Volume 24, Number 4, September 2023 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

Yousof H. M., Ali M. M., Aidi K., Ibrahim M., The modified Bagdonavičius-Nikulin goodness-of-fit test statistic for the right censored distributional validation with applications in medicine and reliability

Kiani S. K., Aslam M., Bhatti M. I., Investigation of half-normal model using informative priors under Bayesian structure

Landmesser-Rusek J., Dudek H., What explains the differences in material deprivation between rural and urban areas in Poland before and during the COVID-19 pandemic?

Patra D., Pal S., Chaudhuri A., Respondent-specific randomized response technique to estimate sensitive proportion

ISSN 1234-7655

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, *September 2023*
Vol. 24, No. 4, pp. III

Submission information for Authors

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

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

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

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

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

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

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

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

STATISTICS IN TRANSITION new series, *September 2023*
Vol. 24, No. 4, pp. V–VI

Policy Statement

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

***

Abstracting and Indexing Databases

Statistics in Transition new series is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

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

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Electronic Journals Library

Elsevier – Scopus

ERIH PLUS (European Reference Index for the Humanities and Social Sciences)

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, September 2023
Vol. 24, No. 4, pp. VII–XI

From the Editor

With the current issue of Statistics in Transition new series *(SiTns)* – one hundred and first since the beginning of its publication (not counting special editions) – we are entering a new decade of the journal's mission carried out so far for 30 years, consisting in promoting the development of statistical sciences while contributing to the scientific life of the broadly conceived international statistical community: authors, reviewers and readers of SiTns.

The September issue of SiTns consists of a collection of twelve articles written by twenty-eight authors from eleven countries, in order of appearance: Egypt, USA, Algeria, Pakistan, Australia, Poland, India, Iran, Nigeria, Thailand and Canada. There are eight articles in the Original Scientific Articles section, two articles in Other Articles section, and two in Research Communicates and Letters section.

Research articles

In the first paper, ***The modified Bagdonavičius-Nikulin goodness-of-fit test statistic for the right censored distributional validation with applications in medicine and reliability***, **Haitham M. Yousof**, **M. Masoom Ali**, **Khaoula Aidi**, and **Mohamed Ibrahim** discuss the modified test statistic ($\mathcal{TT}rr$, $\varepsilon\varepsilon 2$) for validation under the right censored cas. Simulation *via* Barzilai-Borwein algorithm is performed for assessing the right-censorship estimation method. Next, four right censored data sets are analyzed under the new modified test statistic for checking the distributional validation. It is shown that according to the modified Bagdonavičius-Nikulin goodness-of-fit test statistic, the new odd log-logistic inverted Weibull model can be used in modelling the censored medicine and reliability real data sets.

Sania Khawar Kiani's, **Muhammad Aslam's**, and **M. Ishaq Bhatti's** paper entitled ***Investigation of half-normal model using informative priors under Bayesian structure*** describes properties of half-normal distribution using informative priors under the Bayesian criterion. It employs the squared root inverted gamma, Chi-square and Rayleigh distributions as the prior distribution to construct the posterior distributions of the respective distributional parameters. Hyperparameters are elicited via prior predictive distribution. The properties of the posterior distribution are studied, and their graphs are presented using a real data set. A comprehensive simulation scheme is

conducted using informative priors. Bayes estimates are obtained using the loss functions (squared error loss function, modified loss function, quadratic loss function and DeGroot loss function). By the comparison of results, with increasing the sample size, the Bayes estimates converge to the parametric values and their risks tend to be smaller.

In the next article, ***What explains the differences in material deprivation between rural and urban areas in Poland before and during the COVID-19 pandemic?*** **Joanna Landmesser-Rusek** and **Hanna Dudek** examine the relationships between the compositional changes in demographic and socioeconomic factors and the changes in the prevalence of material deprivation in rural and urban areas in Poland. Using the European Union Statistics on Income and Living Conditions (EU-SILC) data for 2019–2020, the authors applied the Fairlie decomposition approach for a logit model. Six items of material deprivation analysing each symptom (item) as a binary variable were considered. Separate models were evaluated for each symptom. It was found that the important characteristics affecting a gap in material deprivation between rural and urban areas are: household equivalized income, the level of education, the type of household, and the presence of disabled or unemployed persons in the household. A non-significant effect of the pandemic on the material deprivation gap between rural and urban areas were observed.

Dipika Patra, **Sanghamitra Pal**, and **Arijit Chaudhuri** in the paper ***Respondent-specific randomized response technique to estimate sensitive proportion*** focus on Randomized Response Techniques and present a more general procedure using five different types of cards. A respondent-specific randomized response technique is also proposed, in which respondents are allowed to build up the boxes according to their own choices. An immediate objective for this change is to enhance the sense of protection of privacy of the respondents. But as by-products higher efficiency in terms of actual coverage percentages of confidence intervals and related features are demonstrated by a simulation study and superior jeopardy levels against divulgence of personal secrecy are also reported to be achievable. The findings described in this study will stimulate researchers and survey practitioners to apply the response-specific RRT in real surveys. Respondents will co-operate freely in the survey methods as they are building their own RR devices.

In the next paper, ***Odd log-logistic generalised Lindley distribution with properties and applications*** **Vahid Ranjbar**, **Abbas Eftekharian**, **Omid Kharazmi,** and **Morad Alizadeh** introduce a new three-parameter lifetime model, called the odd log-logistic generalised Lindley (OLLGL) distribution. The statistical properties of the OLLGL distribution including the hazard function, quantile function, moments, incomplete moments, generating functions, mean deviations and maximum likelihood estimation

for the model parameters are given. The new density function can be expressed as a linear mixture of exponentiated. Different methods are discussed to estimate the model parameters. Simulation studies were conducted to examine the performance of this distribution. The importance and flexibility of the new model are also illustrated empirically by means of two real-life data sets. Finally, Bayesian analysis and Gibbs sampling are performed based on the two data sets.

Rasaki Olawale Olanrewaju's, Sodiq Adejare Olanrewaju's, and Omodolapo Waliyat Isamot's article ***Hyper-parametric Generalized Autoregressive Scores (GASs): an application to the price of United States cooking gas*** outlines the framework of the Generalized Autoregressive Score (GAS) model with a variety of symmetric conditional densities of different time-varying hyper-parameters. The distinctive trait and goal of the observation-driven GAS model is to use its score and information functions as the compeller of time-variation via hyper-parameters of conditional densities. The score and Hessian functions (via location, scale, skewness, and shapes parameters) are of paramount interest due to their capability to curtail the lacuna of heaviness in the tail of normal distribution and possibility of skewed observations. Due to the flexibility of the GAS model to several statistical distributions, an empirical application to financial data of the price of the United State cooking gas was subjected to the GAS model with ten different conditional densities. Each of the conditional density subjected to the GAS model via the application of the price of cooking gas from 2005 to 2020 was driven by the mechanism of time-varying score and Hessian functions of their embedded hyper-parameters.

In the article entitled ***Testing the annual rainfall dispersion in Chaiyaphum Thailand, by using confidence intervals for the coefficient of variation of an inverse gamma distribution*** **Wararit Panichkitkosolkul** proposes two statistics for testing the CV of an IG distribution based on the Score and Wald methods. An evaluation of their performances is made using the Monte Carlo simulations conducted under several shape parameter values for an IG distribution based on empirical type I error rates and powers of the tests. The simulation results reveal that the Wald-method test statistic performed better than the Score-method one in terms of the attained nominal significance level, and is thus recommended for analysis in a similar context. Furthermore, the efficacy of the proposed test statistics was illustrated by applying them to the annual rainfall amounts in Chaiyaphum. The researchers can apply the proposed methods for testing the population CV in an IG distribution with other data sets fitted well to an IG distribution. For example, the IG distribution has been used for the hitting time distribution of a Wiener process. Future research could focus on the one-tailed hypothesis testing.

Kumari Priyanka's and **Pidugu Trisandhya's** paper ***Advances in estimation by the item sum technique in two move successive sampling*** contains a proposal of an estimator for the estimation of dynamic sensitive population mean using the Item Sum Technique (IST) and non-sensitive auxiliary information in the two-move successive sampling. Possible allocation designs for allocating long-list and short-list samples pertaining to the IST have been elaborated. The comparison between various allocation designs has been carried out. Theoretical considerations have been integrated with numerical as well as simulation studies to show the working version of the proposed IST estimators in the two-move successive sampling. It was concluded that IST is an alternative technique to deal with sensitive issues in successive sampling. In IST setup, the estimator utilizing additional auxiliary variable is proved to be more efficient than the estimator in which no additional auxiliary variable is used. Out of the two allocation designs for allocating LL and SL samples, the IST class of estimators using optimum allocation design is coming out to be more efficient than the estimator using general allocation deign.

Other articles

XXXI Scientific Conference of the Classification and Data Analysis Section (SKAD 2022)

Aneta Ptak-Chmielewska and **Agnieszka Chłoń-Domińczak** present ***Analysis of social and economic conditions of microenterprises based on taxonomy methods.*** In the article, a unique data set of the situation of SMEs in the Kujawsko-Pomorskie region was used to assess the changes of the characteristics of the microenterprise sector at the local level in Poland between 2019 and 2020, that is during the first years of the COVID pandemic. The analysis shows that there are visible changes in the microenterprise sector and the economic conditions under which microenterprises operated. In the largest clusters of gminas, there is a drop in the number of microenterprises per 10 000 population. There is also a significant decline in average revenues reported to tax authorities. This data is consistent with other national statistics, and also with observations at the European level on the drop in revenues and financial situation as one of the most important risks faced by the SME sector.

Dorota Żebrowska-Suchodolska discusses ***Elimination of characteristics concerning the performance of open-ended equity funds using PCA.*** The aim of the research was to apply principal component analysis (PCA) to reduce the dimension of the indicators that help the investor in selecting a fund, and to find the main factors determining the choice of an appropriate investment fund in terms of its performance and risk. The subject of the study was 15 equity funds that had been on the Polish market for many years. The research showed that it is possible to reduce the primary

variables to two dimensions. 13 groups were selected for the study. The groups were selected in terms of correlation of indicators. They contained from 1 to 10 indicators. The pairs of indicators included in the principal components have been placed in other parts of the circle, allowing the investor to assess the fund from the point of view of completely different information. The resulting indicators found in each group are based on a combination of classical and non-classical measures.

Research Communicates and Letters

Yong You's paper entitled ***An empirical study of hierarchical Bayes small area estimators using different priors for model variances*** describes hierarchical Bayes (HB) estimators based on different priors for small area estimation. In particular, the inverse gamma and flat priors for variance components in the HB small area models of You and Chapman (2006) and You (2021) were used. The authors evaluate the HB estimators through a simulation study and real data analysis. The results indicate that using the inverse gamma prior for the variance components in the HB models can be very effective. The simulation study and real data analysis demonstrate that proper IG prior should be used in the HB small area models for variance components. For future work, informative priors such as IG prior with parameter values based on previous survey data could also be used in the model to improve the HB small area estimators.

In the last paper, **Sławomir Dorocki** and **Mariusz Cembruch-Nowakowski** consider ***Application of statistical methods in socio-economic geography and spatial management based on selected scientific journals listed in the Web of Sciences database.*** The authors present an analysis of the use of statistical methods and tools in scientific articles related to socio-economic geography and spatial management published in the years 2012–2021. They focused on papers published in three selected journals relating to social geography (Geoforum), economic geography (Applied Geography) and spatial management (Landscape and Urban Planning). There is no doubt that conducting research with the support of statistical methods increases the credibility and reliability of their results as well as ensures the correctness of inference. This is particularly important for the analysis of spatial phenomena, which is becoming more and more complex. The conclusions presented in the text are based on the analysis of the representative but relatively small sample of the literature resources available.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 1–18, https://doi.org/10.59170/stattrans-2023-048
Received – 18.06.2021; accepted – 09.02.2023

The modified Bagdonavičius-Nikulin goodness-of-fit test statistic for the right censored distributional validation with applications in medicine and reliability

Haitham M. Yousof[1], M. Masoom Ali[2], Khaoula Aidi[3], Mohamed Ibrahim[4]

Abstract

A modified version of Bagdonavičius-Nikulin goodness-of-fit statistical test is presented for validation under the right censor case. Simulation via Barzilai-Borwein algorithm is performed for assessing the right-censorship estimation method. Four right censored data sets are analyzed under the new modified test statistic for checking the distributional validation.

Key words: inverted Weibull distribution, censored validation, Bagdonavičius-Nikulin, goodness-of-fit testing

1. Introduction

The Nikulin-Rao-Robson statistic, which is based on the differences between two estimators of the probabilities to fall into grouping intervals, is a well-known modification of the classical chi-squared tests in the case of complete data. One estimator is based on the empirical distribution function, and the other on maximum likelihood estimators of unknown parameters of the tested model using initial non-grouped data (see Nikulin (1973a), Nikulin (1973b), Nikulin (1973c) and Rao and Robson (1974) for more details and Goual et al. (2019), Goual and Yousof (2020a), Goual and Yousof (2020b), Yousof et al. (2021c) for more relevant applications).

However, methods for testing the censored validity of parametric distributions are in increasing development, but the presence of censorship makes them unavailable.

[1] Department of Statistics, Mathematics and Insurance, Benha University, Benha 13518, Egypt. E-mail: haitham.yousof@fcom.bu.edu.eg. ORCID: https://orcid.org/ 0000-0003-4589-4944.

[2] Department of Mathematical Sciences Ball State University, Muncie, Indiana 47306, USA. E-mail: mali@bsu.edu. ORCID: https://orcid.org/ 0000-0002-0120-9442.

[3] Laboratory of probability and statistics LaPS, University Badji Mokhtar, Annaba, Algeria. E-mail: khaoula.aidi@yahoo.fr. ORCID: https://orcid.org/ 0000-0002-4756-3690.

[4] Department of Applied, Mathematical and Actuarial Statistics, Faculty of Commerce, Damietta University, Damietta, Egypt. E-mail: mohamed_ibrahim@du.edu.eg. ORCID: https://orcid.org/ 0000-0003-4893-9669.

Habib and Thomas (1986) and Hollander and Pena (1992) proposed a modified Chi-squared test for randomly censored data based on the well-known Kaplan–Meyer estimators. Galanova (2012) considered some nonparametric modifications to the Anderson–Darling statistic, Kolmogorov–Smirnov statistic and the Cramer-Von-Mises statistic for accelerate failure models. Bagdonavičius-Nikulin (2011a) presented a new Chi-squared goodness-of-fit test statistic for the right censored data. (see also Bagdonavičius and Nikulin (2011b)). The Chi-squared goodness-of-fit test statistic of Bagdonavičius-Nikulin is applied for distributional validation under the right-censorship case.

In this paper, a modified Chi-squared goodness-of-fit test statistic based on the Bagdonavičius-Nikulin test is presented and applied accordingly for validation under a new odd log-logistic inverted Weibull distribution using the right censor case. First, a simulation study under the right censor case via the Barzilai-Borwein (BB) algorithm is performed for assessing the right censored estimation method. Numerous domains of optimization have paid substantial attention to the Barzilai and Borwein gradient approach. This is as a result of its practical effectiveness, affordability of computing, and simplicity. This study establishes root-linear global convergence for the Barzilai and Borwein method for strictly convex quadratic problems given in infinite-dimensional Hilbert spaces using spectral analysis techniques. It is shown how these findings may be used to two optimization issues governed by partial differential equations.

Following Ravi and Gilbert (2009), Hamedani et al. (2023) and Emam et al. (2023) and using the BB algorithm, we generated N=10,000 with different size (q=15,25,50,130,350,500) from the new odd log-logistic inverted Weibull model using some carefully selected initial values. The mean square error (MSEs) are used for assessing the performance of the censored maximum likelihood. Second, the modified Bagdonavicius-Nikulin test is applied using four right censored real data sets for distributional validation. The following censored real data are considered:

I. Data of bone marrow transplant (38 patients).
II. Data of allogeneic bone marrow transplant where the Histocompatibility Leukocyte Antigen (HLA) homolog marrow was used to rebuild their immune systems (50 patients).
III. Lymphoma data (times up to death) (31 patients).
IV. Strength data of some cords having resisted for a determined time were studied (38 pieces).

The Bagdonavicius-Nikulin goodness-of-fit statistical test proved that the new model can be used as a suitable alternative for analyzing four right censored data sets.

In this context, we will mention some recent studies that applied or presented new, modified extensions to the Bagdonavicius-Nikulin goodness-of-fit test. It is worth noting that the browser for statistical literature on this subject (Bagdonavicius-Nikulin goodness-of-fit test) will not find many new Bagdonavicius-Nikulin goodness-of-fit extensions and will find few research studies that applied this test, especially since Bagdonavicius-Nikulin goodness-of-fit test has precise requirements and strict procedures and requires censored data. As is well known, it is not easy to obtain new censored data to apply to and highlight the importance of the new test. In the next few lines, we will review some recent studies that were concerned with applying this test to actual data subject to censorship from the right, with a summary of what each study provided separately.

Mansour et al. (2020a) applied Bagdonavicius-Nikulin goodness-of-fit test for a novel log-logistic model utilizing for distributional validation. For the "right censored" real data set of survival times, the modified test is used. The new test's components are all clearly deduced and presented. For testing the adaptability and significance of the new model under the unfiltered framework, three actual data applications are provided. For filtered validation, two more genuine data sets are examined.

Mansour et al. (2020b) suggested and implemented a modified Chi-square goodness-of-fit test for the Burr X Weibull model using the Bagdonavicius-Nikulin technique for the right censored validation. The appropriate censored real data set is subjected to the modified goodness-of-fit statistics test. The modified goodness-of-fit test recovers the information loss based on the censored maximum likelihood estimators on the initial data, while the grouped data follows the Chi-square distribution The components of the modified criterion tests are derived. Under the unfiltered approach, validation is an actual data application.

Recently, an adapted Chi-square type test for distributional validity with applications to right-censored reliability and medical data by Yousof et al. (2021a). In this study, A modified version of the Bagdonaviius-Nikulin goodness-of-fit test statistic, also known as the modified Chi-square goodness-of-fit test, is researched, and then used for distributional validation in the appropriate censored scenario. The updated goodness-of-fit test is introduced and used with the appropriate censored data sets. Through a thorough simulation analysis, the censored Barzilai-Borwein algorithm is used to evaluate the new test's validity. Four actual and right censored data sets are subjected to the modified Bagdonaviius-Nikulin test. The new modified Bagdonaviius-Nikulin goodness-of-fit test statistic is used to compare a new distribution against a large number of other competing distributions.

The validity of the Bagdonaviius and Nikulin goodness-of-fit test statistic for the right censor case under the double Burr type X distribution is shown in a new updated

form. The method of maximum likelihood estimation in the case of censored data is used and implemented. Simulations using the Barzilai-Borwein algorithm are run to determine the best censored estimation technique. For the purpose of testing the null hypothesis, another simulation study is provided using a modified version of the statistical goodness-of-fit test developed by Bagdonaviius and Nikulin. For the purpose of examining the distributional validity, four right censored data sets are examined using the new modified test statistic (see Aidi et al. (2021)).

Finally, Ibrahim et al. (2022) presented a novel modified version of the Bagdonavicius and Nikulin goodness-of-fit statistical test, and its distributional validity is examined for both the right censor case and the exponentiated Rayleigh generalized Chen distribution. Simulations using the Barzilai-Borwein algorithm are run to determine the best censored estimating technique. For the purpose of examining the distributional validity, four right censored data sets are examined using the new modified test statistic. For more details, information, applications, and new extensions of this test in the case of censored data from the right, see: Yousof et al. (2021b) (for a new parametric lifetime model along with modified Chi-square type test for right censored distributional validation, characteristics and many estimation methods), Ibrahim et al. (2021) (for a new exponential generalized log-logistic model with the Bagdonavičius and Nikulin testing for distribution validation and some non-Bayesian estimation methods), see also Ibrahim et al. (2019 and 2020) and Yadav et al. (2022) for some related details about the Nikulin-Rao-Robson goodness-of-fit test.

