformatics, machine learning, non-parametric methods, statistical computation, time series and econometrics. Professor Xu has published 70 research papers in international/national journals. Currently he is an active member of multiple editorial boards: Statistical Papers, Statistical Analysis and Data Mining, Journal of Statistical Computation and Simulation, and BioMed Central (BMC) Research Notes.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* https://sit.stat.gov.pl/ForAuthors.

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **Bold**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper.
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, (**1.1.**, **1.2.** …), **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).