2. Censored distributional validation

2.1. Maximum likelihood censored data

Consider the new odd log-logistic (NOLL) family (Cordeiro et al. (2016)). Then, for the inverted Weibull (IW) baseline model, the probability density function (PDF) of the new odd log-logistic inverted Weibull (NOLLIW) model can be derived as

$$f_{\underline{\boldsymbol{V}}}(x) = \frac{\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2\boldsymbol{\varsigma}_3 x^{-(\varsigma_3+1)}\, exp[-\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2 x^{-\varsigma_3}]\{1-exp[-\boldsymbol{\varsigma}_2 x^{-\varsigma_3}]\}^{\varsigma_1-1}}{(exp[-\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2 x^{-\varsigma_3}]+\{1-exp[-\boldsymbol{\varsigma}_2 x^{-\varsigma_3}]\}^{\varsigma_1})^2}, x>0, \quad (1)$$

where $\underline{\boldsymbol{V}} = (\boldsymbol{\varsigma}_1, \boldsymbol{\varsigma}_2, \boldsymbol{\varsigma}_3)$, $\boldsymbol{\varsigma}_1 > 0$, $\boldsymbol{\varsigma}_2 > 0, \boldsymbol{\varsigma}_3 > 0$ are three shape parameters. The survival function $S_{\underline{\boldsymbol{V}}}(x_i)$ can be written as

$$S_{\underline{\boldsymbol{V}}}(x) = 1 - \frac{exp(-\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2 x^{-\varsigma_3})}{exp(-\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2 x^{-\varsigma_3}) + [1-exp(-\boldsymbol{\varsigma}_2 x^{-\varsigma_3})]^{\varsigma_1}}. \quad (2)$$

In reliability studies and survival analysis, data are often censored. If $X_1, X_2, \ldots, X_{\boldsymbol{q}}$ is a censored sample from the NOLLIW distribution, each observation can be written as

$$x_i = min(X_i, \boldsymbol{C}_i)|_{(i=1,..,\boldsymbol{q})},$$

where X_i are failure times and $\boldsymbol{C}_i$ censoring times. Using (1) and (2), the log likelihood function is

$$L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})) = \log\left[\prod_{i=1}^{\boldsymbol{q}} f_{\underline{\boldsymbol{V}}}(x_i)^{\boldsymbol{\Delta}_i} S_{\underline{\boldsymbol{V}}}(x_i)^{1-\boldsymbol{\Delta}_i}\right]|_{(\boldsymbol{\Delta}_i=1_{X_i<\boldsymbol{c}_i})},$$

which can then be can be written as

$$L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}}) = \sum_{i=1}^{\boldsymbol{q}} \boldsymbol{\Delta}_i \log f\,(x_i) + \sum_{i=1}^{\boldsymbol{q}} (1-\boldsymbol{\Delta}_i) \log S\,(x_i)\,|_{(\boldsymbol{\Delta}_i=1_{X_i<\boldsymbol{c}_i})}. \tag{3}$$

Let $d_i = exp\left(-\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3}\right)$, and $\delta_i = 1 - exp\left(-\boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3}\right)$. Then,

$$L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}}) = \sum_{i=1}^{\boldsymbol{q}} \boldsymbol{\Delta}_i \begin{bmatrix} log(\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2\boldsymbol{\varsigma}_3) - (\boldsymbol{\varsigma}_3 - 1)\, log\, x_i \\ -\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3} + (\boldsymbol{\varsigma}_1 - 1)\, log\, \delta_i - 2\, log(\, d_i + \delta_i) \end{bmatrix}$$
$$+ \sum_{i=1}^{\boldsymbol{q}} (1-\boldsymbol{\Delta}_i)[\boldsymbol{\varsigma}_1\, log\, \delta_i - log(\, d_i + \delta_i)].$$

The MLEs for parameters $\boldsymbol{\varsigma}_1, \boldsymbol{\varsigma}_2$ and $\boldsymbol{\varsigma}_3$ are derived from solving the following nonlinear system of $\frac{\partial L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})}{\partial \boldsymbol{\varsigma}_1}$, $\frac{\partial L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})}{\partial \boldsymbol{\varsigma}_2}$ and $\frac{\partial L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})}{\partial \boldsymbol{\varsigma}_3}$ where

$$\frac{\partial L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})}{\partial \boldsymbol{\varsigma}_1} = \sum_{i=1}^{\boldsymbol{q}} \left[\frac{1}{\boldsymbol{\varsigma}_1} - \boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3} + log\, \delta_i + \frac{2\left(d_i \boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3} - \delta_i^{\boldsymbol{\varsigma}_1}\, log\, \delta_i\right)}{d_i + \delta_i}\right]$$
$$+ \sum_{i=1}^{\boldsymbol{q}} (1-\boldsymbol{\Delta}_i)\left[log\, \delta_i + \frac{\boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3} d_i - \delta_i^{\boldsymbol{\varsigma}_1}\, log\, \delta_i}{d_i + \delta_i}\right],$$

$$\frac{\partial L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})}{\partial \boldsymbol{\varsigma}_2} = \sum_{i=1}^{\boldsymbol{q}} \boldsymbol{\Delta}_i \left[\frac{1}{\boldsymbol{\varsigma}_2} - \boldsymbol{\varsigma}_1 x_i^{-\boldsymbol{\varsigma}_3} + (\boldsymbol{\varsigma}_1 - 1)\frac{x_i^{-\boldsymbol{\varsigma}_3}(1-\delta_i)}{\delta_i} \right.$$
$$\left. + \frac{2\begin{pmatrix} \boldsymbol{\varsigma}_1 x_i^{-\boldsymbol{\varsigma}_3} d_i \\ -\boldsymbol{\varsigma}_1 x_i^{-\boldsymbol{\varsigma}_3}(1-\delta_i)\delta_i^{\boldsymbol{\varsigma}_1 - 1} \end{pmatrix}}{d_i + \delta_i}\right]$$
$$+ \sum_{i=1}^{\boldsymbol{q}} (1-\boldsymbol{\Delta}_i)\left[\begin{matrix} \frac{\boldsymbol{\varsigma}_1 x_i^{-\boldsymbol{\varsigma}_3}(1-\delta_i)}{\delta_i} \\ + \frac{\boldsymbol{\varsigma}_1 x_i^{-\boldsymbol{\varsigma}_3} d_i - \boldsymbol{\varsigma}_1 x_i^{-\boldsymbol{\varsigma}_3}(1-\delta_i)\delta_i^{\boldsymbol{\varsigma}_1 - 1}}{d_i + \delta_i} \end{matrix}\right],$$

and

$$\frac{\partial L_{i,q}(\underline{\boldsymbol{V}})}{\partial \boldsymbol{\varsigma}_3} \sum_{i=1}^{q} \boldsymbol{\Delta}_i \left[\frac{1}{\boldsymbol{\varsigma}_3} + \left[\boldsymbol{\varsigma}_1 \boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3} - 1\right] log\, x_i - (\boldsymbol{\varsigma}_1 - 1) \frac{\boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3}\, log\, x_i\, exp[-\boldsymbol{\varsigma}_1 \boldsymbol{\varsigma}_2 x^{-\boldsymbol{\varsigma}_3}]}{\delta_i} \right.$$
$$\left. - \frac{2\left(\boldsymbol{\varsigma}_1 \boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3}\, log\, x_i\, d_i - \boldsymbol{\varsigma}_1 \boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3}\, log\, x_i\, (1-\delta_i)\delta_i^{\boldsymbol{\varsigma}_1 - 1}\right)}{d_i + \delta_i} \right]$$
$$- \sum_{i=1}^{q} (1 - \boldsymbol{\Delta}_i) \left[\frac{\boldsymbol{\varsigma}_1 \boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3}\, log\, x_i\, exp[-\boldsymbol{\varsigma}_1 \boldsymbol{\varsigma}_2 x^{-\boldsymbol{\varsigma}_3}]}{\delta_i} \right.$$
$$\left. + \frac{\left(\boldsymbol{\varsigma}_1 \boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3}\, log\, x_i\, d_i - \boldsymbol{\varsigma}_1 \boldsymbol{\varsigma}_2 x_i^{-\boldsymbol{\varsigma}_3}\, log\, x_i\, (1-\delta_i)\delta_i^{\boldsymbol{\varsigma}_1 - 1}\right)}{d_i + \delta_i} \right].$$

The explicit form of $\hat{\boldsymbol{\varsigma}}_1$, $\hat{\boldsymbol{\varsigma}}_2$ and $\hat{\boldsymbol{\varsigma}}_3$ cannot be obtained, so we use numerical methods.

2.2. Test criteria for the new model

Let $X_1, X_2, \ldots, X_{\boldsymbol{q}}$ be grouped in r sub-intervals $I_1, I_2, \cdots, I_r$ as

$$I_{\boldsymbol{j}} =]a_{(\boldsymbol{j}-1)}; a_{(\boldsymbol{j})}] \mid_{(\boldsymbol{j}=1,2,\cdots,r)},$$

which are mutually disjoint. The limits $a_{(\cdot)}$ of the intervals $I_{\boldsymbol{j}}$ are calculated such that:

$$\hat{\rho}_{\boldsymbol{j}} = \hat{\rho}_{\boldsymbol{j}}(\underline{\boldsymbol{V}}) = \int_{a_{(\boldsymbol{j}-1)}}^{a_{(\boldsymbol{j})}} f_{\underline{\boldsymbol{V}}}(x_i) dx, \tag{4}$$

$$a_{(\boldsymbol{j})} = F^{-1}\left(\frac{\boldsymbol{j}}{r}\right) \mid_{(\boldsymbol{j}=1,2,\cdots,r-1)},$$

and

$$0 < a_{(0)} < a_{(1)} < \ldots < a_{(\boldsymbol{j}-1)} < a_{(\boldsymbol{j})} < +\infty.$$

The new test statistic $\boldsymbol{T}_{r,\varepsilon}^2$ is defined by

$$\boldsymbol{T}_{r,\varepsilon}^2 = \sum_{\boldsymbol{j}=1}^{q} \frac{1}{\boldsymbol{U}_{\boldsymbol{j}}} \left(\boldsymbol{U}_{\boldsymbol{j}} - e_{\boldsymbol{j}}\right)^2 + \boldsymbol{Q}, \tag{5}$$

where $\boldsymbol{e}_{\boldsymbol{j}}$ is the number of expected failures (NEF) in the grouped intervals and $\boldsymbol{U}_{\boldsymbol{j}}$ is the number of observed failures (NOF) in grouping intervals where

$$\boldsymbol{Q} = \boldsymbol{W}^T \hat{\mathbf{G}}^{-} \boldsymbol{W},$$
$$\widehat{\boldsymbol{W}} = \hat{\boldsymbol{C}} \hat{\boldsymbol{A}}^{-1} \boldsymbol{Z} = \left(\widehat{\boldsymbol{W}}_1, \ldots, \widehat{\boldsymbol{W}}_s\right)^T,$$
$$\hat{\mathbf{G}} = [\hat{\mathbf{g}}_{\mathcal{L}\mathcal{L}'}]_{s \times s},$$

$$\boldsymbol{W}_{\mathcal{L}} = \sum_{j=1}^{r} \widehat{\boldsymbol{C}}_{\mathcal{L}j} \widehat{\boldsymbol{\mathcal{A}}}_{j}^{-1} \boldsymbol{Z}_{j},$$

$$\boldsymbol{Z}_{j} = \frac{1}{\sqrt{\boldsymbol{q}}} (\boldsymbol{U}_{j} - e_{j}),$$

$$\hat{\mathbf{g}}_{\mathcal{L}\mathcal{L}} = \hat{\imath}_{\mathcal{L}\mathcal{L}'} - \sum_{j=1}^{r} \widehat{\boldsymbol{C}}_{\mathcal{L}j} \widehat{\boldsymbol{C}}_{\mathcal{L}'j} \widehat{\boldsymbol{\mathcal{A}}}_{j}^{-1}, \; j = 1,2,\dots,r, i = 1,2,\dots,\boldsymbol{q},, \mathcal{L}, \mathcal{L}' = 1,2,\dots,s.$$

The elements of $\widehat{\boldsymbol{C}}$ defined by

$$\widehat{\boldsymbol{C}}_{\mathcal{L}j} = \frac{1}{\boldsymbol{q}} \sum_{i:x_i \in I_j}^{\boldsymbol{q}} \boldsymbol{\Delta}_{i} \frac{\partial}{\partial \widehat{\underline{\mathbf{V}}}_{\mathcal{L}}} \ln H_{\underline{\widehat{V}}}(x_i), \tag{6}$$

where $H_{\underline{\widehat{V}}}(x_i)$ refers to the cumulative hazard rate function (CHRF) of the NOLLIW distribution.

2.3. Test quadratic form of the modified criteria

The quadratic form of the modified test statistic can be written as

$$\boldsymbol{\mathcal{T}}_{r,\varepsilon}^{2} = \sum_{j=1}^{r} \frac{1}{\boldsymbol{u}_{j}} (\boldsymbol{U}_{j} - e_{j})^{2} + \widehat{\boldsymbol{W}}^{T} \left[\hat{\imath}_{ll'} - \sum_{j=1}^{r} \widehat{\boldsymbol{C}}_{lj} \widehat{\boldsymbol{C}}_{l'j} \widehat{\boldsymbol{\mathcal{A}}}_{j}^{-1}\right]^{-1} \widehat{\boldsymbol{W}}, \tag{7}$$

where matrices $\widehat{\boldsymbol{W}}$, $\widehat{\boldsymbol{C}}$ are defined in Abouelmagd et al. (2019a,b), Mansour et al. (2020a,b) and Yadav et al. (2020) with more details and $\hat{\mathbf{I}}$ is the estimated information matrix.

2.4. Simulations via the BB algorithm

In this subsection we are interested in solving the nonlinear system of equations

$$0 = \frac{\partial}{\partial \boldsymbol{\varsigma}_{k}} L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})|_{k=1,2,3}$$

where the three functions

$$\frac{\partial}{\partial \boldsymbol{\varsigma}_{k}} L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})|_{\mathbf{J}(p):\Re^{p}\times\Re^{p}\to\Re^{p}}$$

refer to nonlinear functions with continuous partial derivative. We are interested in situations where p is large, and where the Jacobian of $\frac{\partial}{\partial \boldsymbol{\varsigma}_{k}} L_{i,\boldsymbol{q}}(\underline{\boldsymbol{V}})$ is either unavailable or requires a prohibitive amount of storage. The Quasi-Newton methods can be used for obtaining an approximation of $\mathbf{J}(p)$, which, along with the vector of solutions, are

updated at each iteration. Using the R statistical software and the BB algorithm of Ravi and Gilbert (2009), we generated $N = 10{,}000$ with different size ($\boldsymbol{q} = 15,\ 25,\ 50,\ 130,\ 350,\ 500$) from the NOLLIW model using the initial values ($\boldsymbol{\varsigma}_1 = 1.5,\ \boldsymbol{\varsigma}_2 = 2,\ \boldsymbol{\varsigma}_3 = 1.3$). Firstly, we compute the MLEs and then the criteria $\boldsymbol{\mathcal{T}}^2_{r,\varepsilon}$. The results which are given in Table 1 is obtained by inverting the CDF of the new model by setting

$$U = F_{\underline{V}}(x) = \frac{exp(-\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2 x^{-\boldsymbol{\varsigma}_3})}{exp(-\boldsymbol{\varsigma}_1\boldsymbol{\varsigma}_2 x^{-\boldsymbol{\varsigma}_3})+[1-exp(-\boldsymbol{\varsigma}_2 x^{-\boldsymbol{\varsigma}_3})]^{\boldsymbol{\varsigma}_1}},$$

where U follows the uniform (0,1) model to obtain the quantile function. Since the quantile function of the NOLLIW model has no closed form, we use the numerical methods to generate the random numbers. Based on Table 1, it is seen that the MSE decreases as n increases. For example, we have the following results:

I. MSE ($\boldsymbol{\varsigma}_1$, $\boldsymbol{q} = 15,\ 25,\ 50,\ 130,\ 350,\ 500$) = (0.0077, 0.0059, 0.0040, 0.0026, 0.0017, 0.00009).
II. MSE ($\boldsymbol{\varsigma}_2$, $\boldsymbol{q} = 15,\ 25,\ 50,\ 130,\ 350,\ 500$) = (0.0115, 0.0092, 0.0083, 0.0075, 0.0058, 0.0029).
III. MSE ($\boldsymbol{\varsigma}_3$, $\boldsymbol{q} = 15,\ 25,\ 50,\ 130,\ 350,\ 500$) = (0.0088, 0.0068, 0.0042, 0.0039, 0.0022, 0.0017).

Table 1: MLEs of a_1, a_2, a_3 and MSE.

$\widehat{\boldsymbol{\varsigma}}_i$, MSEs↓ & n→	$\boldsymbol{q}_1$=15	$\boldsymbol{q}_2$=25	$\boldsymbol{q}_3$=50	$\boldsymbol{q}_4$=130	$\boldsymbol{q}_5$=350	$\boldsymbol{q}_6$=500
$\boldsymbol{\varsigma}_1$	1.4632	1.4696	1.4775	1.4854	1.4967	1.49980
MSE	0.0077	0.0059	0.0040	0.0026	0.0017	0.00009
$\boldsymbol{\varsigma}_2$	1.9442	1.9359	1.9372	1.9278	1.9187	1.9955
MSE	0.0115	0.0092	0.0083	0.0075	0.0058	0.0029
$\boldsymbol{\varsigma}_3$	1.3239	1.3251	1.3196	1.3113	1.3079	1.3016
MSE	0.0088	0.0068	0.0042	0.0039	0.0022	0.0017

2.5. Test statistic $\mathcal{T}_{r,\varepsilon}^2$ under censored simulations

For testing H_0 that a certain right censored data came from NOLLIW model, we compute $\mathcal{T}_{r,\varepsilon}^2$ for $N = 10{,}000$ simulated samples with $\boldsymbol{q} = 25, 50, 130, 350, 500$. Then, we calculate empirical levels of significance when $\mathcal{T}_{r,\varepsilon}^2 > \chi_{r-np,\varepsilon}^2$ corresponding to theoretical levels of significance ($\varepsilon = 1\%, 5\%, 10\%$), where np is the number of parameters. The results which are given in Table 2 is obtained by inverting the CDF of the new model by sitting $U = F_{\underline{V}}(x)$, where U follows the uniform (0,1) model to obtain the quantile. Since the quantile function has no closed form, we use the numerical methods to generate the random numbers. The generated samples are considered as a right censored data. Then using equation (5), the value of the new test ($\mathcal{T}_{r,\varepsilon}^2$) is calculated for $\boldsymbol{q} = 25, 50, 130, 350, 500$ under $\varepsilon = 1\%, 5\%, 10\%$. Finally the values of the $\mathcal{T}_{r,\varepsilon}^2$ test is compared with the corresponding significance level. The following results can be highlighted:

I. $\mathcal{T}_{r,\varepsilon}^2$ =0.0071, 0.0076, 0.0084, 0.0092, 0.0098< $\varepsilon_1 = 1\%$.

II. $\mathcal{T}_{r,\varepsilon}^2$ =0.0442, 0.0466, 0.0474, 0.0483, 0.0499< $\varepsilon_2 = 5\%$.

III. $\mathcal{T}_{r,\varepsilon}^2$ =0.0935, 0.0952, 0.0961, 0.0970, 0.0992< $\varepsilon_2 = 5\%$.

The results are reported in Table 2 which means that the test proposed can be employed for fitting data from NOLLIW version. Based on Table 2, the new model performs well under the new test.

Table 2: Simulated levels of significance for $\mathcal{T}_{r,\varepsilon}^2$ test for NOLLIW model.

N=10,000	$\boldsymbol{q}_1$=25	$\boldsymbol{q}_2$=50	$\boldsymbol{q}_3$=130	$\boldsymbol{q}_4$=350	$\boldsymbol{q}_5$=500
$\varepsilon_1 = 1\%$	0.0071	0.0076	0.0084	0.0092	0.0098
$\varepsilon_2 = 5\%$	0.0442	0.0466	0.0474	0.0483	0.0499
$\varepsilon_3 = 10\%$	0.0935	0.0952	0.0961	0.0970	0.0992

2.6. Censored real data analysis

In this section, we present three applications to distribute NOLLIW into three real data sets. First, we consider the bone marrow transplant data (data set **I**) (Klein and Moeschberger (2003)) for patients suffering from acute lymphoblastic leukemia. These data consists of time (in days) to death or on study time after allogenic bone marrow transplant for 38 patients. The bone marrow transplant is a standard treatment for acute leukemia. Recovery following bone marrow transplantation is a complex process. Immediately following transplantation, patients have depressed platelet counts and have higher hazard rate for the development of infections but as the time passes the

hazard decreases, where starred values denote to censored observations. Below is the time to death (in days) data after bone marrow transplant:
(1,86,107,110,122,156,162,172,194, 243,262, 262, 269,276, 350*, 371, 417, 418, 466,487,526,530*,716,781,996∗,1111*,1167*,1182*,1199∗,1 79, 1330*, 1377*, 1433*, 1462*,1496*, 1602*,2081*,226∗).

The second data set (data set **II**) consists of sample data from 50 patients with acute myeloid leukemia, reported to the International Register of Bone Marrow Transplants of John et al. (1997). These patients had an allogeneic bone marrow transplant where the HLA (Histocompatibility Leukocyte Antigen) homolog marrow was used to rebuild their immune systems. The data required for this study are shown below:
(0.030, 0.493, 0.855, 1.184, 1.283, 1.480, 1.776, 2.138, 2.5, 2.763, 2.993, 3.224, 3.421, 4.178, 4 .441*, 5.691, 5.855* ,6.941*, 6.941, 7.993*, 8.882, 8.882, 9.145*, 11.480, 11.513, 12.105*, 12.796, 12.993*, 13.849*, 16.612*, 17.138*, 20.066, 20.329*, 22.368*,26.776*, 28.717*,28.717* ,32.928*,33.783*,34.211*,34.770*, 39.539*, 41.118*, 45.033*, 46.053*, 46.941*, 48.289*, 57.401*, 58.322*, 60.625*).

The third data (data set **III**) are called the lymphoma data set and consisting of times (in months) from diagnosis stage up to death for 31 individuals with the advanced non-Hodgkin's lymphoma clinical symptoms. Among these 31 observations 11 of the times are censored, because those patients were still alive at the last time of follow-up:
(2.5, 4.1, 4.6, 6.4, 6.7, 7.4, 7.6, 7.7, 7.8, 8.8, 13.3, 13.4, 18.3, 19.7, 21.9, 24.7, 27.5, 29.7, 30.1*, 32.9, 33.5, 35.4*, 37.7*, 40.9*, 42.6*, 45.4*, 48.5*, 48.9*, 60.4*, 64.4*, 66.4*). The test statistic is used to verify if the lymphoma data can be modelled by NOLLIW distribution.

Finally, we consider the censored reliability data of (Crowder et al. (1991)). In their experiment, Crowder et al. (1991) obtained information about the strengths of a certain type of braided cord after the weather, the forces of 48 pieces of cord having resisted for a determined time were studied. The right censored force values observed are given below:
(26.8*, 29.6*, 33.4*, 35*, 36.3, 40*, 41.7, 41.9*, 42.5*, 43.9, 49.9, 50.1, 50.8,51.9, 52.1, 52.3, 52.3, 52.4, 52.6, 52.7, 53.1, 53.6, 53.6, 53.9, 53.9, 54.1, 54.6, 54.8, 54.8,55.1, 55.4, 55.9, 56, 56.1, 56.5, 56.9, 57.1, 57.1, 57.3, 57.7, 57.8, 58.1, 58.9, 59, 59.1, 59.6, 60.4, 60.7).

All results of this application are summarized in Table 3 and Table 4. Table 3 gives the results of $\hat{\rho}_j$, $\boldsymbol{U}_j$, $\widehat{\boldsymbol{C}}_{1j}, \widehat{\boldsymbol{C}}_{2j}, \widehat{\boldsymbol{C}}_{3j}$ and e_j for the four real data sets where $\hat{\rho}_j$, $\boldsymbol{U}_j$, $\widehat{\boldsymbol{C}}_{1j}, \widehat{\boldsymbol{C}}_{2j}, \widehat{\boldsymbol{C}}_{3j}$ and e_j are the main components of the modified test statistic. However, Table 4 gives the values of $\boldsymbol{T}_{r,\varepsilon}^2$ versus $\chi_{r,\varepsilon}^2$.

The values of $\boldsymbol{T}_{r,\varepsilon}^2$ in Table 4 are calculated based on the corresponding values obtained in Table 3. Since $\chi_{\varepsilon=\%5}^2 = 11.0705$, the four values of $\boldsymbol{T}_{r,\varepsilon}^2$ are 10.956, 7.6235,

6.8580 and 6.8956. These results show that the NOLLIW distribution can be used in modelling the four mentioned real data sets according to the modified Bagdonavičius -Nikulin goodness-of-fit test statistic for right censored validation.

Table 3: Values of $\hat{\rho}_j, \boldsymbol{u}_j, \widehat{\boldsymbol{C}}_{1j}, \widehat{\boldsymbol{C}}_{2j}, \widehat{\boldsymbol{C}}_{3j}$ and e_j.

Data set	$\hat{\rho}_j$	$\boldsymbol{u}_j$	$\widehat{\boldsymbol{C}}_{1j}$	$\widehat{\boldsymbol{C}}_{2j}$	$\widehat{\boldsymbol{C}}_{3j}$	e_j
I & r=5	185.5	8	1.0236	0.5632	1.8289	7.6
	320.5	7	0.9532	0.2351	1.7421	7.6
	510.5	6	0.8124	-1.523	1.0231	7.6
	1220.5	9	1.1526	0.6310	2.0513	7.6
	2081	8	1.0856	0.5231	1.9045	7.6
II & r=5	1.923	7	1.3526	0.9238	2.1235	10
	8.562	13	2.0956	1.6485	2.9425	10
	16.432	9	1.4526	0.8231	2.4513	10
	34.526	11	1.748	1.5237	2.7412	10
	60.625	10	1.6245	1.4032	2.6143	10
III & r=4	7.500	6	0.9352	0.5631	0.5417	7.75
	15.65	6	0.4587	-0.4581	0.4689	7.75
	31.45	7	1.3026	-0.4956	0.7864	7.75
	66.40	12	0.3145	0.2031	0.2153	7.75
IV & r=5	43.5	9	0.9326	0.5326	1.4362	9.6
	52.9	11	1.1306	0.7541	1.5962	9.6
	55.2	10	1.0053	0.6138	1.4012	9.6
	57.2	8	0.7654	0.4319	1.2312	9.6
	60.7	10	1.0103	0.5322	1.3496	9.6

Table 4: The values of $\boldsymbol{\mathcal{T}}_{r,\varepsilon}^2$ and $\chi_{r-np,\varepsilon}^2$ for each data.

Data set	r	$\widehat{\mathbf{V}}$			$\boldsymbol{\mathcal{T}}_{r,\varepsilon}^2$ & $\chi_{r-np,\varepsilon}^2$	Rank
		$\hat{\varsigma}_1$	$\hat{\varsigma}_2$	$\hat{\varsigma}_3$		
I	5	1.8235	1.2856	1.5982	$10.956<\chi_{2,0.05}^2=11.0705$	3
II	5	1.5032	1.0203	1.1052	$11.936<\chi_{2,0.05}^2=11.0705$	4
III	4	1.6385	1.6230	1.2865	$7.2365<\chi_{1,0.05}^2= 9.4877$	1
IV	5	1.3746	0.8263	1.4256	$9.8569<\chi_{2,0.05}^2=11.0705$	2

Based on Table 3 and Table 4, we conclude that:

1. For the right censored data of bone marrow transplant which contains from 38 patients, it is seen that $\boldsymbol{\mathcal{T}}_{5,0.05}^2 = 10.956$ however $\chi_{2,0.05}^2$=11.0705. Since $\boldsymbol{\mathcal{T}}_{5,0.05}^2$ is less that $\chi_{2,0.05}^2$, we can say by accepting the null hypothesis that the bone marrow transplant data follow the odd log-logistic inverted Weibull distribution as well and that odd new log-logistic inverted Weibull distribution can be used and applied in modelling the bone marrow transplant data.
2. Data of allogeneic bone marrow transplant (50 patients), it is seen that $\boldsymbol{\mathcal{T}}_{5,0.05}^2 = 11.936$ however $\chi_{2,0.05}^2$=11.0705. Since $\boldsymbol{\mathcal{T}}_{5,0.05}^2$ is less that $\chi_{2,0.05}^2$, we can say by accepting the null hypothesis that the acute myeloid leukemia data follow the odd log-logistic inverted Weibull distribution as well and that new odd log-logistic inverted Weibull distribution can be used and applied in modelling the acute myeloid leukemia data.
3. Lymphoma data (times up to death) (31 patients), it is noted that $\boldsymbol{\mathcal{T}}_{5,0.05}^2 = 7.2365$ however $\chi_{2,0.05}^2$=9.4877. Which means that $\boldsymbol{\mathcal{T}}_{5,0.05}^2$ is less that $\chi_{2,0.05}^2$. Therefore, we can say by accepting the null hypothesis that the leukemia data follow the odd log-logistic inverted Weibull distribution as well and that odd log-logistic inverted Weibull distribution can be used and applied in modelling the leukemia data.
4. Strength data (38 pieces), it is concluded that $\boldsymbol{\mathcal{T}}_{5,0.05}^2 = 9.8569$ however $\chi_{2,0.05}^2$=11.0705). Hence, we can say by accepting the null hypothesis that the strength data follow the odd log-logistic inverted Weibull distribution as well and that odd log-logistic inverted Weibull distribution can be used and applied in modelling the strength data.

Table 5: Comparing the NOLLIW and the two-parameters Weibull under the $\mathcal{T}^2_{r,\varepsilon}$ test.

Data set	r	Models and testing results	Decision
		NOLLIW model	
I	5	$10.956<\chi^2_{2,0.05}=11.0705$	Accept H_0
II	5	$11.936<\chi^2_{2,0.05}=11.0705$	Accept H_0
III	4	$7.2365<\chi^2_{1,0.05}= 9.4877$	Accept H_0
IV	5	$9.8569<\chi^2_{2,0.05}=11.0705$	Accept H_0
		The two-parameters Weibull model	
I	5	$12.5362>\chi^2_{2,0.05}=11.0705$	Reject H_0
II	5	$13.4521 < \chi^2_{2,0.05}=11.0705$	Reject H_0
III	5	$10.5236<\chi^2_{2,0.05}=11.0705$	Accept H_0
IV	5	$9.5236<\chi^2_{2,0.05}=11.0705$	Accept H_0

Based on Table 5, we conclude that:

I. The bone marrow transplant data (data set **I**) follow the odd log-logistic inverted Weibull distribution. However, these data does not follow the two-parameters Weibull model since $12.5362>\chi^2_{2,0.05}=11.0705$.

II. The acute myeloid leukemia data (data set **II**) also follow the odd log-logistic inverted Weibull distribution. However, these data does not follow the two-parameters Weibull model since $13.4521 > \chi^2_{2,0.05}=11.0705$.

III. The leukemia data (data set **III**) can be modelled using the odd log-logistic inverted Weibull distribution and the two-parameters Weibull model. However, the odd log-logistic inverted Weibull distribution is better than the two-parameters Weibull model since $7.2365<\chi^2_{1,0.05}= 9.4877$ for the odd log-logistic inverted Weibull distribution and $10.5236<\chi^2_{2,0.05}=11.0705$ for two-parameters Weibull model.

IV. The strengths data (data set **IV**) can be modelled using the odd log-logistic inverted Weibull distribution and the two-parameters Weibull model. However, the odd log-logistic inverted Weibull distribution is better than the two-parameters Weibull model since $9.8569<\chi^2_{2,0.05}=11.0705$ for the odd log-logistic inverted Weibull distribution and $9.5236<\chi^2_{2,0.05}=11.0705$ for two-parameters Weibull model.

3. Concluding remarks

In this paper, a modified Bagdonavičius -Nikulin goodness-of-fit test statistic is presented and applied for distributional validation under the right censor case. The modified test statistic ($\boldsymbol{\mathcal{T}}^2_{r,\varepsilon}$) is given along with all its relevant components. Four right censored data sets are analyzed under the new modified test statistic for checking the distributional validation. According to the modified Bagdonavičius -Nikulin goodness-of-fit test statistic, the new odd log-logistic inverted Weibull model can be used in modelling the censored medicine and reliability real data sets. Below are the results of the modified test statistic under the right censor data sets:

I. Data of bone marrow transplant (38 patients):
$\boldsymbol{\mathcal{T}}^2_{5,0.05} = 10.956$ ($<\chi^2_{2,0.05}$=11.0705). By accepting the null hypothesis, we can conclude that the bone marrow transplant data also follow the odd log-logistic inverted Weibull distribution and that the bone marrow transplant data can be modelled using the odd new log-logistic inverted Weibull distribution.

II. Data of allogeneic bone marrow transplant (50 patients):
$\boldsymbol{\mathcal{T}}^2_{5,0.05} = 11.936$ ($<\chi^2_{2,0.05}$=11.0705). By accepting the null hypothesis, we can deduce that the acute myeloid leukemia data also follow the odd log-logistic inverted Weibull distribution, and that the acute myeloid leukemia data can be modelled through using new odd log-logistic inverse Weibull distribution.

III. Lymphoma data (times up to death) (31 patients):
$\boldsymbol{\mathcal{T}}^2_{5,0.05} = 7.2365$ ($<\chi^2_{1,0.05}$=9.4877). By accepting the null hypothesis, we can claim that the leukemia data also follow the odd log-logistic inverted Weibull distribution, and that the leukemia data can be modelled using the odd log-logistic inverted Weibull distribution.

IV. Strength's data (38 pieces):
$\boldsymbol{\mathcal{T}}^2_{5,0.05} = 9.8569$ ($<\chi^2_{2,0.05}$=11.0705). Therefore, if the null hypothesis is accepted, we can infer that the strength data also follow the odd log-logistic inverted Weibull distribution and that the odd log-logistic inverted Weibull distribution may be used to describe the right censored strength data.

V. The bone marrow transplant data follows the odd log-logistic inverted Weibull distribution. However, these data does not follow the two-parameters Weibull model since 12.5362>$\chi^2_{2,0.05}$=11.0705.

VI. The acute myeloid leukemia data follows the odd log-logistic inverted Weibull distribution. However, these data does not follow the two-parameters Weibull model since $13.4521 < \chi^2_{2,0.05}$=11.0705.

VII. The leukemia data can be modelled using the odd log-logistic inverted Weibull distribution and the two-parameters Weibull model. However, the odd log-logistic inverted Weibull distribution is better than the two-parameters Weibull model since $7.2365<\chi^2_{1,0.05}= 9.4877$ for the odd log-logistic inverted Weibull distribution and $10.5236<\chi^2_{2,0.05}=11.0705$ for two-parameters Weibull model.

VIII. The strengths data can be modelled using the odd log-logistic inverted Weibull distribution and the two-parameters Weibull model. However, the odd log-logistic inverted Weibull distribution is better than the two-parameters Weibull model since $9.8569<\chi^2_{2,0.05}=11.0705$ for the odd log-logistic inverted Weibull distribution and $9.5236<\chi^2_{2,0.05}=11.0705$ for two-parameters Weibull model.

References

Abouelmagd, T. H. M., Hamed, M. S., Hamedani, G. G., Ali, M. M., Goual, H., Korkmaz, M. C., Yousof, H. M., (2019). The zero truncated Poisson Burr X family of distributions with properties, characterizations, applications, and validation test. *Journal of Nonlinear Sciences and Applications*, Vol. 12, pp. 314–336.

Abouelmagd, T. H. M., Hamed, M. S., Handique, L., Goual, H., Ali, M. M., Yousof, H. M., Korkma, M. C., (2019). *A new class of distributions based on the zero truncated Poisson distribution with properties and applications*, Vol. 12, pp. 152–164.

Aidi, K., Butt, N. S., Ali, M. M., Ibrahim, M., Yousof, H. M., Shehata, W. A. M., (2021). A Modified Chi-square Type Test Statistic for the Double Burr X Model with Applications to Right Censored Medical and Reliability Data. *Pakistan Journal of Statistics and Operation Research*, Vol. 17, pp. 615–623.

Bagdonavičius, V., Nikulin, M., (2011a). Chi-squared Goodness-of-fit test for right censored Data. *Int. J. Appl. Math. Stat.*, Vol. 24, pp. 30–50.

Bagdonavičius, V., Nikulin, M., (2011b). Chi-squared tests for general composite hypotheses from censored samples Comptes Rendus de lácadémie des Sciences de Paris. *Mathématiques*, Vol. 349, pp. 219–223.

Cordeiro, G. M., Alizadeh, M., Ozel, G., Hosseini, B., Ortega, E. M. M., Altun, E., (2016). The generalized odd log-logistic family of distributions: properties, regression models and applications. *Journal of Statistical Computation and Simulation*, Vol. 87, pp. 908–932.

Crowder M. J., Kimber A. C., Smith R. L and Sweeting, T. J., (1991). *Statistical analysis of reliability data*, CHAPMAN &HALL/CRC.

Galanova, N., Lemeshko, B. Y., Chimitova, E. V., (2012). Using Nonparametric Goodness-of-Fit Tests to Validate Accelerated Failure Time Models. *Optoelectron. Instrum. Data Process*, Vol. 48, pp. 580–592.

Goual, H. and Yousof, H. M., (2020). Validation of Burr XII inverse Rayleigh model via a modified chi-squared goodness-of-fit test. *Journal of Applied Statistics*, Vol. 47, pp. 393–423.

Goual, H., Yousof, H. M. and Ali, M. M., (2020). Lomax inverse Weibull model: properties, applications, and a modified Chi-squared goodness-of-fit test for validation. *Journal of Nonlinear Sciences* & *Applications* (JNSA), Vol. 13, pp. 330–353.

Goual, H., Yousof, H. M. and Ali, M. M., (2019). Validation of the odd Lindley exponentiated exponential by a modified goodness of fit test with applications to censored and complete data. *Pakistan Journal of Statistics and Operation Research*, Vol. 15, pp. 745–771.

Habib, M. G., Thomas, D. R., (1986). Chi-squared goodness-of-fit tests for randomly censored Data. *Ann. Stat.*, Vol. 14, pp. 759–765.

Hamedani, G.G., Goual, H., Emam, W., Tashkandy, Y., Ahmad Bhatti, F., Ibrahim, M., Yousof, H.M., (2023). A New Right-Skewed One-Parameter Distribution with Mathematical Characterizations, Distributional Validation, and Actuarial Risk Analysis, with Applications. *Symmetry*, Vol. 15, pp. 7451297.

Hollander, M., Pena, E., (1992). Chi-square goodness-of-fit test for randomly censored data. *JASA*, Vol. 87, pp. 458–463.

Ibrahim, M., Aidi, K., Ali, M. M. and Yousof, H. M., (2022). A Novel Test Statistic for Right Censored Validity under a new Chen extension with Applications in Reliability and Medicine. *Annals of Data Science, forthcoming.* doi.org/10.1007/s40745-022-00416-6

Ibrahim. M., Aidi, K., Ali, M. M. and Yousof, H. M., (2021). The Exponential Generalized Log-Logistic Model: Bagdonavičius-Nikulin test for Validation and Non-Bayesian Estimation Methods. *Communications for Statistical Applications and Methods*, Vol. 29, pp. 681–705.

Ibrahim, M., Altun, E., Goual, H., and Yousof, H. M., (2020). Modified goodness-of-fit type test for censored validation under a new Burr type XII distribution with

different methods of estimation and regression modelling. *Eurasian Bulletin of Mathematics*, Vol. 3, pp. 162–182.

Ibrahim, M., Yadav, A. S., Yousof, H. M., Goual, H., & Hamedani, G. G., (2019). A new extension of Lindley distribution: modified validation test, characterizations and different methods of estimation. *Communications for Statistical Applications and Methods*, Vol. 26, pp. 473–495.

Klein J. P. and Moeschberger M. L., (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Statistics for Biology and Health.

Klein J. P. and Moeschberger M. L., (2003). Survival Analysis: Techniques for Censored and Truncated Data. *Springer*, New York.

Mansour, M. M., Ibrahim, M., Aidi, K., Shafique Butt, N., Ali, M. M., Yousof, H. M., & Hamed, M. S. (2020a). A New Log-Logistic Lifetime Model with Mathematical Properties, Copula, Modified Goodness-of-Fit Test for Validation and Real Data Modelling. *Mathematics,* Vol. 8, pp. 1508.

Mansour, M., Rasekhi, M., Ibrahim, M., Aidi, K., Yousof, H. M., & Elrazik, E. A., (2020b). A New Parametric Life Distribution with Modified Bagdonavičius–Nikulin Goodness-of-Fit Test for Censored Validation, Properties, Applications, and Different Estimation Methods. *Entropy*, Vol. 22, pp. 592.

Nikulin. M. S., (1973a). Chi-squared test for normality. *In: proceedings of the International Vilnius Conference on Probability Theory and Mathematical Statistics*, Vol. 2, pp. 119–122.

Nikulin. M. S., (1973b). Chi-squared test for continuous distributions with shift and scale parameters. *Theory of Probability and its Applications*, Vol. 18, pp. 559–568.

Nikulin. M. S., (1973c). On a Chi-squared test for continuous distributions. *Theory of Probability and its Applications*, Vol. 19, pp. 638–639.

Rao, K. C., Robson, D. S., (1974). A Chi-square statistic for goodness-of-fit tests within the exponential family. *Communication in Statistics*, Vol. 3, pp. 1139–1153.

Ravi, V., and Gilbert, P. D., (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function, *J. Statist. Software*, 32(4).

Emam, W., Tashkandy, Y., Goual, H., Hamida, T., Hiba, A., Ali, M.M., Yousof, H.M., Ibrahim, M., (2023). A New One-Parameter Distribution for Right Censored Bayesian and Non-Bayesian Distributional Validation under Various Estimation Methods. *Mathematics 2023*, Vol. 11, pp. 897.

Yadav, A. S., Goual, H., Alotaibi, R. M., Ali, M. M., Yousof, H. M., (2020). Validation of the Topp-Leone-Lomax model via a modified Nikulin-Rao-Robson goodness-of-fit test with different methods of estimation. *Symmetry*, Vol. 12, pp. 57.

Yadav, A. S., Shukla, S., Goual, H., Saha, M. and Yousof, H. M., (2022). Validation of xgamma exponential model via Nikulin-Rao-Robson goodness-of- fit test under complete and censored sample with different methods of estimation. *Statistics, Optimization & Information Computing*, Vol. 10, pp. 457–483.

Yousof, H. M., Al-nefaie, A. H., Aidi, K., Ali, M. M., Ibrahim, M., (2021a). A Modified Chi-square Type Test for Distributional Validity with Applications to Right Censored Reliability and Medical Data: A Modified Chi-square Type Test. *Pakistan Journal of Statistics and Operation Research*, Vol. 17, pp. 1113–1121.

Yousof, H. M., Aidi, K., Hamedani, G. G and Ibrahim, M., (2021b). A new parametric lifetime distribution with modified Chi-square type test for right censored validation, characterizations and different estimation methods. *Pakistan Journal of Statistics and Operation Research*, Vol. 17, pp. 399–425.

Yousof, H. M., Ali, M. M., Goual, H. and Ibrahim. M., (2021c). A new reciprocal Rayleigh extension: properties, copulas, different methods of estimation and modified right censored test for validation. *Statistics in Transition new series*, Vol. 23, pp. 1–23.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 19–36, https://doi.org/10.59170/stattrans-2023-049
Received – 16.02.2021; accepted – 11.01.2023

Investigation of half-normal model using informative priors under Bayesian structure

Sania Khawar Kiani[1], Muhammad Aslam[2], M. Ishaq Bhatti[3]

Abstract

This paper considers properties of half-normal distribution using informative priors under the Bayesian criterion. It employs the squared root inverted gamma, Chi-square and Rayleigh distributions as the prior distribution to construct the Posterior distributions of the respective distributional parameters. Hyperparameters are elicited via prior predictive distribution. The properties of posterior distribution are studied, and their graphs are presented using a real data set. A comprehensive simulation scheme is conducted using informative priors. Bayes estimates are obtained using the loss functions (squared error loss function, modified loss function, quadratic loss function and Degroot loss function). Statistical inferences interval estimates and Bayesian hypothesis testing are presented to demonstrate the usefulness of the study.

Key words: informative prior, squared root inverted gamma distribution (SRIG), Bayesian hypothesis testing, loss functions.

1. Introduction

Bayesian Inference is an approach to Statistical Inference, which is distinct from frequentist inference. Bayesian statistics, named for Thomas Bayes, is a set of fields of statistics in which the evidence about the true state of the world is expressed in terms of degrees of belief or, more specifically, Bayesian probabilities. Moravveji etal. (2019) presents a Bayesian approach for the estimation of the parameters of two-piece scale mixtures of normal distributions. This is a rich family of light/heavy-tailed symmetric/asymmetric distributions that includes, as a special case, the heavy-tailed scale mixtures of normal distributions, and is flexible in computations for modelling

[1] Riphah College of Rehabilitation Sciences, Riphah International University, Islamabad, Pakistan.
E-mail: sania.khawar@riphah.edu.pk. ORCID: https://orcid.org/0000-0002-7744-3593.
[2] Department of Mathematics and Statistics, Riphah International University, Islamabad, Pakistan.
E-mail: m.aslam@riphah.edu.pk, aslamsdqu@yahoo.com. ORCID: https://orcid.org/0000-0003-3355-2330.
[3] SBE-UBD, Universiti Brunei Darussalam, Brunei & LBS, La Trobe University, Melbourne, Australia.
E-mail: i.Bhatti@latrobe.edu.au & ishaq.bhatti@ubd.edu.bn. ORCID: https://orcid.org/0000-0002-5027-7871.

symmetric and asymmetric data. A Bayesian approach is possible from the specification of hierarchical representations of the proposed family[4].

The half-normal distribution (HND) is linked with skewed positive data in describing lifetime process under fatigue. Various studies are done on the characteristics of HND under Bayesian with the choice of various priors. For example, Bland and Altman (1999) studied the half-normal model for dealing with the relationships between measurement and magnitude error whereas Cohen (1992) studied the problem of inference of truncated distributions, including the truncated normal through a classical approach. Classical inference for half-normal model is examined by Pewsey (2002, 2004). Later, Cooray and Ananda (2008) defined the generalized HND derived from a model for static fatigue, which is then followed by Gauss et al. (2012), who study the Kumaraswamy generalized half-normal distribution for modelling skewed positive data. Gupta (2018) estimates the location parameter of a HND is considered. Some unbiased as well as biased estimators are derived. Admissibility and minimaxity of Pitman estimator are proved. A complete class of estimators among multiples of the maximum likelihood estimator is obtained.

Dobler (2015) developed Stein's method for HND and applied it to derive rates of convergence in distributional limit theorems for three statistics of the simple symmetric random walk: the maximum value, the number of returns to the origin and the number of sign changes up to a given time 'n'. Dobler compares the characterizing operator of the limiting HND with suitable characteristics of the discrete approximating distributions. Jeniffer et al. (2014) study the extended generalized half-normal distribution for modelling skewed fatigue life data. The new model contains as special cases the half-normal and generalized half-normal (Cooray and Ananda, 2008) distributions. Several of its structural properties are derived, including the density function, moments, quantile and generating functions, mean deviations and order statistics. They investigate maximum likelihood estimation of the model parameters. Alzaatreh and Knight (2013) propose the gamma-HND. Various structural properties of the gamma-HND are derived. The shape of the distribution may be unimodel or bimodal. Results for moments, limit behaviour, mean deviations and Shannon entropy are provided. To estimate the model parameters, the method of maximum likelihood estimation is proposed. Three real-life data sets are used to illustrate the applicability of the gamma-HND. For the first time, Cordeiro (2012) study the Kumaraswamy generalized HND for modelling skewed positive data. The half-normal and generalized half-normal (Cooray and Ananda, 2008) distributions are special cases of the new model. Several of its structural properties are derived, including explicit expressions for

[4] For recent applications, one can refer to the recent work by Shrivastava et al. (2019), Montagna et al. (2020), Ariyo et al. (2022) and Sindhu and Hussain 2022, among others.

the density function, moments generating and quantile functions, mean deviations, and moments of the order statistics.

Some recent important works related to simulation, the choice of complex priors related to HND, are done by various authors including Van Erp and Brown (2020) and Al Amer et al. (2021), Sindhu and Hussain (2022), Ariyo et al. (2022), Bruch and Felderer (2022), Martin et al. (2022), among others. Here, we would like to summarize their work for the ready reference of the readers. For example, Ariyo et al. (2022) explored the performance of three Bayesian model-selection criteria when vague priors are used for the covariance parameters of the random effects in a linear mixed-effects model using simulation study. They considered five different specifications of inverse-Wishart (IW) prior, five different separation priors and a joint prior. The results show that marginals perform far better over the conditional and the superiority of joint and separation priors over IW in all settings with selection criteria on a practical data set. Second is the work of Bruch and Felderer (2022), who considered prior choice for the variance parameter in multilevel regression and poststratification selective data and their Monte Carlo simulation study was done on the similar way as that of ours. They observed that prior choices are *challenging* when data results from selective inclusion mechanism which may be subject to bias in the estimation of a proportion based on a sample that is subject to a highly selective inclusion mechanism.

Moreover, similar work is done by Martin et al. (2022) using Python instead of SAS. They explored Bayesian modelling and computation in Python with the aim to help beginner Bayesian practitioners to become intermediate modellers. Beside SAS, they used PyMC3, Tensor-flow Probability and Arvi-Z approaches and other libraries focusing on the practice of applied statistics with a summary of references to the package used in, whereas Sindhu and Hussain (2022) derived and performed predictive inference and parameter estimation from the half-normal distribution for the left censored data. They also derive the posterior and predictive distribution in conjunction with informative vis-à-vis uninformative priors. They used SAS and simulated left censored samples from a half-normal distribution are utilized to interpret the results.

In this paper, the posterior distributions of the parameter using informative priors are derived in Section 2. The prior predictive distributions are derived in Section 3. Section 4 presents the elicitation of the hyperparameters via prior predictive distribution. The graphs of posterior distributions using a real data set are drawn in Section 5. In Section 6, the expressions of Bayes estimates under different loss functions are obtained. Section 7 presents Bayes estimates and Posterior risks using real data set. Section 8 contains credible intervals and hypothesis testing using a real data set. A simulation study is conducted using Mathematica and SAS packages[5] in Section 9. Section 10 contains some concluding remarks.

[5] For the use of other software and computing subroutines one can refer the work of Martin et al. (2022)

2. Posterior Distribution of the Parameter Using Informative Priors

A random variable X is said to be half-normal distribution with location parameter zero and unknown scale parameter θ if its p.d.f is:

$$f(x;\theta) = \sqrt{\frac{2}{\pi}}\frac{1}{\theta} exp\left\{-\left(\frac{x^2}{2\theta^2}\right)\right\}, \quad \theta > 0, 0 < x < \infty \tag{2.1}$$

Let $x_{1,}x_{2,}\ldots,x_n$ be a random sample taken from HND with unknown parameter θ and its likelihood function is:

$$L(\theta\ \mathbf{x}) = \left(\sqrt{\frac{2}{\pi}}\right)^n \frac{1}{\theta^n} exp\left\{-\left(\frac{\sum_{i=1}^n x^2}{2\theta^2}\right)\right\} \tag{2.2}$$

2.1. Posterior Distribution using Informative Priors

The posterior distribution using informative priors, i.e. squared root inverted gamma prior, inverted chi-square prior and inverse Raleigh prior, are presented in the following sections.

2.1.1. Posterior Distribution Using Squared Root Inverted Gamma Prior

The squared root inverted gamma (SRIG) with hyperparameters 'a' and 'b' is defined as:

$$p(\theta) = \frac{2b^a}{\Gamma(a)}\theta^{-(2a+1)} exp\left\{-\left(\frac{b}{2\theta^2}\right)\right\}, \ a, b, \theta > 0 \tag{2.3}$$

Using equations (2.2) and (2.3), the posterior distribution of the parameter θ given data **x** is:

$p(\theta|\mathbf{x}) \propto p(\theta)L(\theta, x)$

$$p(\theta|\mathbf{x}) \propto \frac{2\left(b+\frac{\sum x^2}{2}\right)^{\frac{n}{2}}}{\Gamma\left(a+\frac{n}{2}\right)}\theta^{-\left[2\left(a+\frac{n}{2}\right)+1\right]} exp\left\{-\left[\frac{1}{\theta^2}\left(b+\frac{\sum x^2}{2}\right)\right]\right\}, \quad 0 < \theta < \infty \tag{2.4}$$

which is the density kernel of (SRIG) distribution, so the posterior distribution of $\theta|\mathbf{x}$ is

$$SRIG(\alpha, \beta)\ where\ \alpha = a + \frac{n}{2} \text{ and } \beta = b + \frac{\sum x^2}{2}.$$

2.1.2. Posterior Distribution using Inverted Chi-square Prior

The inverted chi-square (IC) with hyperparameter 'ν' and is defined as:

$$p(\theta) = \frac{\left(\frac{1}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\theta^{-\left(\frac{\nu}{2}+1\right)} exp\left\{-\left(\frac{1}{2\theta^2}\right)\right\}, \quad \nu, \theta > 0 \tag{2.5}$$

Using equations (2.2) and (2.5), the posterior distribution of the parameter $\theta|\mathbf{x}$ is:

$$p(\theta|\mathbf{x}) \propto \frac{2\left(\frac{\sum x^2+1}{2}\right)^{\frac{\nu}{4}+\frac{n}{2}}}{\Gamma\left(\frac{\nu}{4}+\frac{n}{2}\right)}\theta^{-\left[2\left(\frac{\nu}{4}+\frac{n}{2}\right)+1\right]} exp - \left[\frac{1}{\theta^2}\left(\frac{\sum x^2+1}{2}\right)\right], \quad 0 < \theta < \infty \tag{2.6}$$

which is the density kernel of (SRIG) distribution, so the posterior distribution of $\theta|\mathbf{x}$ is

$$SRIG(\alpha, \beta)\ where\ \alpha = \frac{\nu}{4} + \frac{n}{2} \text{ and } \beta = \frac{\sum x^2+1}{2}.$$

2.1.2. Posterior Distribution using Inverse Rayleigh Prior

The inverse Rayleigh (IR) with Hyperparameter 'c' is defined as:

$$p(\theta) = \frac{2c}{\theta^3} exp - \left(\frac{c}{\theta^2}\right), \quad c, \theta > 0 \tag{2.7}$$

Using equations (2.2) and (2.7), the posterior distribution of the parameter $\theta|\mathbf{x}$ is:

$$p(\theta|\mathbf{x}) \propto \frac{2\left(\frac{\sum x^2}{2}+c\right)^{\frac{n}{2}+1}}{\Gamma\left(\frac{n}{2}+1\right)} \theta^{-\left[2\left(\frac{n}{2}+1\right)+1\right]} exp - \left[\frac{1}{\theta^2}\left(\frac{\sum x^2}{2}+c\right)\right], \quad 0 < \theta < \infty \tag{2.8}$$

which is the density kernel of (SRIG) distribution, so the posterior distribution of $\theta|\mathbf{x}$ is

$$SRIG(\alpha, \beta) \; where \; \alpha = \frac{n}{2} + 1 \text{ and } \beta = \frac{\sum x^2}{2} + c.$$

3. Prior Predictive Distribution Using Informative Priors

The prior predictive distribution is the model predicts over the observed variables before any of data are considered. The prior predictive distribution is also known as marginal distribution of an unobserved value which is the prior distribution of θ and single variable p.d.f integrating out this parameter. The derivations of the prior predictive distribution using informative priors are given below. Let Y be the random variable having the HND with unknown parameter θ.

The prior predictive distribution can be obtained by the following equation

$$p(y) = \int_0^\infty p(\theta) f(y, \theta) \, d\theta \tag{3.1}$$

where y represents future random variable.

3.1. Prior Predictive Distribution using Squared root Inverted Gamma Prior

The prior predictive distribution using equation (2.3) and (3.1) is:

$$p(y) = \sqrt{\frac{2}{\pi}} \frac{b^a \Gamma\left(a+\frac{1}{2}\right)}{\Gamma(a)\left(b+\frac{y^2}{2}\right)^{a+\frac{1}{2}}}, y > 0 \tag{3.2}$$

The above equation is used for the elicitation of hyperparameters '*a*' and '*b*'.

3.2. Prior Predictive Distribution using Inverted Chi-Square Prior

The prior predictive distribution using equation (2.5) and (3.1) is:

$$p(y) = \sqrt{\frac{2}{\pi}} \left(\frac{1}{2}\right)^{\frac{\nu}{2}} \frac{b^a \Gamma\left(\frac{\nu}{4}+\frac{1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\left(\frac{y^2+1}{2}\right)^{\frac{\nu}{4}+\frac{1}{2}}}, y > 0 \tag{3.3}$$

The above equation is used for the elicitation of Hyperparameter 'ν'.

3.3. Prior Predictive Distribution using Inverse Rayleigh Prior

The prior predictive distribution using equation (2.7) and (3.1) is:

$$p(y) = \frac{3c}{2\sqrt{2}\left(c^2+\frac{y^2}{2}\right)^{\frac{3}{2}}}, y > 0 \tag{3.4}$$

The above equation is used for the elicitation of Hyperparameter '*c*'.

4. Elicitation of Hyperparameters

The methods of elicitation through prior predictive distribution are defined by Aslam (2003). For the elicitation of the hyperparameters of the informative priors, we use prior predictive distributions given in Ssection 3 and consider the intervals that are used in the elicitation.

4.1. Elicitation of Hyperparameters of Squared root inverted Gamma Prior

Using the prior predictive distribution given in equation (3.2), expert's probabilities are to be 0.15 and 0.10, which are associated with the intervals $0.01 \leq y \leq 0.5$ and $3 \leq y \leq 5$ respectively.

$$\int_{0.01}^{0.05} \sqrt{\frac{2}{\pi}} \frac{b^a \Gamma\left(a+\frac{1}{2}\right)}{\Gamma(a)\left(b+\frac{y^2}{2}\right)^{a+\frac{1}{2}}} dy = 0.15$$

$$\int_{3}^{5} \sqrt{\frac{2}{\pi}} \frac{b^a \Gamma\left(a+\frac{1}{2}\right)}{\Gamma(a)\left(b+\frac{y^2}{2}\right)^{a+\frac{1}{2}}} dy = 0.10$$

To elicit the hyperparameters 'a' and 'b', the above equations are simultaneously solved through the program developed in SAS package using 'PROC SYSNLIN' commands and the values of the hyperparameters 'a' and 'b' are found to be 0.7136 and 0.1330 respectively .

4.2. Elicitation of Hyperparameter of Inverted chi-square prior

Using the prior predictive distribution given in equation (3.3).The expert's probability for the interval (0, 0.5) is to be 0.5.

$$\int_{0}^{0.5} \sqrt{\frac{2}{\pi}} \left(\frac{1}{2}\right)^{\frac{\nu}{2}} \frac{b^a \Gamma\left(\frac{\nu}{4}+\frac{1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\left(\frac{y^2+1}{2}\right)^{\frac{\nu}{4}+\frac{1}{2}}} = 0.18$$

The above equation is used to elicit the hyperparameter of inverted chi-square 'ν' by applying 'PROC SYSNLIN' and the value of the hyperparameter 'ν' is found to be 0.8963.

4.3. Elicitation of Hyperparameter of Inverse Rayleigh Prior

Using the prior predictive distribution given in equation (3.4), expert's probability is to be 0.08, which is associated with the interval $4 \leq y \leq 6$.

$$\int_{4}^{6} \frac{3c}{2\sqrt{2}\left(c^2+\frac{y^2}{2}\right)^{\frac{3}{2}}} = 0.08$$

The above equation is used to elicit the hyperparameter of inverse Rayleigh 'c' by applying 'PROC SYSNLIN' and the value of the hyperparameter 'c' is found to be 0.8531.

5. Graphs of Posterior Distribution Using Real Data Set

This section represents the graphs of the posterior distribution using informative priors. We draw graphs in SAS package.

5.1. Real Data Set

The real data set is used for analysis. From Serge et al. (2010), the data set of maximum flood levels (in millions cubic feet per second) for the Susquehanna River at Harrisburg, Pennsylvania over four-year periods. We have the following 20 observations:

0.654, 0.613, 0.402, 0.379, 0.269, 0.740, 0.416, 0.338, 0.315, 0.449, 0.297, 0.423, 0.379, 0.3235, 0.418, 0.412, 0.494, 0.392, 0.484, 0.268.

The mean, variance and CV of the above data are as follows.

$$\bar{X} = 0.423 \;\; \sigma^2 = 0.016 \text{ CV} = 0.295$$

5.1.1. Graphs of Posterior Distributions

The graphs of posterior distribution using SRIG prior with parameters $\alpha_{SRIG} = 10.7136$, $\beta_{SRIG} = 2.07245$, IC prior with parameters $\alpha_{IC} = 10.224075$, $\beta_{IC} = 2.3945$, and IR prior with parameters $\alpha_{IR} = 11, \beta_{IR} = 2.79255$ are presented below in Figures 5.1, 5.2 and 5.3.

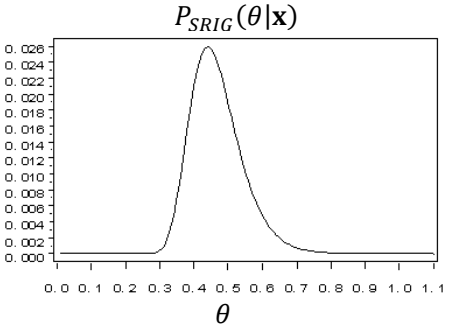


Figure 5.1: Graph using SRIG prior

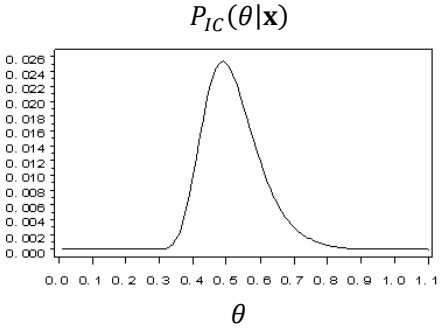


Figure 5.2: Graph using IC prior

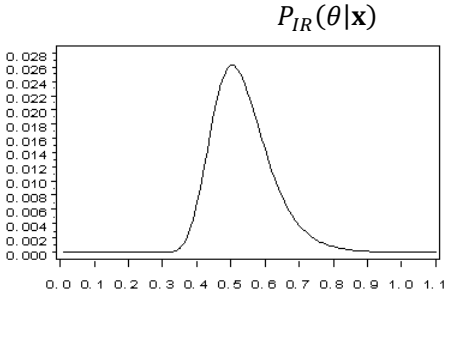


Figure 5.3: Graph using IR prior

The graphs of posterior distributions using informative priors in Figures 5.1, 5.2 and 5.3 are similar and positively skewed.

5.2 Properties of Posterior Distribution Using Real Data Set

The properties of posterior distribution using a real data set mentioned in 5.1 are determined and given below.

Table 5.1: Properties of Posterior Distribution

n=20	Mean	Variance	Mode	C.V
SRIG Prior	1.3724	0.0054	0.4299	5.3638%
IC Prior	1.4854	0.2342	0.4769	32.5835%
IR Prior	1.5950	0.1414	0.4927	23.5793%

From the above Table 5.1, if we compare informative priors, squared root inverted Gamma prior is more efficient than other priors, as variance is minimum using Squared root inverted Gamma prior.

6. Bayes Estimates Under Different Loss Functions

In statistics, typically a loss function is used for parameter estimation, and the event in question is a function of the difference between estimated and true values for an instance of data. In this section, we have used four different loss functions. The details are given below.

6.1. Squared Error Loss Function

The loss function: $L(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$ is called squared error loss function (SELF), where θ is the parameter and $\hat{\theta}$ is an estimator.

By minimizing the risk function $\rho(\hat{\theta}) = EL(\theta,\hat{\theta})$ with respect to θ, we have the Bayes estimator

$$\hat{\theta} = E(\theta) \tag{6.1}$$

which is the posterior mean under SELF.

The Bayes posterior risk is

$$\rho(\hat{\theta}) = E(\theta^2) - \{E(\theta)\}^2 \tag{6.2}$$

which is the posterior variance, and it is the Bayes posterior risk under SELF.

6.2. Quadratic Loss Function

The loss function: $L(\theta,\hat{\theta}) = \left(1 - \frac{\hat{\theta}}{\theta}\right)^2$ is called quadratic loss function (QLF).

By minimizing the risk function, we have $\hat{\theta} = \frac{E(\theta^{-1})}{E(\theta^{-2})}$, which is the Bayes estimator under QLF.

The Bayes posterior risk is

$\rho(\hat{\theta}) = 1 - \frac{\{E(\theta^{-1})\}^2}{E(\theta^{-2})}$. This is the Bayes posterior risk under quadratic loss function.

6.3. Modified Loss Function

The loss function $L(\theta,\hat{\theta}) = \frac{(\theta-\hat{\theta})^2}{\theta}$ is called modified loss function (MLF).

By minimizing the risk function, we have $\hat{\theta} = \frac{1}{E(\theta^{-1})}$, which is the Bayes estimator under MLF.

The Bayes posterior risk is $\rho(\hat{\theta}) = E(\theta) - \frac{1}{E(\theta^{-1})}$.This is the Bayes posterior risk under modified loss function.

6.4. Degroot Loss Function

The loss function $L(\theta,\hat{\theta}) = \left(\frac{\theta-\hat{\theta}}{\hat{\theta}}\right)^2$ is called Degroot loss function (DLF).

By minimizing the risk function, we have $\hat{\theta} = \frac{E(\theta^2)}{E(\theta)}$, which is the Bayes estimator under DLF.

The Bayes posterior risk is

$$\rho(\hat{\theta}) = \frac{Var(\theta)}{E(\theta)} \tag{6.3}$$

This is the Bayes posterior risk under Degroot loss function. The expressions of Bayes estimators and posterior risks using SRIG, IC and IR priors are given in Tables 6.1, 6.2 and 6.3 respectively.

Table 6.1: Bayes Estimators and Posterior Risks Assuming SRIG Prior

Loss Functions	Bayes Estimators	Posterior Risks
SELF	$\hat{\theta} = \sqrt{b+\frac{\sum x^2}{2}}\frac{\Gamma\left(a+\frac{n-1}{2}\right)}{\Gamma\left(a+\frac{n}{2}\right)}$	$\rho(\hat{\theta}) = \frac{2b+\sum x^2}{2a+n-2} - \left(\sqrt{b+\frac{\sum x^2}{2}}\frac{\Gamma\left(a+\frac{n-1}{2}\right)}{\Gamma\left(a+\frac{n}{2}\right)}\right)^2$
QLF	$\hat{\theta} = \sqrt{b+\frac{\sum x^2}{2}}\frac{\Gamma\left(a+\frac{n+1}{2}\right)}{\Gamma\left(a+\frac{n+2}{2}\right)}$	$\rho(\hat{\theta}) = 1 - \left[\left(\frac{1}{a+\frac{n}{2}}\right)\left(\frac{\Gamma\left(a+\frac{n+1}{2}\right)}{\Gamma\left(a+\frac{n+2}{2}\right)}\right)^2\right]$
MLF	$\hat{\theta} = \sqrt{b+\frac{\sum x^2}{2}}\frac{\Gamma\left(a+\frac{n}{2}\right)}{\Gamma\left(a+\frac{n+1}{2}\right)}$	$\rho(\hat{\theta}) = \sqrt{b+\frac{\sum x^2}{2}}\left[\left(\frac{\Gamma\left(a+\frac{n-1}{2}\right)}{\Gamma\left(a+\frac{n}{2}\right)} - \frac{\Gamma\left(a+\frac{n}{2}\right)}{\Gamma\left(a+\frac{n+1}{2}\right)}\right)\right]$
DLF	$\hat{\theta} = \sqrt{b+\frac{\sum x^2}{2}}\frac{\Gamma\left(a+\frac{n-2}{2}\right)}{\Gamma\left(a+\frac{n-1}{2}\right)}$	$\rho(\hat{\theta}) = \frac{\frac{2b+\sum x^2}{2a+n-2} - \left[\sqrt{b+\frac{\sum x^2}{2}}\frac{\Gamma\left(a+\frac{n-1}{2}\right)}{\Gamma\left(a+\frac{n}{2}\right)}\right]^2}{\sqrt{b+\frac{\sum x^2}{2}}\frac{\Gamma\left(a+\frac{n-2}{2}\right)}{\Gamma\left(a+\frac{n}{2}\right)}}$

Table 6.2: Bayes Estimators and Posterior Risks Assuming IC Prior

Loss Functions	Bayes Estimators	Posterior Risks
SELF	$\hat{\theta} = \sqrt{\frac{\sum x^2 + 1}{2}} \frac{\Gamma\left(\frac{v}{4} + \frac{n-1}{2}\right)}{\Gamma\left(\frac{v}{4} + \frac{n}{2}\right)}$	$\rho(\hat{\theta}) = \frac{\sum x^2 + 1}{v + 2n - 2} - \left(\sqrt{\frac{\sum x^2 + 1}{2}} \frac{\Gamma\left(\frac{v}{4} + \frac{n-1}{2}\right)}{\Gamma\left(\frac{v}{4} + \frac{n}{2}\right)}\right)^2$
QLF	$\hat{\theta} = \sqrt{\frac{\sum x^2 + 1}{2}} \frac{\Gamma\left(\frac{v}{4} + \frac{n+1}{2}\right)}{\Gamma\left(\frac{v}{4} + \frac{n+2}{2}\right)}$	$\rho(\hat{\theta}) = 1 - \left[\left(\frac{1}{\frac{v}{4} + \frac{n}{2}}\right)\left(\frac{\Gamma\left(\frac{v}{4} + \frac{n+1}{2}\right)}{\Gamma\left(\frac{v}{4} + \frac{n}{2}\right)}\right)^2\right]$
MLF	$\hat{\theta} = \sqrt{\frac{\sum x^2 + 1}{2}} \frac{\Gamma\left(\frac{v}{4} + \frac{n}{2}\right)}{\Gamma\left(\frac{v}{4} + \frac{n+1}{2}\right)}$	$\rho(\hat{\theta}) = \sqrt{\frac{\sum x^2 + 1}{2}}\left[\left(\frac{\Gamma\left(\frac{v}{4} + \frac{n-1}{2}\right)}{\Gamma\left(\frac{v}{4} + \frac{n}{2}\right)} - \frac{\Gamma\left(\frac{v}{4} + \frac{n}{2}\right)}{\Gamma\left(\frac{v}{4} + \frac{n+1}{2}\right)}\right)\right]$
DLF	$\hat{\theta} = \sqrt{\frac{\sum x^2 + 1}{2}} \frac{\Gamma\left(\frac{v}{4} + \frac{n-2}{2}\right)}{\Gamma\left(\frac{v}{4} + \frac{n-1}{2}\right)}$	$\rho(\hat{\theta}) = \frac{\frac{\sum x^2+1}{v+2n-4} - \left[\sqrt{\frac{\sum x^2+1}{2}} \frac{\Gamma\left(\frac{v}{4}+\frac{n-1}{2}\right)}{\Gamma\left(\frac{v}{4}+\frac{n}{2}\right)}\right]^2}{\sqrt{\frac{\sum x^2+1}{2}} \frac{\Gamma\left(\frac{v}{4}+\frac{n-2}{2}\right)}{\Gamma\left(\frac{v}{4}+\frac{n}{2}\right)}}$

Table 6.3: Bayes Estimators and Posterior Risks Assuming IR Prior

Loss Functions	Bayes Estimators	Posterior Risks
SELF	$\hat{\theta} = \sqrt{\frac{\sum x^2 + c}{2}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2} + 1\right)}$	$\rho(\hat{\theta}) = \frac{\sum x^2 + 2c}{n} - \left(\sqrt{\frac{\sum x^2 + c}{2}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2} + 1\right)}\right)^2$
QLF	$\hat{\theta} = \sqrt{\frac{\sum x^2}{2} + c} \frac{\Gamma\left(\frac{n+3}{2}\right)}{\Gamma\left(\frac{n}{2} + 2\right)}$	$\rho(\hat{\theta}) = 1 - \left[\left(\frac{1}{\frac{n}{2}+1}\right)\left(\frac{\Gamma\left(\frac{n+3}{2}\right)}{\Gamma\left(\frac{n}{2} + 2\right)}\right)^2\right]$
MLF	$\hat{\theta} = \sqrt{\frac{\sum x^2}{2} + c} \frac{\Gamma\left(\frac{n}{2} + 1\right)}{\Gamma\left(\frac{n+3}{2}\right)}$	$\rho(\hat{\theta}) = \sqrt{\frac{\sum x^2}{2} + c}\left[\left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2} + 1\right)} - \frac{\Gamma\left(\frac{n}{2} + 1\right)}{\Gamma\left(\frac{n+3}{2}\right)}\right)\right]$
DLF	$\hat{\theta} = \sqrt{\frac{\sum x^2}{2} + c} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}$	$\rho(\hat{\theta}) = \frac{\frac{\sum x^2+2c}{n} - \left[\left(\sqrt{\frac{\sum x^2}{2} + c} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}+1\right)}\right)^2\right]}{\sqrt{\frac{\sum x^2}{2} + c} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}+1\right)}}$

We have simulated the values of the Bayes estimators and posterior risks given in Appendix under different loss functions. If we compare the results of the Bayes estimates and posterior risks, we can see the simulated values are closer to the true parameter as we increase our sample sizes under different loss functions. Bayes estimates and posterior risks have minimum values for SRIG prior, which shows SRIG prior has more efficient results than other priors. While comparing the loss functions, the SELF has more precise results than all other loss functions. We can conclude that among all prior distributions SRIG has better results.

7. Bayes Estimation and Posterior Risks Using Real Data Set

By using the above loss functions, the Bayes estimates and posterior risks of the parameter through informative priors, i.e. SRIG, IC and IR priors, are as follows, with posterior risks in parentheses.

Table 7.1: Bayes Estimates and Posterior Risk under Different loss Functions

n=20	Prior Distributions		
Loss Functions	SRIG	IC	IR
SELF	1.3724(0.0067)	1.4854(0.0070)	1.5951(0.0068)
QLF	1.3781(0.0231)	1.4922(0.0242)	1.7470(0.0225)
MLF	1.5067(0.0110)	1.6382(0.0128)	1.7470(0.0122)
DLF	1.3691(0.0028)	1.4815(0.0053)	1.5915(0.0041)

If we compare informative priors, we observe that posterior risk using Squared root inverted gamma prior is less than other priors hence SRIG prior gives more efficient results. We observe that MLF performance in terms of posterior risk is better than other loss functions.

8. Bayesian Point and Interval Estimates Using Real Data Sets

In this section, we obtained Bayesian point and interval estimates. The Bayesian analog of a classical confidence is called a credible set. For details about credible sets, see Saleem and Aslam (2009), Lynn et al. (2003) and Saleem and Raza (2011), among others. The Bayesian credible intervals are obtained by using the posterior distribution of the respective parameter of interest.

8.1. Credible Intervals

A credible interval or Bayesian confidence interval is an interval in which domain of a posterior probability distribution is used for interval estimation. Credible intervals are not unique on a posterior distribution.

The credible intervals are constructed as:

$$1-\alpha = p\left\{\chi^2_{\left(1-\frac{\alpha}{2}2p\right)} < \frac{2(\beta)}{A} < \chi^2_{\left(\frac{\alpha}{2},2p\right)}\right\}$$

We have,

$$\left[C_L{}^{(\theta)}, C_U{}^{(\theta)}\right] = \left[\sqrt{\frac{2(\beta)}{\chi^2_{\left(1-\frac{\alpha}{2}2(\alpha)\right)}}}, \sqrt{\frac{2(\beta)}{\chi^2_{\left(\frac{\alpha}{2}2(\alpha)\right)}}}.\right] \quad 8.1$$

Thus $\left(C_L{}^{(\theta)} < \theta < C_U{}^{(\theta)}\right)$ is the $(1-\alpha)$ 100% credible interval where 'α' and 'β' are the respective parameters of posterior distribution.

The Credible intervals for real data set by using equation (8.1) are given in Table 8.1.

Table 8.1: Credible Intervals using Informative Priors

Prior Distributions	90% Credible Interval	95% Credible Interval	99% Credible Interval
SRIG	(0.3452,0.4788)	(0.3358,0.4960)	(0.3185,0.5326)
IC	(0.3349,0.5054)	(0.3532,0.5238)	(0.3349,0.5630)
IR	(0.3724,0.5155)	(0.3622,0.5339)	(0.3437,0.5731)

In comparison, we can observe that 90% credible intervals are narrower than 99% and 95%.When we compare informative priors' credible intervals under squared root inverted gamma prior are shorter than all other priors.

8.2 Bayesian Hypothesis Testing

Hypothesis testing has been subject to polemic since its early formulation by the Neyman and Pearson in the 1930s. It is more difficult to carry out a point null hypothesis test in a Bayesian analysis. Bayesian model comparison is a method of selection based on the Bayes factors. Bayes Factor is ratio of probabilities for null and alternative hypotheses.

Jeffreys (1961) gives the following typology for comparing $H_a\ vs\ H_b$ where H_a is used for null hypothesis and H_b is used for alternative hypothesis. (i) $B > 1$ H_a is supported, (ii) $10^{-\frac{1}{2}} \leq B \leq 1$ Minimal evidence against H_a (iii) $10^{-1} \leq B \leq 10^{-\frac{1}{2}}$ Substantial evidence against H_a.

(iv) $10^{-2} \leq B \leq 10^{-1}$ Strong evidence against H_a (v) $B < 10^{-2}$ Decisive evidence againstH_a.

Table 8.2: Hypothesis testing using Real Data Set

$H_a\ vs\ H_b$	Using SRIG Prior	Using IC Prior	Using IR Prior
	B.F	B.F	B.F
$H_a: \theta \leq 0.34$ $H_b: \theta > 0.34$	0.0268	0.0031	0.0009
$H_a: \theta \leq 0.43$ $H_b: \theta > 0.43$	0.6696	0.2049	0.1280
$H_a: \theta \leq 0.55$ $H_b: \theta > 0.55$	8.5737	2.7522	2.1075
$H_a: \theta \leq 0.68$ $H_b: \theta > 0.68$	115.713	26.6151	21.592

The above Table 8.2 shows:

- While considering the hypothesis

$$H_a: \theta \leq 0.34 \text{ Versus } H_b: \theta > 0.34$$

Bayes factor using squared root inverted gamma priors lies between $10^{-2} \leq B \leq 10^{-\frac{1}{2}}$. So we conclude that there is substantial evidence against the posterior distribution under H_a, and $B \leq 10^{-2}$so we conclude decisive evidence against the posterior distribution under H_a.

- While considering the hypothesis

$$H_a: \theta \leq 0.43 \text{ Versus } H_b: \theta > 0.43$$

As $10^{-\frac{1}{2}} \leq B \leq 1$ we have minimal evidence against $H_a$for all priors.

- While considering the hypothesis

$$H_a: \theta \leq 0.55 \text{ Versus } H_b: \theta > 0.55$$

As $B > 1$, so we strongly supported H_a using all informative priors.

- While considering the hypothesis

$$H_a: \theta \leq 0.68 \text{ Versus } H_b: \theta > 0.68$$

As $B > 1$, so we strongly supported $H_a$using all informative priors.

9. Properties of Posterior Distribution using Simulation Study

Simulation is the process of imitating a real phenomenon with a set of mathematical formulas. Here, we discuss some properties of posterior distribution through a simulation study of parameter θ. We have done all simulations in Mathematica package.

Table 9.1: Properties of Posterior Distribution under SRIG Prior

n	$\theta = 2$			$\theta = 4$		
	Mean	Variance	Mode	Mean	Variance	Mode
50	1.9907	1.2314	1.9980	3.9425	1.2268	3.9792
100	1.9964	1.2088	1.9996	3.9945	1.2096	3.9987
500	2.0050	1.1984	2.0048	4.0510	1.1914	4.0070
1000	2.0007	1.1871	2.0006	4.0014	1.1804	4.0003

Table 9.2: Properties of Posterior Distribution under IC Prior

n	$\theta = 2$			$\theta = 4$		
	Mean	Variance	Mode	Mean	Variance	Mode
50	1.9867	1.2173	1.9825	3.9839	1.2040	3.9971
100	1.9969	1.2054	1.9963	3.9914	1.2039	3.9997
500	2.0968	1.1978	2.0963	4.0775	1.1907	4.0553
1000	2.0012	1.1882	2.0010	4.0003	1.1847	4.0072

Table 9.3: Properties of Posterior Distribution under IR Prior

n	$\theta = 2$			$\theta = 4$		
	Mean	Variance	Mode	Mean	Variance	Mode
50	1.9867	1.2173	1.9825	3.9839	1.2040	3.9971
100	1.9969	1.2054	1.9963	3.9914	1.2039	3.9997
500	2.0968	1.1978	2.0963	4.0775	1.1907	4.0553
1000	2.0012	1.1882	2.0010	4.0003	1.1847	4.0072

From the Tables 9.1, 9.2 and 9.3, it is observed that as we increase our sample sizes, our simulated values through mean tend to true values of parameter. Similarly, mode is closely to the true parameter as we increase sample sizes. Squared root inverted gamma prior is more precise than all other priors in the case of comparing informative priors. We have also simulated values of variances, which can show as we increase the sample sizes it becomes less.

10. Concluding Remarks

We have presented the Bayesian analysis of half-normal model using informative (squared root inverted gamma, inverted chi-square and inverse Rayleigh) priors. Initially, we derive posterior distributions using informative priors. The SAS package is used to draw graphs of posterior distributions. The properties of posterior distribution (mean, median, mode, variance and coefficient of variation) are discussed through simulation as well as real data set. The credible intervals for 90%, 95%, and 99% using informative priors are constructed and the Bayes factors of different hypothesis are computed. By the comparison of results, with increasing the sample size the Bayes estimates converge to the parametric values and their risks tend to be smaller. As under informative priors the Bayes risks for the estimates under SRIG are smaller than the Bayes risks assuming IC and IR priors, thus SRIG is more suitable prior. If we compare the Bayes risk under different loss functions, namely SELF, QLF, MLF and DLF, then the MLF is a better loss function for estimating the parameter θ.

References

Al Amer, F. M., Thompson, C. G. and Lin, L., (2021). Bayesian methods for meta-analyses of binary outcomes: implementations, examples, and impact of priors. *International journal of environmental research and public health*, 18(7), p. 3492.

Ayman and Kristen, (2013). On the gamma-half normal distribution and its applications. *Journal of Modern Applied Statistical Methods*, 12(1), p.15.

Allan, A. T., Hill, R. A., (2021). Definition and interpretation effects: how different vigilance definitions can produce varied results. *Animal Behaviour*, 180, pp.197–208.

Ariyo, O., Lesaffre, E., Verbeke, G. and Quintero, A., (2022). Model selection for Bayesian linear mixed models with longitudinal data: sensitivity to the choice of priors. *Communications in statistics-simulation and computation*, 51(4), pp. 1591–1615.

Aslam, M., (2003). An Application of the Prior Predictive Distribution to Elicit the Prior Density. *Journal of Statistical Theory and Applications*, 2(1), pp. 70-83.

Aslam M., Saleem, M., (2009). On Bayesian Analysis of the Rayleigh Survival Time Assuming the Random Censor Time. *Pakistan Journal of Statistics*, 25(2), pp. 71–82.

Berger, O. J., (1985). Statistical Decision Theory and Bayesian Analysis. 2nd edition, Springer Series in Statistics, ISBN–10: 0-387-96098-8 and –13: 978-0387-96098-2.

Bland, J. M., Altman, D., G., (1999). Measuring agreement in method comparison studies. *Stat Methods Med* Res 8, pp. 135–160.

Bruch, C., Felderer, B., (2022). Prior Choice for the Variance Parameter in the Multilevel Regression and Post stratification Approach for Highly Selective Data. A Monte Carlo Simulation Study. *Austrian Journal of Statistics*, 51(4), pp. 76–95.

Casella, L., Elberly, G., (2003). Estimating Bayesian Credible Intervals. *Journal of the Statistical Planning and Inference,* 112, pp. 115–32.

Cohen, A. C., (1991). *Truncated and censored samples: theory and applications*. CRC press.

Cordeiro, G. M., Pescim, R. R. and Ortega, E. M., (2012). The Kumaraswamy generalized half-normal distribution for skewed positive data. *Journal of Data Science*, 10(2), pp. 195–224.

Cooray, K., Ananda, M. M. A., (2008). A generalization of the half-normal distribution with applications to lifetime data. *Communication in Statistics − Theory and Methods,* 37, pp. 1323–1337.

Dobler, C., (2015). Stein's method for the half-normal distribution with applications to limit theorems related to the simple symmetric random walk. *ALEA: Latin American Journal of Probability and Mathematical Statistics*, 20(109), p. 34.

Jeffreys, H., (1998). *The theory of probability*. OUP Oxford.

Martin, O. A., Kumar, R. and Lao, J., (2022). *Bayesian Modeling and Computation in Python*. Chapman and Hall/CRC.

Moravveji, B., Khodadadi, Z. and Maleki, M. A., (2019). Bayesian Analysis of Two-Piece Distributions Based on the Scale Mixtures of Normal Family. *Iran J Sci Technol Trans Sci.,* 43**,** pp. 991–1001.

Patra, L. K., Kumar, S. and Gupta, N., (2018). Estimation of the Location Parameter of a General Half-Normal Distribution. *International Conference on Mathematics and Computing Springer*, Singapore, pp. 281–293.

Pewsey, A., (2002). Large-sample inference for the general half-normal distribution. *Communications in Statistics-Theory and Methods*, 31(7), pp. 1045–1054.

Pewsey, A., (2004). Improved likelihood based inference for the general half-normal distribution. *Communications in Statistics-Theory and Methods*, 33(2), pp.197–204.

Robert, C. P., Casella, G., (1994). Distance weighted losses for testing and confidence set evaluation. *Test*, 3(1), pp. 163–182.

Saleem, M., Raza, A., (2011). On Bayesian Analysis of the Exponential Survival Time Assuming the Exponential Censor time. *Pakistan Journal of Science*, 63(1).

Provost, S. B., Mabrouk, I., (2010). A generalized exponential-type distribution. *Pak. J. Statist*, 26(1), pp. 97–110.

Shrivastava, A., Chaturvedi, A. and Bhatti, M. I., (2019). Robust Bayesian analysis of a multivariate dynamic model. *Physica A: Statistical Mechanics and its Applications*, 528, p. 121451.

Sindhu, T. N., Hussain, Z., (2022). Predictive Inference and Parameter Estimation from the Half-Normal Distribution for the Left Censored Data. *Annals of Data Science*, 9(2), pp. 285–299.

Silvia, M., Vanessa, O. and Argiento, R., (2020). *Bayesian isotonic logistic regression via constrained splines: an application to estimating the serve advantage in professional tennis*.

Sanchez, J. J. D., da Luz Freitas, W. W. and Cordeiro, G. M., (2016). The extended generalized half-normal distribution. *Brazilian Journal of Probability and Statistics*, pp. 366–384.

Van Erp, S., Browne, W. J., (2021). Bayesian Multilevel Structural Equation Modeling: An Investigation into Robust Prior Distributions for the Doubly Latent Categorical Model. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(6), pp. 875–893.

Appendix

Bayes Estimates and posterior risks under Different Loss Functions

Table 1: Bayes Estimates of Informative Priors using SELF

Priors	SRIG Prior		IC Prior		IR Prior	
N	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$
50	2.0083	3.9706	2.0506	3.9758	2.0167	3.9662
100	2.0044	3.9901	2.0224	3.9964	2.0426	3.9912
500	1.9923	4.0637	1.9849	4.0396	1.9429	4.0826

Table 2: Bayes Estimates of Informative Priors using QLF

Priors	SRIG Prior		IC Prior		IR Prior	
n	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$
50	1.9698	3.9856	2.0720	3.9937	1.9897	3.9927
100	1.9904	3.9942	2.0892	4.0463	1.9983	3.9987
500	2.0097	4.0468	2.0048	4.0930	2.0562	4.0963

Table 3: Bayes Estimates of Informative Priors using MLF

Priors	SRIG Prior		IC Prior		IR Prior	
n	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$
50	2.0387	4.0204	1.9835	3.9913	2.0214	4.0443
100	2.0932	4.0439	1.9991	3.9978	2.0610	4.0728
500	2.0083	4.0992	2.0437	4.0933	1.9886	4.0027
1000	2.0037	4.0070	2.0679	4.0013	1.9989	4.0002

Table 4: Bayes Estimates of Informative Priors using DLF

Priors	SRIG Prior		IC Prior		IR Prior	
n	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$
50	1.9862	3.9999	2.0160	3.9915	1.9925	4.0202
100	1.9999	4.0537	2.0028	3.9996	1.9957	4.0210
500	2.0193	4.0860	2.0921	4.0928	2.0868	4.0114
1000	2.0641	4.0025	2.0944	4.0047	2.0029	3.9993

Table 5: Posterior Risks of Informative Priors using SELF

Priors	SRIG Prior		IC Prior		IR Prior	
n	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$
50	1.2116	1.2376	1.2070	1.2112	1.2296	1.2314
100	1.2048	1.2092	1.1927	1.2042	1.2059	1.2094
500	1.1881	1.1880	1.1885	1.1905	1.1914	1.1902
1000	1.1809	1.1847	1.1864	1.1886	1.1880	1.1888

Table 6: Posterior Risks of Informative Priors using QLF

Priors	SRIG Prior		IC Prior		IR Prior	
n	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$
50	1.2225	1.2427	1.2143	1.2189	1.2281	1.2412
100	1.2153	1.2338	1.2095	1.2161	1.2109	1.2218
500	1.1967	1.2113	1.1918	1.1905	1.1912	1.1912
1000	1.1871	1.1883	1.1882	1.1879	1.1871	1.1863

Table 7: Posterior Risks of Informative Priors using MLF

Priors	SRIG Prior		IC Prior		IR Prior	
n	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$
50	1.2151	1.2363	1.2556	1.2690	1.2352	1.2252
100	1.2058	1.2287	1.2340	1.2566	1.2253	1.2007
500	1.1979	1.2052	1.2075	1.2116	1.1943	1.1946
1000	1.1873	1.1964	1.1991	1.1923	1.1838	1.1877

Table 8: Posterior Risks of Informative Priors using DLF

Priors	SRIG Prior		IC Prior		IR Prior	
n	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$	$\theta = 2$	$\theta = 4$
50	1.2431	1.2492	1.2151	1.2158	1.2226	1.2456
100	1.2364	1.2146	1.2007	1.2097	1.2193	1.2134
500	1.2079	1.1983	1.1878	1.1912	1.1992	1.1886
1000	1.1875	1.1876	1.1822	1.1884	1.1954	1.1805

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 37–52, https://doi.org/10.59170/stattrans-2023-050
Received – 28.08.2022; accepted – 05.05.2023

What explains the differences in material deprivation between rural and urban areas in Poland before and during the COVID-19 pandemic?

Hanna Dudek[1], **Joanna Landmesser-Rusek**[2]

Abstract

We examined the relationships between the compositional changes in demographic and socioeconomic factors and the changes in the prevalence of material deprivation in rural and urban areas in Poland. Using the European Union Statistics on Income and Living Conditions (EU-SILC) data for 2019–2020, we applied the Fairlie decomposition approach for a logit model. We found that the important characteristics affecting a gap in material deprivation between rural and urban areas are: equivalized income, the level of education, the type of household, and the presence of disabled or unemployed people in the household. A non-significant effect of the pandemic on the material deprivation gap between rural and urban areas was observed.

Key words: rural-urban differences, COVID-19, logit model, Fairlie decomposition, EU-SILC.

1. Introduction

In recent years, there has been growing interest in material deprivation. This issue refers to the inability to satisfy needs considered basic in European conditions. By focusing on the financial inability to satisfy these needs, the analysis of material deprivation enables a more direct measurement of the population's standard of living than income indicators.

The relevancy of material deprivation research in the European Union (EU) has grown significantly since 2010 as a result of the adoption of the Europe 2020 Strategy (Guio et al., 2016). As of this time, material deprivation indicators have been used by all EU Member States and the European Commission to monitor national and EU

[1] Warsaw University of Life Sciences, Poland. E-mail: hanna_dudek@sggw.edu.pl. ORCID: https://orcid.org/0000-0001-8261-2745.

[2] Warsaw University of Life Sciences, Poland. E-mail: joanna_landmesser@sggw.edu.pl. ORCID: https://orcid.org/0000-0001-7286-8536.

progress towards the EU's social protection and social inclusion objectives (Fusco et al., 2013; Guio, 2018). Apart from reports of statistical offices, there are currently many scientific papers devoted to material deprivation. Research literature includes studies relating to material deprivation in individual countries (Šoltés and Ulman, 2015; Dudek and Szczesny, 2021a) as well as in the entire EU (Bárcena-Martín et al., 2014; Bedük, 2018; Dudek, 2019; Łuczak and Kalinowski, 2020).

The literature indicates an existing poverty gap between rural and urban areas in Central and Eastern European countries (Bernard, 2019; Swain, 2016). When it comes to Poland, Dudek and Szczesny (2021a) found a higher level of material deprivation among rural than urban households. However, they found that in regression models including typical socioeconomic variables and degree of urbanisation of the place of residence, rural-urban differences were statistically insignificant. Thus, it is essential to pinpoint the causes of the rural-urban gap in material deprivation. This problem was taken up in the present study.

The primary source of information on material deprivation in the EU is the EU Statistics on Income and Living Conditions (EU-SILC) survey. As the EU-SILC survey in 2020 was conducted in Poland in the fall, it is possible to investigate the effect of the COVID-19 pandemic on material deprivation. Thus, we use pre-COVID data (2019) and 2020 data to examine the impact of the COVID-19 pandemic on rural-urban differences in material deprivation.

In this study, we pose two research questions:

(i) How do rural-urban differences in material deprivation vary according to socioeconomic factors?

(ii) Has the COVID-19 pandemic affected rural-urban differences?

In other words, the main aim of this paper is to identify factors influencing rural-urban differences in material deprivation of Polish households. In addition, the study aims to investigate the extent to which these differences are explained by given socioeconomic features. For these purposes, it proposes using the Fairlie decomposition approach. This approach works by decomposing the difference in proportions based on a probit or logit binary model.

Our paper contributes to the literature by exploiting new information collected in 2019–2020 through the EU-SILC survey to provide a snapshot of material deprivation among Polish households when the country continues struggling with the COVID-19 pandemic. The paper also contributes to the literature by providing the first econometric evidence for factors affecting the rural-urban gap in material deprivation in Poland using the Fairlie decomposition approach.

2. Methodology

2.1. Fairlie decomposition method

Decomposition techniques are most commonly used in studying gender pay gaps using linear regression models (see Słoczyński, 2012; Zajkowska, 2013; Śliwicki and Ryczkowski, 2014; Landmesser et al., 2015, Landmesser, 2017). Such studies mainly use the Blinder-Oaxaca decomposition technique, dividing the group differences into two parts: a part explained by compositional differences and a part that is 'attributable to the coefficients' (Blinder 1973; Oaxaca 1973). This technique can be extended to non-linear models, including models with binary dependent variables.

Following the Blinder-Oaxaca concept, Fairlie (2005) proposed the idea of decomposition for binary probit and logit models. Fairlie initially used this method to analyse racial differences in the digital divide (Fairlie, 2005) and race differences among business owners (Fairlie and Robb, 2007). Over the past five years, the Fairlie technique has been widely used in various fields of science, e.g. in analyses of the gender gap in food insecurity (Broussard, 2019), rural-urban inequalities in health (Rahimi et al., 2021), gender differences in saving behaviour (Boto-García et al., 2022). However, it has not been used in material deprivation analysis before.

Thus, this article provides the first results on rural-urban differences in material deprivation using the Fairlie approach. Below we present the concept of this approach.

The standard Blinder-Oaxaca decomposition for a linear regression can be expressed as:

$$\bar{Y}^A - \bar{Y}^B = (\bar{X}^A - \bar{X}^B)\,\hat{\beta}^A + \bar{X}^B\left(\hat{\beta}^A - \hat{\beta}^B\right) \tag{1}$$

where

A – group A,
B – group B,
$\bar{Y}^A, \bar{Y}^B$ – the average values of the dependent variable for group A and B, respectively,
$\bar{X}^A, \bar{X}^B$ – row vectors of average values of independents variables for group A and B, respectively,
$\hat{\beta}^A, \hat{\beta}^B$ – vectors of parameter estimates for group A and B, respectively.

Unlike in linear models, where $\bar{Y}^A = \bar{X}^A\,\hat{\beta}^A$ and $\bar{Y}^B = \bar{X}^B\,\hat{\beta}^B$, which formula (1) implies, in models with a nonlinear function F, $\bar{Y}^A$ does not necessarily equal $F(\bar{X}^A)$ and $\bar{Y}^B$ does not necessarily equal $F(\bar{X}^B)$. However, the following dependencies occur in the logit model:

$$\bar{Y}^A = \sum_{i=1}^{N^A} \frac{F(X_i^A \hat{\beta}^A)}{N^A} \text{ and } \bar{Y}^B = \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^B)}{N^B} \tag{2}$$

where N_A and N_B are the sample size for group A and B, respectively.

Thus, following Fairlie (2005), the decomposition can be written as:

$$\bar{Y}^A - \bar{Y}^B = \left[\sum_{i=1}^{N^A} \frac{F(X_i^A \hat{\beta}^A)}{N^A} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^A)}{N^B}\right] + \left[\sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^A)}{N^B} - \sum_{i=1}^{N^B} \frac{F(X_i^B \hat{\beta}^B)}{N^B}\right] \quad (3)$$

The first term in brackets measures the disparity due to the differences in characteristics (the 'characteristic effect'), and the second term in brackets measures the disparity due to the different effects of the observed characteristics (the 'coefficient effect'). The second term also captures the portion of the binary outcome variable gap due to group differences in unmeasurable or unobserved endowments.

The estimation of the total contribution is the difference between the average values of the predicted probabilities. Assuming that N_A=N_B, using parameter estimates from the logit model for the pooled sample, $\hat{\beta}^*$, the contribution of X_1 to the rural-urban gap can be written as:

$$\frac{1}{N^B}\sum_{i=1}^{N^B} F(\hat{\alpha}^* + X_{1i}^A \hat{\beta}_1^* + X_{2i}^A \hat{\beta}_2^*) - F(\hat{\alpha}^* + X_{1i}^B \hat{\beta}_1^* + X_{2i}^A \hat{\beta}_2^*) \quad (4)$$

Similarly, the independent contribution of X_2 can be expressed as:

$$\frac{1}{N^B}\sum_{i=1}^{N^B} F(\hat{\alpha}^* + X_{1i}^B \hat{\beta}_1^* + X_{2i}^A \hat{\beta}_2^*) - F\left(\hat{\alpha}^* + X_{1i}^B \hat{\beta}_1^* + X_{2i}^B \hat{\beta}_2^*\right) \quad (5)$$

Standard errors for (4) and (5) can be calculated by the delta method (Fairlie, 2005).

In practice, the sample sizes of the two groups (N_A and N_B) may differ. In such a case, a one-to-one matching of observations from the two samples is needed to be calculated. To address this problem, random subsamples of equal sizes are drawn.

Fairlie (2017) recommended the replication of the decomposition from a minimum of 1000 subsamples and finding the mean values of estimates from each decomposition to obtain an accurate decomposition estimate. More detailed information in this regard was provided by Fairlie (2017).

In our study, we used the STATA program (module) written by Jann (2006) to carry out the analysis. This program enables:

- to draw a hypothetical sample with replacement from both groups, whereby the probability of being selected from the sample is proportionate to the sampling weight;
- to solve the path dependence problem in the detailed decomposition, with multiple estimations of material deprivation with randomised order of the independent variables being performed, and the obtained effects being averaged over all possible orderings.

Thus, using the 'Fairlie' module in STATA (Stata-Corp, College Station, Texas, United States of America), we carried out the decomposition analysis to enable the

quantification of how much of the gap between the rural and urban groups is attributable to differences in specific measurable characteristics.

In the Fairlie decomposition technique, a positive coefficient would result in a positive contribution to the rural-urban gap in material deprivation, and it is interpreted as supporting (increasing) the rural-urban material deprivation inequality (if the disparity is positive). A negative coefficient would similarly yield a negative contribution to the material deprivation inequality and consequently works to decrease the inequality if the inequality is positive.

2.2. Data

To study the ways in which differences between rural areas (thinly populated areas) and urban areas (densely populated or intermediate populated areas) in material deprivation in Poland were affected via various mechanisms, we use 2019–2020 data from the EU-SILC cross-sectional files. EU-SILC provides annual population representative information on material deprivation and several demographic and socioeconomic variables regarding EU countries.

Usually, the EU-SILC survey in Poland is carried out in April – June. However, in 2020 it was conducted by the Central Statistical Office and 16 statistical offices from September 28 to December 4, 2020. The change of the survey date was dictated by the appearance of a pandemic threat during the survey conducted so far (Statistics Poland, 2022). Thus, the surveyed Poles in the fall of 2020 had already experienced some effects of the pandemic caused by the lockdown.

The analysed sample includes 19,874 Polish households from the 2019 wave and 15,281 Polish households from the 2020 wave.

The study considers those indicators that were considered both under the Europe 2020 and Europe 2030 strategies (see: Poverty in Poland…, 2021). All indicators analysed are binary indicators corresponding to given material deprivation items.

These items relate to the inability of a household to:

1) avoid arrears on mortgage or rent, utility bills, hire purchase instalments or other loan payments (the feature short name in our analysis: ‘arrears’);
2) afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day (the short name: ‘food’);
3) face unexpected financial expenses (the short name: ‘unexpected’);
4) afford to keep the home adequately warm (the short name: ‘warm’);
5) afford a one-week annual holiday away from home (the short name: ‘holiday’); and
6) afford a car/van for private use (the short name: ‘car’).

We consider the following set of socioeconomic factors for the decomposition:

- the natural logarithm of annual equivalized household disposable income (using the modified OECD equivalence scale) from the previous year, with 2020 prices (the continuous variable 'income');
- household type (single-person, 2 young adults (age<65), 2 older adults, 2 adults with 1 dependent child, 2 adults with 2 dependent children, 2 adults with at least 3 dependent children, single parent with children, other with dependent children, other without dependent children; the 'household type' variable in the models);
- region of Poland (7 regions; the 'region' variable);
- presence of persons in households whose activities were limited due to health reasons (the 'disability' variable);
- presence of unemployed in the household (the 'unemployed' variable);
- presence of retirees in the household (the 'retired' variable);
- the average age of household's members (the 'mean age' variable in the models);
- highest education level of household members (the 'max education' variable with three response categories: tertiary, upper secondary, lower than upper secondary).

To assess the impacts of the COVID-19 pandemic in 2020 we use the dummy variable 'year 2020' defined as 1 for 2020 sample observations, and 0 for 2019.

3. Results

The results of a preliminary analysis of income changes in Poland and the EU in 2019–2020 showed an increase in median annual equivalized income (measured in purchasing power standards – PPS). Poland saw a rise in income (12,366 in PPS in 2019, 13,381 in 2020), while the EU (27-country area) saw a decrease (17,478 in PPS in 2019, 17,337 in 2020) (Eurostat, 2022). Also, in Poland, the percentage of households showing various symptoms of material deprivation fell in 2020 compared to 2019 (in contrast to the EU as a whole) (Eurostat, 2022).

The main aim of this paper is to identify factors influencing rural-urban differences in material deprivation among Polish households. A preliminary comparison between rural and urban households in terms of given material deprivation items is presented in Figure 1.

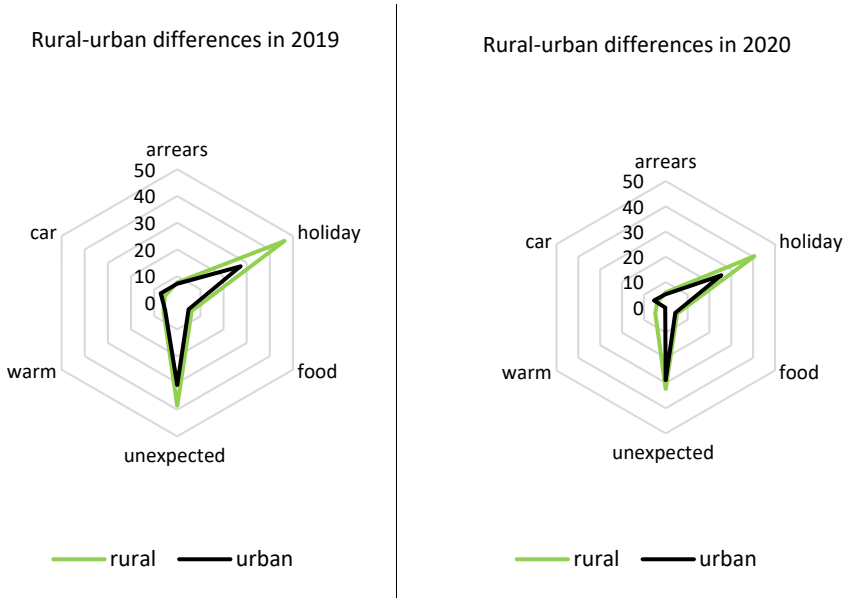


Figure 1: Prevalence of material deprivation among rural and urban households.

Source: Own work.

Based on Figure 1, it can be noticed that material deprivation did not increase in the first pandemic year. Probably, it happened due to the improved income situation of Poles. However, in both analysed years, Polish households were mostly vulnerable regarding the items 'holiday' and 'unexpected'. There were also visible rural-urban differences in these two items. The detailed results of the prevalence of deprivation presented in the Appendix revealed that statistically significant differences between rural and urban households occurred with respect to 'holiday', 'unexpected', 'warm', 'food' and 'car' (the corresponding 95% intervals do not overlap). For the first four items, deprivation in the countryside is greater than in the city. Only deprivation regarding the ability to afford a car was significantly higher among urban households. For the remaining item 'arrears', this difference is not significant at the 0.05 level.

The Fairlie decomposition for binary logit models was conducted to examine the factors affecting the rural-urban gap for each of the six deprivation symptoms. In our study, the process of randomly sampling households and estimating each variable's contribution to the gap was repeated 1,000 times. The order of the variables was randomized on each run to address the issue of path dependence.

The estimated parameters of the logit models enabled the identification of factors influencing the experience of material deprivation. Applying the decomposition technique to the logit model allowed us to extract the factors explaining the observed

differences in material deprivation. Table 1 shows detailed results for the decomposition regarding the two items for which the highest deprivation occurred – i.e. inability to pay for holidays and unexpected financial expenses by place of residence (rural vs urban areas).

Table 1: Results for Fairlie decomposition of differences between rural and urban households concerning 'holiday' and 'unexpected' items

Non-linear decomposition by place of residence						
Specification	holiday			unexpected		
Sample size			34 767			34 687
The sample size for rural areas			11 773			11 745
The sample size for urban areas			22 994			22 942
Deprivation rate in rural areas			0.3879			0.3377
Deprivation rate in urban areas			0.2296			0.2869
Deprivation rate difference	0.1583		100%	0.0508		100%
Explained difference	0.1022		65%	0.0623		122%
Explained part						
Explanatory variable	Coef.			Coef.		
Income	0.0440	**	**28%**	0.0403	**	**79%**
Household type	0.0101	**	**6%**	-0.0229	**	**-45%**
Region	0.0030	**	2%	-0.0014		-3%
Disability	0.0083	**	5%	0.0058	**	11%
Unemployed	0.0049	**	3%	0.0062	**	12%
Retired	-0.0004		0%	-0.0007		-1%
Mean age	-0.0004		0%	0.0000		0%
Max education	0.0328	**	**21%**	0.0351	**	**69%**
Year 2020	0.0000		0%	0.0000		0%
Total explained	0.1022	**	65%	0.0623	**	122%

** – significant at 0.05 level; higher impacts in bold.

Source: own work.

There is a positive difference in deprivation rates between rural and urban households both in their inability to pay for holidays and coping with unexpected financial expenses. The explained effect is high (65% and 122%, respectively). The inequalities examined should be assigned in the majority to the differentiation of individual household characteristics (rather than to parameters in the estimated models).

Regarding differences in deprivation concerning the inability to pay for holidays, the variables that significantly affect the magnitude of deprivation are equivalized income ('income'), household type ('household type'), region of Poland, the presence

of disabled or unemployed in the household, and the highest education level ('max education'). Differences in deprivation due to unexpected expenses are affected by similar variables, except the 'region' variable. The variable denoting the year 2020 was non-significant at the 0.05 level, suggesting an insignificant impact of the pandemic in explaining the rural-urban gap.

The estimated positive coefficients indicate a positive contribution to the rural-urban gap in material deprivation. The values of the variables standing by their support (increase) the observed rural-urban inequality in material deprivation. For example, for the 'holiday' item, the different educational levels of rural and urban residents account for 21% of the observed inequality.

Negative coefficients yield a negative contribution to material deprivation inequality. Thus, the observed differences in deprivation regarding unexpected financial expenses are reduced by the dissimilarity of rural and urban household types (by 45%).

The results of Fairlie decomposition for all analysed symptoms of material deprivation in an aggregate manner are presented in Table 2 and Table 3.

Table 2: Results of Fairlie decomposition for all material deprivation symptoms by place of residence (rural vs urban areas) – influence directions of variables

Specification	arrears	holiday	food	unexpected	warm	car
Pr(Y=1/G=rural)	0.063	0.388	0.053	0.338	0.052	0.048
Pr(Y=1/G=urban)	0.065	0.230	0.045	0.287	0.042	0.068
Difference	-0.002	0.158	0.008	0.051	0.010	-0.020
Explained part	0.011	0.102	0.012	0.062	0.010	0.009
Influence directions of variables						
Income	+	+	+	+	+	+
Household type	–	+	–	–	–	–
Region	ns	+	ns	ns	ns	–
Disability	+	+	+	+	+	+
Unemployed	+	+	+	+	+	+
Retired	ns	ns	ns	ns	ns	ns
Mean age	ns	ns	ns	ns	ns	ns
Max education	+	+	+	+	+	+
Year 2020	ns	ns	ns	ns	ns	ns

+/– means positive/negative contribution of the variable in the explained part at the significance level of 0.05, ns – nonsignificant

Source: own work.

A positive difference in deprivation rates between rural and urban households is found in their inability to pay for holidays, afford food, face unexpected financial expenses, and keep the home warm. A negative difference occurs in the case of the

inability to avoid mortgage or rent arrears and the inability to afford a car. The explained effect is always positive, meaning that the household characteristics included in the analysis magnify the observed rural-urban differences in deprivation.

Equivalized income, disability, unemployment, education level, and household type were the significant variables affecting differences in all symptoms of material deprivation by place of residence. The impact of the region of the household's residence was relevant only for some deprivation items. It was also observed that the presence of retirees and the mean age of household members have no effect. Moreover, 2020 shows non-significance (at the significance level of 0.05), which allows us to conclude that the pandemic has no impact on the material deprivation gap in rural and urban areas.

Table 3: Results of Fairlie decomposition for all material deprivation symptoms by place of residence – percentages of the explained part

Specification	arrears	holiday	food	unexpected	warm	car
Difference	-0.002	0.158	0.008	0.051	0.01	-0.02
Unexplained part	-0.013	0.056	-0.004	-0.011	0	-0.029
Explained part	0.011	0.102	0.012	0.062	0.01	0.009
%Unexplained	650%	35%	-50%	-22%	0%	145%
%Explained	-550%	65%	150%	122%	100%	-45%
Components of the explained part						
Income	-309%	28%	85%	79%	62%	-31%
Household type	197%	6%	-52%	-45%	-41%	26%
Region	34%	2%	13%	-3%	6%	5%
Disability	-145%	5%	19%	11%	14%	-6%
Unemployed	-116%	3%	21%	12%	18%	-10%
Retired	37%	0%	-4%	-1%	-3%	2%
Mean age	73%	0%	0%	0%	-1%	5%
Max education	-308%	21%	74%	69%	49%	-35%
Year 2020	8%	0%	0%	0%	-1%	0%

Source: own work.

The inequalities examined should be assigned in the majority to the differentiation of individual households' characteristics in the inability to pay for holidays, afford food, face unexpected financial expenses, and keep the home warm. Most of the gap is attributed to the parameters in estimated models in the case of the inability to avoid mortgage arrears and the inability to afford a car.

It can be noted that the following variables contribute the most to explaining the observed differences in material deprivation by place of residence: income, education level, household type, and presence of disabled or unemployed people. Variables that provide a negligible explanation for the observed differences are 'region', 'mean age', 'retired', and 'year 2020'.

4. Discussion

Several authors undertook the problem of rural-urban differences in Poland. For example, Landmesser (2009) compared the economic activity of people concerning their place of residence, Sompolska-Rzechula and Kurdys-Kujawska (2020) analysed subjective assessment of the quality of life of rural and urban residents, Kalinowski (2022) investigated poverty in the countryside, Głowicka-Wołoszyn et al. (2019) compared housing conditions and Wołoszyn and Wysocki (2020) focused on income inequalities among rural and urban households. Generally, the mentioned authors found a worse situation in rural areas compared to urban areas.

When it comes to material deprivation, there needs to be more literature on comparisons between rural and urban households. Moreover, most of the research concern a comprehensive analysis of material deprivation (Dudek and Szczesny, 2021a and 2022b; Šoltés and Ulman, 2015). In our study, however, we analyse each symptom separately. This turned out to be important, as urban households were not better off in all items than in the countryside. We found that there was statistically greater deprivation in the urban areas due to the 'car' item.

Not surprisingly, the most important factor influencing the rural-urban gap is income. However, as indicated in the paper (Dudek and Szczesny, 2021b), material deprivation does not coincide with income poverty. Therefore, it is worth considering demographic and socioeconomic factors that may be important in explaining this phenomenon.

This study has some key strengths. Firstly, it uses the newest EU-SILC of nationally representative data. Secondly, it investigates the factors influencing differences in material deprivation between rural and urban areas. For this purpose, it proposes the Fairlie decomposition approach, which has not been used in material deprivation analysis before. Thus, this paper contributes to the literature by providing the first evidence for factors affecting the households' material deprivation in Poland using the Fairlie approach. The main concern with the non-linear model is sensitivity to the order of independent variables included in the decomposition process (path dependency). The Fairlie method solves this problem by randomly ordering the variables across replications of the decomposition.

The limitations of this study relate to the fact that the dependent variables are self-reported and are likely to have reporting bias. Moreover, the data used are cross-sectional and, therefore, we cannot establish any causality between material deprivation and different socioeconomic variables. Despite these limitations, this study gives an understanding and quantification of the drivers and magnitude of rural-urban inequalities in material deprivation.

5. Conclusions

The analysis revealed that both before and during the first year of the COVID-19 pandemic, a significant proportion of Polish households exhibited symptoms of material deprivation. However, material deprivation in Poland decreased during the period studied. Probably, it was the increase in average income that mainly contributed to the decrease in deprivation.

The study focuses on rural-urban differences in material deprivation. Decomposition analysis provided in-depth information about the phenomenon under study. We considered six items of material deprivation analysing each symptom (item) as a binary variable. Separate models were evaluated for each symptom. We used the Fairlie method as it was developed for non-linear regression models, including the logit and probit models. This method basically tests how much of the difference in material deprivation between rural and urban areas is due to differences in the variables included in the analysis. It also goes further to estimate the contribution of each variable to the explained material deprivation difference between rural and urban areas.

It was found that for items 'holiday', 'unexpected', 'food' and 'warm' rural households were significantly more vulnerable than urban, however, a greater prevalence of 'car deprived' was in urban areas. Moreover, for 'arrears' there was no statistical difference in this regard. This means that it is worth analysing each material deprivation item separately.

The detailed decomposition carried out revealed that the important characteristics affecting the occurrence of the rural-urban material deprivation gap are equivalized income, level of education, type of household, and the presence of disabled or unemployed people in the household. However, the results obtained allow us to conclude that there is a statistically insignificant effect of the pandemic in explaining rural-urban differences. It is important to monitor the pandemic effect in the coming years. This would allow the most vulnerable groups of households to be recognized and specific implications for social policy analysis and evaluation to be identified. This issue is crucial as reducing any form of poverty and social exclusion is one of the most important goals of the EU social policy.

Acknowledgement

We thank Eurostat for accessing EU-SILC microdata (research proposal 38/2017-EU-SILC). The results and their interpretation are the authors' responsibility.

References

Bárcena-Martín, E., Lacomba, B., Moro-Egido, A.I., Perez-Moreno, S., (2014). Country Differences in Material Deprivation in Europe. *Review of Income and Wealth*, Vol. 60(4), pp. 802–820.

Bedük, S., (2018). Understanding Material Deprivation for 25 EU Countries: Risk and Level Perspectives, and Distinctiveness of Zeros. *European Sociological Review*, Vol. 34(2), pp. 121–137.

Bernard, J., (2019). Where Have All the Rural Poor Gone? Explaining the Rural–Urban Poverty Gap in European Countries. *Sociologia Ruralis*, Vol. 59 (3), pp. 369-392.

Blinder, A. S., (1973). Wage Discrimination: Reduced Form and Structural Variables. *Journal of Human Resources*, Vol. 8, pp. 436-455.

Boto-García, D., Bucciol, A., Manfrè, M., (2022). The role of financial socialization and self-control on saving habits. *Journal of Behavioral and Experimental Economics*, Vol. 100, 101903.

Broussard, N. H., (2019). What Explains Gender Differences in Food Insecurity. *Food Policy*, Vol. 83, pp. 180–194.

Dudek, H., (2019). Country-level Drivers of Severe Material Deprivation Rates in the EU. *Ekonomický časopis,* Vol. 67(1), pp. 33–51.

Dudek, H., Szczesny, W., (2021a). Multidimensional material deprivation in Poland: a focus on changes in 2015–2017. *Quality* & *Quantity*, Vol. 55, pp. 741–763.

Dudek, H., Szczesny, W., (2021b). Dudek, H., Szczesny, W., (2021). Multi-Dimensional Material Deprivation in the Visegrád Group: Zero-Inflated Beta Regression Modelling. In: Betti, G., Lemmi, A. (eds.), *Analysis of Socio-Economic Conditions: Insights from a Fuzzy Multidimensional Approach*. London and New York: Routledge, pp. 151–165.

Eurostat, (2022). https://ec.europa.eu/eurostat/data/database. Accessed March 17, 2022.

Fairlie, R. W., (2005). An Extension of the Blinder-Oaxaca Decomposition Technique to Logit and Probit Models. *Journal of Economic and Social Measurement*, Vol. 30(4), pp. 305–316.

Fairlie, R. W., Robb, A. M., (2007). Why are Black-Owned Businesses Less Successful than White-Owned Businesses: The Role of Families, Inheritances, and Business Human Capital. *Journal of Labor Economics*, Vol. 25(2), pp. 289–323.

Fairlie, R. W., (2017). Addressing Path Dependence and Incorporating Sample Weights in the Nonlinear Blinder-Oaxaca Decomposition Technique for Logit, Probit and Other Nonlinear Models. *SIEPR Discussion Paper*, No. 17–013. https://siepr.stanford.edu/sites/default/files/publications/17-013.pdf. Accessed February 7, 2022.

Fusco, A., Guio, A.-C., Marlier, E., (2013). Building a Material Deprivation Index in a Multinational Context: Lessons from the EU Experience. In: Berenger, V., Bresson, F. (eds.) *Poverty and Social Exclusion around the Mediterranean Sea*, pp. 43–71. Springer, New York.

Głowicka-Wołoszyn, R., (2019). Multi-Dimensional Assessment of Housing Conditions of the Population in Rural and Urban Areas of the Wielkopolskie Voivodeship. *Annals of the Polish Association of Agricultural and Agribusiness Economists*, T. 21(2), pp. 79–87,

Guio, A.-C., Marlier, E., Gordon, D., Fahmy, E, Nandy, S., Pomati, M., (2016). Improving the Measurement of Material Deprivation at the European Union Level. *Journal of European Social Policy*, Vol. 26(3), pp. 219–333.

Guio, A.-C., (2018). Multidimensional Poverty and Material Deprivation: Empirical Findings. In: D'Ambrosio, C. (eds.) *Handbook of Research on Economic and Social Well-Being*, pp. 171–192. Edward Elgar Publishing, Cheltenham.

Jann, B., (2006). Fairlie: Stata Module to Generate Nonlinear Decomposition of Binary Outcome Differentials. Available from http://ideas.repec.org/c/boc/bocode/s456727.html. Accessed March 21, 2022.

Kalinowski, S., (2022). Ubóstwo i wykluczenie na wsi. In: J. Wilkin, A. Hałasiewicz (eds.), *Polska Wieś 2022. Raport o stanie wsi*, Wyd. FDPA, WN Scholar: Warszawa, pp. 153–169.

Landmesser, J., (2009). The Survey of Economic Activity of People in Rural Areas – the Analysis Using the Econometric Hazard Models. *Acta Universitatis Lodziensis, Folia Oeconomica*, No. 228, pp. 385–392.

Landmesser, J. M., Karpio, K., Łukasiewicz, P., (2015). Decomposition of Differences Between Personal Incomes Distributions in Poland. *Quantitative Methods in Economics*, Vol. XVI(2), pp. 43–52.

Landmesser, J. M., (2017). Differences in Income Distributions for Men and Women in Poland – an Analysis Using Decomposition Techniques. *Acta Scientiarum Polonorum. Oeconomia*, Vol. 16(4), pp. 103–112.

Łuczak, A., Kalinowski, S., (2020). Assessing the Level of the Material Deprivation of European Union Countries. *PLoSONE,* Vol. 15(9), e0238376.

Oaxaca, R., (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, Vol. 14, pp. 693–709.

Poverty in Poland in 2019 and 2020, (2021). Statistics Poland, Warsaw.

Rahimi, E., Hashemi Nazari, S. S., (2021). A Detailed Explanation and Graphical Representation of the Blinder-Oaxaca Decomposition Method with its Application in Health Inequalities. *Emerging Themes* in *Epidemiology,* Vol. 18**,** 12.

Słoczyński, T., (2012). Wokół międzynarodowego zróżnicowania międzypłciowej luki płacowej. *International Journal of Management and Economics*, Vol. 34, pp. 169–185.

Sompolska-Rzechula, A., Kurdys-Kujawska, A., (2020). Quality of Life of Rural and Urban Population in Poland: Evaluation and Comparison. *European Research Studies Journal*, Vol. 23(3), pp. 645–656.

Statistics Poland, (2022). Dochody i warunki życia ludności Polski – raport z badania EU-SILC 2020. Główny Urząd Statystyczny, Warszawa. Available from https://stat.gov.pl/obszary-tematyczne/warunki-zycia/dochody-wydatki-i-warunki-zycia-ludnosci/dochody-i-warunki-zycia-ludnosci-polski-raport-z-badania-eu-silc-2020,6,14.html. Accessed April 2, 2022.

Swain, N., (2016). Eastern European Rurality in a Neo-Liberal. European Union World. *Sociologia Ruralis*, Vol. 56 (4) pp. 574–596.

Śliwicki, D., Ryczkowski, M., (2014). Gender Pay Gap in the Micro Level – Case of Poland. *Quantitative Methods in Economics*, Vol. XV(1), pp. 159–173.

Šoltés, E., Ulman, P., (2015). Material deprivation in Poland and Slovakia – a comparative analysis. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, Vol. 947(11), pp. 19–36.

Wołoszyn, A., Wysocki, F., (2020). Income Inequality of Polish Rural and Urban Households in 2010–2017. *Annals of Polish Association of Agricultural Economists and Agribusiness*, Vol. 22(1), pp. 360–368.

Zajkowska, O., (2013). Gender Pay Gap in Poland – Blinder-Oaxaca Decomposition. *Quantitative Methods in Economics*, Vol. XIV(2), pp. 272–278.

APPENDIX

Table A1: 95% confidence intervals for the proportions of households experiencing given material deprivation item in rural and urban households

Material deprivation item	Rural		Urban	
	LCI	UCI	LCI	UCI
arrears	6.29	7.33	5.77	6.61
holiday	42.41	44.45	25.65	27.14
food	5.24	6.12	4.30	4.98
unexpected	34.44	36.37	29.03	30.58
warm	4.94	5.81	4.22	4.89
car	4.22	5.05	5.84	6.66

Source: own work.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 53–70, https://doi.org/10.59170/stattrans-2023-051
Received – 16.11.2021; accepted – 19.10.2022

Respondent-specific randomized response technique to estimate sensitive proportion

Dipika Patra[1], **Sanghamitra Pal**[2], **Arijit Chaudhuri**[3]

Abstract

In estimating the proportion of people bearing a stigmatizing characteristic in a community of people, randomized response techniques are plentifully available in the literature. They are implemented essentially using boxes of similar cards of two distinguishable types. In this paper, we propose a more general procedure using five different types of cards. A respondent-specific randomized response technique is also proposed, in which respondents are allowed to build up the boxes according to their own choices. An immediate objective for this change is to enhance, sense of protection of privacy of the respondents. But as by-products, higher efficiency in terms of actual coverage percentages of confidence intervals and related features are demonstrated by a simulation study, and superior jeopardy levels against divulgence of personal secrecy are also reported to be achievable.

AMS subject classification: 62D05

Key words: protection of privacy, randomized response, sensitive issues, varying probability sampling.

1. Introduction

Paying heed to Chaudhuri's (2011) text, we consider varying probability sampling designs in surveying finite survey population with the purpose of estimating the proportion of people bearing a sensitive feature like tax evasion, bearing criminal antecedents, etc. in a community of persons. Randomized Response (RR) Techniques (RRTs) with standard procedures given by Warner (1965), Simmons (his URL model), Boruch (his Forced Response model) (vide Chaudhuri (2011) for each) and others are well-known and documented. They essentially employ boxes filled with several

[1] Department of Statistics, Seth Anandram Jaipuria College, Kolkata, India.
E-mail: dipika.patra1988@gmail.com. ORCID: https://orcid.org/0000-0003-4318-1123.
[2] Corresponding author. Department of Statistics, West Bengal State University, India.
E-mail: mitrapal2013@gmail.com. ORCID: https://orcid.org/0000-0002-5752-8282.
[3] Applied Statistics Unit, Indian Statistical Institute, Kolkata, India. E-mail: arijitchaudhuri1@rediffmail.com.
ORCID: https://orcid.org/0000-0002-4305-7686.

identically designed cards with two distinct types of visible marks. Standard procedures of unbiased estimation of the proportion of people bearing the sensitive characteristic, say, A, along with variance formulae and unbiased estimators thereof are available in the above location cited. In recent surveys also, RR technique is quite popular. Treating illegal waste disposal as a sensitive attribute, Chong et al. (2019) analyzed the social problem of waste disposal with RR technique. Arnab and Mothupi (2015) assessed the sexual habits of the University Students, using Warner's (1965) and Greenberg et al.'s (1969) RR techniques. Barabesi et al. (2013) employed RR setups for the estimation of the size of hidden gang and the distribution function of a sensitive variable for the members of the group. Van der Heijden et al. (2000) applied the Forced Response technique and Kuk's (1990) RR technique to obtain reliable data on welfare and unemployment benefits fraud which is highly relevant to policy decisions.Together with various applications of RRT, statistical tools were also developed to analyze RR data. For instance, Hout et al. (2007) discussed the univariate and multivariate logistic regression models to measure sensitive feature. They presented univariate model as a generalized linear model and introduced multivariate model to deal with several RR response variables. Also, Fox et al. (2018) considered a generalized linear model and generalized linear mixed model for RR design. The literature to be cited below is rich giving procedures to provide methods and measures of levels of protection verifiable for the respondents' disclosures of privacy.

In the existing literature of RRT, the interviewer constructs the RR device(s) and the respondents are requested to participate in the RR survey. In practice, respondents hesitate to participate in RR survey. Anticipating more participation in such survey, a new RR survey theory has been proposed, in which respondents are allowed to construct the RR devices. This proposed RR technique is termed as *respondent-specific randomized response technique*.

The paper is organized as follows. Section 2 provides certain basics for RRT in the context of qualitative sensitive features. A brief description of the protection of privacy measures is included there. Section 3 is constructed to propose two general RR techniques covering varying probability sampling design. Section 3.1 describes Model 1 in which five distinct types of cards are used in RR device. Section 3.2 proposes a novel RR device in which respondents are asked to build up the RR boxes according to their own choices. Section 4 is devoted to the measure the respondents' privacy protection. Privacy is protected only for a RR-specific parametric combination and such a feature will be seen in this section. The effectiveness and competitiveness of the proposed RRTs are narrated through numerical findings, in Section 5. This article is ended with some concluding remarks in Section 6.

2. The Early Works

Taking a cue from the pioneering work of Warner (1965), Greenberg et al. (1969) recommended unrelated question model with two questions of which one is about the sensitive characteristic A and the other question is unrelated to the sensitive characteristic. The idea of this RR device is originated by Walt R. Simmons. The reason behind this extension is that like A, it is a complement, i.e. A^c may be a sensitive characteristic. In that case, the respondents may hesitate to give out their true nature. Chaudhuri (2011) developed the RR devices and the estimation procedures for general sampling design. The extensions of the work are narrated in Chaudhuri (2011) (chapter 3), Chaudhuri et al. (2016). Boruch's (1972) Forced RRT considers the RR device with three distinct types of cards. Instead of the unrelated question, he suggested to include the cards marked as "Yes" and "No". Taking a cue from them, a new RR technique has been suggested in this paper with five different options in the RR device. In another proposed RR technique, respondents are allowed to build their RR devices choosing different cards according to their own choices. Then, the respondents will be comfortable to participate in RR survey.

Several authors including Lanke (1975,1976), Leysieffer and Warner (1976), Anderson (1975 a,b,c), Diana and Perii (2013) have drawn the attention of many survey practitioners to measure the degree of protection of the responses. However, their measures are confined to Simple Random Sampling (SRS) with replacement. Chaudhuri, Christofides and Saha (2009) covered the protection of privacy measure for RRTs using general sampling design. With the approach of Chaudhuri et al. (2009) and Pal et al. (2020) the protection of privacy measure has been derived here for the proposed generalized RR techniques.

3. Proposed RR Techniques Using Five-types of Cards

A potentially useful generalized RRT is proposed in sub-section 3.1 as Model 1. Additionally, a respondent-specific randomized response technique is also introduced in the sub-section 3.2 as Model 2.

3.1. Model 1: Generalized RR technique

An ameliorated RR technique is proposed here employing two boxes filled with several identically designed cards with 5 distinct types of visible marks as, " I possess A", "I possess A^c", "I possess innocuous character B", "Yes" and "No" having proportions $p_k, (1-p_k)w_2, (1-p_k)w_3, (1-p_k)w_4$ and $(1-p_k)(1-w_2-w_3-w_4)$ respectively in the $k^{th}(k=1,2)$ box $(p_1 \neq p_2, w_2+w_3+w_4 < 1,\ 0 < p_1,\ p_2,\ w_2,\ w_3,\ w_4 < 1)$.

Let $U = (1,2 \ldots N)$ be a finite population on which the variables y and x are defined. The variables y and x are introduced relating to the sensitive attribute A and the innocuous characteristic B respectively.

Thus, for i^{th} $(i \in U)$ person,

$$y_i = \begin{cases} 1 & ,if\ i^{th}\ person\ bears\ A \\ 0 & ,if\ i^{th} person\ bears\ A^c \end{cases}$$

and

$$x_i = \begin{cases} 1 & ,if\ i^{th}\ person\ bears\ B \\ 0 & ,if\ i^{th} person\ bears\ B^c. \end{cases}$$

The aim is to estimate the population proportion $\theta = \frac{1}{N}\sum_{i=1}^{N} y_i\ ;\ \theta \in [0,1]$.

A sample s of size n is drawn from the population U by any sampling design $P(s)$. A sampled person i $(i \in s, i = 1,2, \ldots, n)$ is requested to draw a card randomly from the 1st box without divulging the card-type. The respondent must give out the truthful response in terms of yes or no according to the card type marked as "I possess A", "I possess A^c" or "I possess innocuous character B". The person is also instructed to report yes or no if the card is marked as "Yes" or "No". Figure 3.1 successfully explains the proposed strategy.

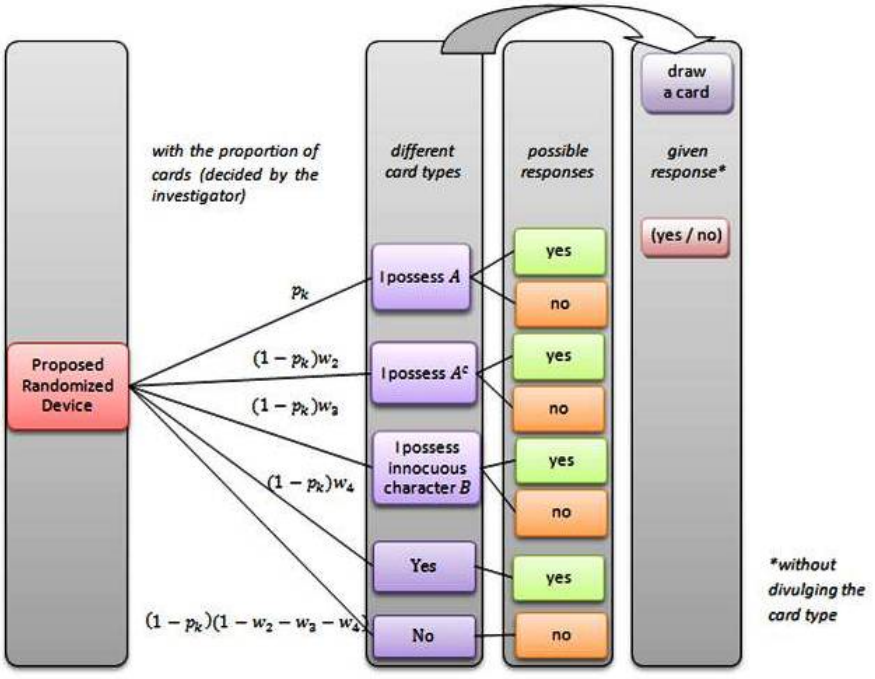


Figure 3.1: Model 1: Generalized RR Technique

Thus, the randomized response from i^{th} $(i \in s)$ person is

$$I_i = \begin{cases} 1 & if\ the\ response\ is\ \ yes \\ 0 & if\ the\ response\ is\ \ no. \end{cases}$$

Therefore,

$$P(I_i = 1) = p_1 y_i + (1 - p_1)\{w_2(1 - y_i) + w_3 x_i + w_4\}$$

and $P(I_i = 0) = p_1(1 - y_i) + (1 - p_1)\{w_2 y_i + w_3(1 - x_i) + (1 - w_2 - w_3 - w_4)\}$.

The person is also requested to report another response described as earlier after drawing a card from the 2nd box, independently.

Therefore, we may denote the 2nd response as,

$$J_i = \begin{cases} 1 & if\ response\ is\ \ yes \\ 0 & if\ response\ is\ \ no \end{cases},\ \ i \in s.$$

Then,

$$P(J_i = 1) = p_2 y_i + (1 - p_2)\{w_2(1 - y_i) + w_3 x_i + w_4\}$$

and $P(J_i = 0) = p_2(1 - y_i) + (1 - p_2)\{w_2 y_i + w_3(1 - x_i) + (1 - w_2 - w_3 - w_4)\}$.

Denoting RR based expectations and variances as E_R and V_R throughout the study, we may write,

$$E_R(I_i) = p_1 y_i + (1 - p_1)\{w_2(1 - y_i) + w_3 x_i + w_4\}$$

and

$$E_R(J_i) = p_2 y_i + (1 - p_2)\{w_2(1 - y_i) + w_3 x_i + w_4\}.$$

Therefore,

$$E_R\big((1 - p_2)I_i - (1 - p_1)J_i\big) = \{p_1(1 - p_2) - p_2(1 - p_1)\}y_i = (p_1 - p_2)y_i$$

$$E_R\left(\frac{(1 - p_2)I_i - (1 - p_1)J_i}{p_1 - p_2}\right) = y_i\,;\quad p_1 \neq p_2$$

leading to

$$r_i = \frac{(1-p_2)I_i - (1-p_1)J_i}{p_1 - p_2},\ p_1 \neq p_2 \tag{1}$$

which is the unbiased estimator for y_i .

An unbiased estimator of the variance $V_R(r_i)$ is given by

$$v_R(r_i) = r_i(r_i - 1) = \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(I_i - J_i)^2, \tag{2}$$

since $y_i^2 = y_i, x_i^2 = x_i, I_i^2 = I_i$ and $J_i^2 = J_i$. The details of the proof are given below.

Considering $v_R^*(r_i) = (1 - p_1)(1 - p_2)(I_i - J_i)^2$ we get,

$$\begin{aligned}
E_R\big(v_R^*(r_i)\big) &= (1 - p_1)(1 - p_2)E_R(I_i - J_i)^2 \\
&= (1 - p_1)(1 - p_2)\big\{E_R\big(I_i^2\big) + E_R\big(J_i^2\big) - 2E_R(I_i)E_R(J_i)\big\} \\
&= (1 - p_2)\{(1 - p_1) - (1 - p_2)\}E_R\big(I_i^2\big) + (1 - p_1)\{(1 - p_2) - (1 - p_1)\}E_R\big(J_i^2\big) \\
&\qquad + E_R\big((1 - p_2)I_i - (1 - p_1)J_i\big)^2 \\
&= (1 - p_2)(p_2 - p_1)E_R\big(I_i^2\big) + (1 - p_1)(p_1 - p_2)E_R\big(J_i^2\big) \\
&\qquad + E_R\big((p_1 - p_2)r_i\big)^2;\ using\ (eq.1) \\
&= -(p_1 - p_2)E_R\big((1 - p_2)I_i - (1 - p_1)J_i\big) + (p_1 - p_2)^2 E_R\big(r_i^2\big);\ since\ I_i^2 = I_i, J_i^2 = J_i \\
&= (p_1 - p_2)^2 E_R\big(r_i^2\big) - (p_1 - p_2)^2 E_R(r_i);\ using\ (eq.\ 1)
\end{aligned}$$

$$= (p_1 - p_2)^2\left(E_R(r_i^2) - y_i\right)$$
$$= \left(p_1 - p_2\right)^2(E_R(r_i^2) - y_i^2);\ since\ y_i = y_i^2$$
$$= (p_1 - p_2)^2(E_R(r_i^2) - (E_R(r_i))^2 = (p_1 - p_2)^2 V_R(r_i).$$

Therefore, $\frac{1}{(p_1-p_2)^2} v_R^*(r_i) = v_R(r_i)$ is the unbiased estimator for $V_R(r_i)$. Employing Horvitz-Thompson (1952) estimator in estimating the population proportion $\theta = \frac{1}{N}\sum_{i\in U} y_i$, the final unbiased estimator can be written as

$$e_{HT} = \frac{1}{N}\sum_{i\in s}\frac{r_i}{\pi_i}. \tag{3}$$

Hence, an unbiased variance estimator is

$$v(e_{HT}) = \frac{1}{N^2}\left[\sum_{i<j\in s}\sum\left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{r_i}{\pi_i}-\frac{r_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_R(r_i)}{\pi_i}\right] \tag{4}$$

where $v_R(r_i)$ is derived in (eq.2).

Composition of Such Randomized Device:

Let the sampled persons be approached with two boxes. In the 1st box, let there be m identically designed cards of which m_1, m_2, m_3 and m_4 cards have visible marks as "I possess A", "I possess A^c", "I possess innocuous character B" and "Yes" respectively. The remaining $(m - m_1 - m_2 - m_3 - m_4)$ cards have a mark "No".

Then, the proportion of cards is

"I possess A": "I possess A^c": "I possess B": "Yes": "No"

$$= m_1 : m_2 : m_3 : m_4 : (m - m_1 - m_2 - m_3 - m_4)$$
$$= \frac{m_1}{m} : \frac{m_2}{m} : \frac{m_3}{m} : \frac{m_4}{m} : \frac{(m-m_1-m_2-m_3-m_4)}{m}$$
$$= \frac{m_1}{m} : \left(\frac{m-m_1}{m}\right)\frac{m_2}{m-m_1} : \left(\frac{m-m_1}{m}\right)\frac{m_3}{m-m_1} : \left(\frac{m-m_1}{m}\right)\frac{m_4}{m-m_1} : \left(\frac{m-m_1}{m}\right)\frac{m-m_1-m_2-m_3-m_4}{m-m_1}.$$

Now, taking $\frac{m_1}{m} = p_1$ and $w_j = \frac{m_j}{m-m_1}; j = 2,3,4,$ the above proportion becomes

$p_1 : (1-p_1)w_2 : (1-p_1)w_3 : (1-p_1)w_4 : (1-p_1)(1-w_2-w_3-w_4)$ where $0 < w_j < 1, j = 2,3,4$ and $w_2 + w_3 + w_4 < 1$ are obvious conditions.

The proportion of the above cards in the 2nd box may be done by changing only (adding or removing) a fixed number of "I possess A" marked cards used in the 1st box. Thus, the number of other-types of cards will remain unchanged as in the 1st box.

Then, one can easily see that the proportion of cards in the 2nd box is now changed to

$$p_2 : (1-p_2)\, w_2 : (1-p_2)\, w_3 : (1-p_2)\, w_4 : (1-p_2)(1-w_2-w_3-w_4)$$

Remark:

1. *The proposed generalized RR technique (Model 1) reduces to Warner's (1965) RRT if* $w_2 = 1,\ w_3 = 0,\ w_4 = 0$.
2. *The proposed Model 1 reduces to Greenberg et al.'s (1969) RRT if* $w_2 = 0, w_3 = 1, m_4 = 0.$
3. *The proposed Model 1 reduces to the Forced RRT if* $w_2 = 0,\ w_3 = 0, w_4 < 1.$

3.2. Model 2: Respondent-specific RR Technique

Intending to enhance the sense of responses' privacy, we modify the RR technique recounted in the previous section (Section 3.1) giving freedom to the respondents to construct their RR devices with the same five distinct types of cards as mentioned earlier. In such a situation, this generalized RR technique, termed as ***Respondent-specific RR technique*** is relevant. This is quite possible that respondents possessing the sensitive characteristic A may prefer any other types of cards except "Yes" marked cards.

With the following illustration, the implementation of this procedure can be easily understandable. Also, Figure 3.2 sheds light on the specifications of the RR devices.

Let a sampled person be approached with two empty boxes and a sufficient number of cards marked as earlier. On request, the person has to build up 1st box (Box 1) by inserting total $m\ (fixed)$ number of cards. In the building process of Box 1, the person will put $m_1\ (fixed\ and > 0)$ number of "I possess A" cards. The rest of the $(m - m_1)$ cards are marked other than "I possess A" marked cards. In other words, there is no restriction on the number of "I possess A^c", "I possess innocuous character B", "Yes" and "No" marked cards. The 2nd box (Box 2) should be built up with $(m + a)$ number of cards in total where the number of "I possess A" cards is $(m_1 + a)$ and the remaining cards are present here in the same number as given in the Box 1. The value of "a" should be decided by the interviewer. Then, the respondents are requested to draw a card randomly from each of the boxes and respond accordingly. The reason to fix up $m\ (> 0), m_1 (< m)$ and $a\ (> 0)$ for all respondents by the interviewer is discussed later.

Therefore, the proportions of "I possess A", "I possess A^c", "I possess innocuous character B", "Yes" and "No" marked cards in the 1st box and 2nd box become $p_1 : (1 - p_1)\ \ w_2 : (1 - p_1)\ \ w_3 : (1 - p_1)\ w_4 : (1 - p_1)(1 - w_2 - w_3 - w_4)$ and $p_2 : (1 - p_2)\ \ w_2 : (1 - p_2)\ \ w_3 : (1 - p_2)\ w_4 : (1 - p_2)(1 - w_2 - w_3 - w_4)$ respectively. It is noteworthy that w_2, w_3 and w_4 are unknown to the interviewer and depend on the choice of the sampled person. However, $p_1\ and\ p_2\ (\neq p_1)$ are known to the interviewer due to the fixed values of m, m_1 and a.

Survey practitioners may use computerized RR devices like a virtual picker wheel. With the help of Google form investigators may record only the answers from the

respondents. The Google form should contain links of virtual RR devices and the options "yes" and "no" for each device. Respondents can enter into a particular RR device clicking on the link, mentioned in the form.

For example, the link https://pickerwheel.com/pw?id=LeCbM only allows respondents to click on the spin button of the virtual RR device. Picker wheel will show the choice or the statement while the spinning of the wheel is stopped. The respondent is requested to select option "yes" ("no") in Google form if the selected choice is "yes" ("no"). Otherwise he/she will provide a truthful response in terms of "yes" or "no" according to the statement visible on the computer screen.

Another link, https://pickerwheel.com/pw?id=aawQs is also a virtual RR device which can be edited by the respondents to construct their own RR device in the case of Model 2. But, they should follow the instructions given by investigators, strictly.

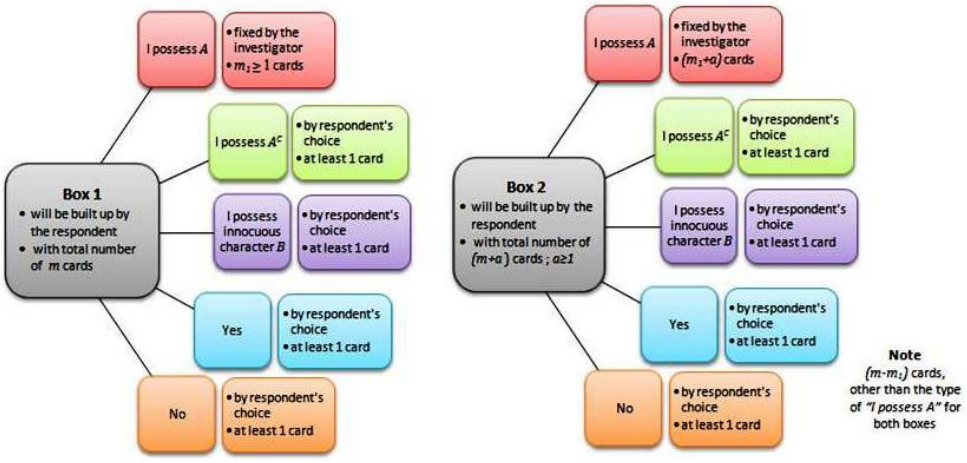


Figure 3.2: Model 2: Respondent-Specific RR Technique

Let I_i' and J_i' be the two independent responses of i^{th} sampled person which are defined as follows:

$$I_i' = \begin{cases} 1 & if\ the\ response\ is\ \ yes \\ 0 & if\ the\ response\ is\ \ no \end{cases}$$

and

$$J_i' = \begin{cases} 1 & if\ the\ response\ is\ \ yes \\ 0 & if\ the\ response\ is\ \ no. \end{cases}$$

Then, taking RR based expectations on I_i' and J_i', we get an unbiased estimator of y_i as

$$r_i' = \frac{(1-p_2)I_i' - (1-p_1)J_i'}{p_1 - p_2};\ \ p_1 \neq p_2\ .$$

Hence, the unbiased estimator of the population proportion θ is given by

$$e'_{HT} = \frac{1}{N}\sum_{i\in s}\frac{r'_i}{\pi_i}.$$

Therefore, the unbiased variance estimator is

$$v(e'_{HT}) = \frac{1}{N^2}\left[\sum_{i<j\in s}\sum\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{r'_i}{\pi_i} - \frac{r'_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_R(r'_i)}{\pi_i}\right]$$

where $v_R(r'_i) = \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(I'_i - J'_i)^2$.

4. Protection of Privacy for the Proposed RRTs

Chaudhuri et al. (2009) investigated the possibility of protecting the privacy of the respondent in RR context using varying probability sampling designs. Pal et al. (2020) recently developed the measure of privacy protection when an opportunity for divulging out the direct response is also given to each respondent along with a specific RR device. The respondents provide either direct response or randomized response, without divulging the type of response so exercised. We refer to Pal et al. (2020) here for a detailed account of the measure of privacy protection with two independent randomized responses.

Denoting the responses of the above respondent specific RR device as (R, R'), the posterior probability and measure of jeopardy of the i^{th} respondent may be written as $L_i(R, R')$ and $J_i(R, R')$ respectively.

Considering the prior probability $L_i\,(0 < L_i < 1)$ for the i^{th} respondent and applying Bayes' theorem, we get

$$Prob\left(A\middle|(R,R')\right) = \frac{L_i P(R|A) P\left(R'\middle|A\right)}{L_i P(R|A) P\left(R'\middle|A\right) + (1-L_i) P(R|A^c) P\left(R'\middle|A^c\right)} = L_i(R,R'). \quad (5)$$

Then, for i^{th} person, the response-specific jeopardy measure

$$J_i(R,R') = \frac{L_i(R,R')/L_i}{1-L_i(R,R')/1-L_i} \quad (6)$$

indicates the risk of divulging the respondent's status due to the responses (R, R').

Consequently, Chaudhuri et al. (2009) and Pal et al. (2020) suggested arithmetic mean $(\bar{J}_i)$ and geometric mean $(\tilde{J}_i)$ respectively as an overall measure of jeopardy.

Here, the possible responses (R, R') are (1,1),(0,0),(1,0) and (0,1).

Then, substituting $(R, R') \equiv (1,1)$ in (eq.5), we get the posterior probability **for the response (1, 1)** as,

$$L_i(1,1) = \frac{L_i P(I_i = 1|y_i = 1) P(J_i = 1|y_i = 1)}{L_i P(I_i = 1|y_i = 1) P(J_i = 1|y_i = 1) + (1 - L_i) P(I_i = 1|y_i = 0) P(J_i = 1|y_i = 0)}$$

$$= \frac{L_i\{p_1 + (1-p_1)(w_3+w_4)\}\{p_2 + (1-p_2)(w_3+w_4)\}}{L_i\{p_1 + (1-p_1)(w_3+w_4)\}\{p_2 + (1-p_2)(w_3+w_4)\} + (1-L_i)(1-p_1)(1-p_2)(w_2+w_3+w_4)^2}. \quad (7.1)$$

and substituting the same in (eq.6), the response specific jeopardy measure **for the response (1, 1)** is

$$J_i(1,1) = \frac{\{p_1+(1-p_1)(w_3+w_4)\}\{p_2+(1-p_2)(w_3+w_4)\}}{(1-p_1)(1-p_2)(w_2+w_3+w_4)^2}. \quad (7.2)$$

Note that $J_i(1,1) \to 1$ if $p_k + (1-p_k)(w_3+w_4) \to (1-p_k)(w_2+w_3+w_4)$; $k = 1,2$.

That imply $p_k \to (1-p_k)w_2$; $k = 1,2$.

i.e. the proportion of "I possess A" cards tending to the proportion of "I possess A^c" cards and $p_1 \to p_2$ are the advisable conditions for protecting the response (1,1) but the condition $p_1 \to p_2$ entails that the variance estimate $v_R(r_i)$ defined in eq. 2 tends to infinite.

Similarly, **for the response (0, 0),**

$$L_i(0,0) = \frac{L_i P(I_i = 0|y_i = 1)P(J_i = 0|y_i = 1)}{L_i P(I_i = 0|y_i = 1)P(J_i = 0|y_i = 1) + (1-L_i)P(I_i = 0|y_i = 0)P(J_i = 0|y_i = 0)}$$

$$= \frac{L_i(1-p_1)(1-p_2)(1-w_4)^2}{L_i(1-p_1)(1-p_2)(1-w_4)^2 + (1-L_i)\{p_1 + (1-p_1)(1-w_2-w_4)\}\{p_2 + (1-p_2)(1-w_2-w_4)\}} \quad (8.1)$$

and

$$J_i(0,0) = \frac{(1-p_1)(1-p_2)(1-w_4)^2}{\{p_1 + (1-p_1)(1-w_2-w_4)\}\{p_2 + (1-p_2)(1-w_2-w_4)\}}. \quad (8.2)$$

The necessary conditions for $J_i(0,0) \to 1$ are $p_k + (1-p_k)(1-w_2-w_4) \to (1-p_k)(1-w_4)$; $\forall\, k = 1,2$. That imply $p_k \to (1-p_k)w_2$; $\forall k = 1,2$.

In other words, the advisable conditions for protecting the response (0,0) are $p_1 \to p_2$ and the proportion of "I possess A" cards tending to the proportion of "I possess A^c" cards. But it entails that $v_R(r_i) \to \infty$.

Now, substituting $(R, R') \equiv (1,0)$ in (eq. 5) and (eq.6), we get

$$L_i(1,0) = \frac{L_i P(I_i = 1|y_i = 1)P(J_i = 0|y_i = 1)}{L_i P(I_i = 1|y_i = 1)P(J_i = 0|y_i = 1) + (1-L_i)P(I_i = 1|y_i = 0)P(J_i = 0|y_i = 0)}$$

$$= \frac{L_i\{p_1 + (1-p_1)(w_3+w_4)\}(1-p_2)(1-w_4)}{L_i\{p_1 + (1-p_1)(w_3+w_4)\}(1-p_2)(1-w_4) + (1-L_i)(1-p_1)(w_2+w_3+w_4)\{p_2 + (1-p_2)(1-w_2-w_4)\}} \quad (9.1)$$

and

$$J_i(1,0) = \frac{\{p_1+(1-p_1)(w_3+w_4)\}(1-p_2)(1-w_4)}{(1-p_1)(w_2+w_3+w_4)\{p_2+(1-p_2)(1-w_2-w_4)\}} \quad (9.2)$$

respectively.

This $J_i(1,0) \to 1$ if $p_1 + (1-p_1)(w_3+w_4) \to (1-p_1)(w_2+w_3+w_4)$ and $p_2 + (1-p_2)(1-w_2-w_4) \to (1-p_2)(1-w_4)$ which imply $p_1 \to (1-p_1)w_2$ and $p_2 \to (1-p_2)w_2$.
i.e. if the proportion of "*I possess A*" cards tends to the proportion of "*I possess* A^c" cards and $p_1 \to p_2$, then $J_i(1,0)$ converges to 1 with $v_R(r_i) \to \infty$.

For the response (0, 1),

$$L_i(1,0) = \frac{L_i P(I_i=0|y_i=1)P(J_i=1|y_i=1)}{L_i P(I_i=0|y_i=1)P(J_i=1|y_i=1) + (1-L_i)P(I_i=0|y_i=0)P(J_i=1|y_i=0)}$$
$$= \frac{L_i\{p_2+(1-p_2)(w_3+w_4)\}(1-p_1)(1-w_4)}{L_i\{p_2+(1-p_2)(w_3+w_4)\}(1-p_1)(1-w_4) + (1-L_i)(1-p_2)(w_2+w_3+w_4)\{p_1+(1-p_1)(1-w_2-w_4)\}} \quad (10.1)$$

and

$$J_i(0,1) = \frac{\{p_2+(1-p_2)(w_3+w_4)\}(1-p_1)(1-w_4)}{(1-p_2)(w_2+w_3+w_4)\{p_1+(1-p_1)(1-w_2-w_4)\}}. \quad (10.2)$$

Now, if $p_k \to (1-p_k)w_2$; $k = 1,2$ then $J_i(0,1)$ converges to 1. But in such a case $v_R(r_i) \to \infty$.

Thus, considering the geometric mean $\tilde{J}_i$ as the overall measure of jeopardy for the proposed RRT, we get

$$\tilde{J}_i = \{J_i(1,1) \times J_i(0,0) \times J_i(1,0) \times J_i(0,1)\}^{1/4}$$
$$= \left\{\frac{p_1+(1-p_1)(w_3+w_4)}{p_1+(1-p_1)(1-w_2-w_4)} \frac{p_2+(1-p_2)(w_3+w_4)}{p_2+(1-p_2)(1-w_2-w_4)} \frac{(1-w_4)^2}{[\![(w]\!]_2+w_3+w_4)^2}\right\}^{1/2}. \quad (11)$$

(eq. 11) converges to 1
if $(1-w_4) \to [\![(w]\!]_2 + w_3 + w_4)$ i.e. $(1-w_2-w_3-w_4) \to w_4$ or

$$\frac{p_k+(1-p_k)(w_3+w_4)}{p_k+(1-p_k)(1-w_2-w_4)} = \frac{(w_2+w_3+w_4)}{1-w_4} \text{ i.e.} p_k \to (1-p_k)w_2 \text{ ; } k = 1,2.$$

In other words, the proposed RR techniques ensure maximum protection if the proportion of "Yes" and "No" cards are the same or the proportion of "I possess A" and "I possess A^c" cards are the same.

Hence, it is advisable to apply this proposed RR techniques with the RR devices having at least one of the following properties:

i) "Yes" and "No" cards are in the same proportion in the devices
ii) "I possess A" and "I possess A^c cards are the same in proportion for both devices.

In Model 1, such restrictions on model parameters may be followed. However, in Model 2 i.e. Respondent-specific RR technique, the restrictions are not guaranteed as RR devices are made by respondents.

In the later section, the measure of jeopardy is calculated numerically for a different combination of p_1 and p_2.

5. Simulation Study

In this section, we investigate the performance of the proposed RRTs using fivetypes of cards through a simulation study. For this, we consider a fictitious data consisting of reckless driving history with weekly expenses of $N = 116$ undergraduate students under 20 years of age. Here, the parameter of interest is the proportion of the students who broke the traffic rules last year. Let the population proportion be defined as $\theta = \frac{1}{N}\sum_{i\in U} y_i$, treating y as a qualitative sensitive variable - *"Breaking the traffic rules"*. For the above study, $\theta = 0.606838$. The innocuous character x is taken here as *"Interested in painting"*. The size measure variable z- *"Weekly expenditure"* is used to draw samples in varying probability sampling scheme.

In order to study the competitiveness concerning the proposed RRTs, the simulation study is performed for different sample sizes and the samples are drawn by Lahiri- Midzuno- Sen [1951, 1952, 1953] sampling strategy, where the first unit is selected with the probability $p_i^* = \frac{z_i}{\sum_U z_i}$ and the remaining units are selected by SRS without replacement from the remaining units in the population after the first draw.

Since the variable y represents the sensitive feature A, it is not directly assessable and is estimated for each respondent through an unbiased estimator defined in eq.1. Then, employing eq. 3 and eq. 4, we get an unbiased estimate for the population proportion and unbiased variance estimate respectively. Here, the 1st order and 2nd order inclusion probabilities for Lahiri- Midzuno- Sen scheme are $\pi_i = p_i^* + \frac{(1-p_i^*)(n-1)}{N-1}$ and $\pi_{ij} = \frac{(n-1)(N-n)(p_i^*+p_j^*)+(n-1)(n-2)}{(N-1)(N-2)}$ respectively.

To judge the efficacy of the RR models, different parametric combinations (p_1, p_2) are taken in Table 5.1, considering 1000 replications of samples for each sample size. Efficacy of the proposed RRTs for different sample sizes are judged by the Average Coverage Probabilities (ACP), the Average Coefficient of Variation (ACV) and the Average Length (AL) of the 95% Confidence Intervals (CI) based on $e_{HT} \pm 1.96\sqrt{v(e_{HT})}$. The point estimator will be judged well if the ACV, the average over 1000 replications of estimated coefficient of variations$\left(CV = 100 \times \frac{\sqrt{v(e_{HT,RR})}}{e_{HT,RR}}\right)$, has a small magnitude, preferably less than 10% or at most 30%. The percentage of cases for which 95% CI covers the true value of the parameter is called ACP. ACP values close to 95% will be preferred. AL is defined as $2 \times 1.96\sqrt{[\![v(e]\!]_{HT}}$. Smaller ACV, AL values along with the ACP value close to 95% are preferred. In addition, absolute relative bias (ARB) of an

unbiased estimator and average variance estimate (AVE) are calculated as $\left|\frac{\bar{e}-\theta}{\theta}\right|$ and $\frac{1}{1000}\sum_{k=1}^{1000} v(e_k)$ respectively, where $\bar{e} = \frac{1}{1000}\sum_{k=1}^{1000} e_k$ is the average of 1000 estimates of θ.

Figures 5.1-5.3, based on Table 5.1, represent the performance of proposed RRTs for different sample sizes. The values of ACV, ACP, AL, ARB and AVE for different parametric combinations are shown in the same graph. To do this, the values of AL, ARB and AVE from Table 5.1 are taken as $100 \times AL$, $1000 \times ARB$ "and" $1000 \times AVE$. The vertical axes of the graphs indicate the values and the horizontal axes indicate different parametric combinations (p_1, p_2) of the RRTs.

As shown in Table 5.1,

i) ACP values are greater than 95%.

ii) ACV and AL values are decreasing as the sample size increases. If p_1 and p_2 are close to each other, ACV and AL values are relatively high.

For example:

if $(p_1, p_2) = (0.4, 0.6)$ and $(0.4, 0.2)$, ACV values are beyond the acceptable range.

iii) Considering the sample size $n = 25$, the parametric combinations (0.4, 0.7) and (0.57, 0.79) are equally competitive and perform moderately as their ACV, AL and AVE values are much lower than others. Figure 5.1 sheds light on this fact.

iv) The parametric combinations (0.4, 0.7) and (0.57, 0.79) perform well in terms of AVE, ACV, ACP, ARB and AL for both sample sizes 30 and 35. Figures 5.2 and 5.3 highlight this finding.

Table 5.1: ACV, ACP, AL, and ARB for the proposed RRTs using fivetypes of cards $(unknown\ \theta = 0.606838)$

p_1	p_2	n	$\bar{e} = \frac{1}{1000}\sum_{k=1}^{1000} e_k$	$AVE = \frac{1}{1000}\sum_{k=1}^{1000} v(e_k)$	ACV	ACP	AL	ARB
0.4	0.6	25	0.61445	0.04718	41.07467	98.7	0.83896	0.01255
0.4	0.6	30	0.61157	0.03981	35.25056	99	0.77313	0.00781
0.4	0.6	35	0.62353	0.03267	30.60223	98.7	0.70227	0.02751
0.4	0.2	25	0.68875	0.08238	46.29767	96.3	1.10524	0.13498
0.4	0.2	30	0.65915	0.07464	43.47351	97.6	1.05719	0.08620
0.4	0.2	35	0.65619	0.06093	42.15731	97.4	0.95809	0.08133
0.4	0.7	25	0.62845	0.01998	23.3216	97.4	0.54935	0.03561
0.4	0.7	30	0.61883	0.01730	21.89433	98	0.51232	0.01976
0.4	0.7	35	0.62145	0.01408	19.54249	98.3	0.46274	0.02408
0.57	0.79	25	0.62094	0.01941	23.46384	96.3	0.54168	0.02324
0.57	0.79	30	0.62819	0.01629	21.01769	96.1	0.49702	0.03519
0.57	0.79	35	0.62929	0.01336	18.78795	98.2	0.45075	0.03699
0.57	0.3	25	0.65181	0.03375	30.09188	96.6	0.71012	0.07411
0.57	0.3	30	0.65319	0.02885	27.31937	96.7	0.65903	0.07639
0.57	0.3	35	0.65078	0.02344	24.54986	96.8	0.59563	0.07241

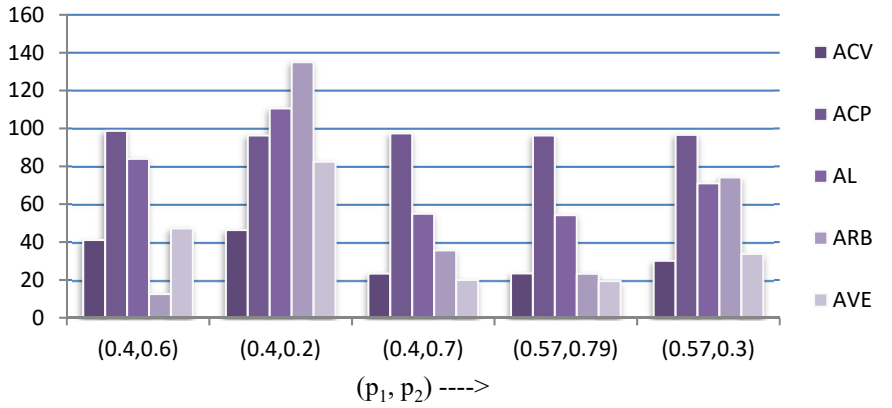


Figure 5.1: Performances of the proposed RRTs for the sample size n=25

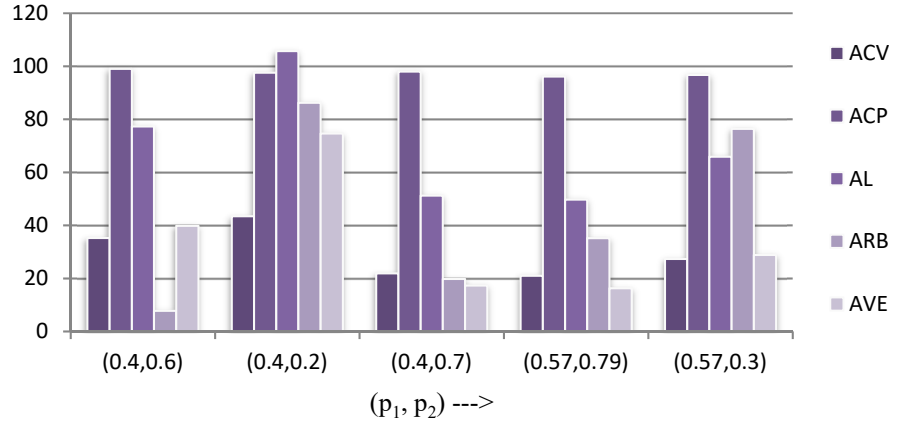


Figure 5.2: Performances of the proposed RRTs for the sample size n=30

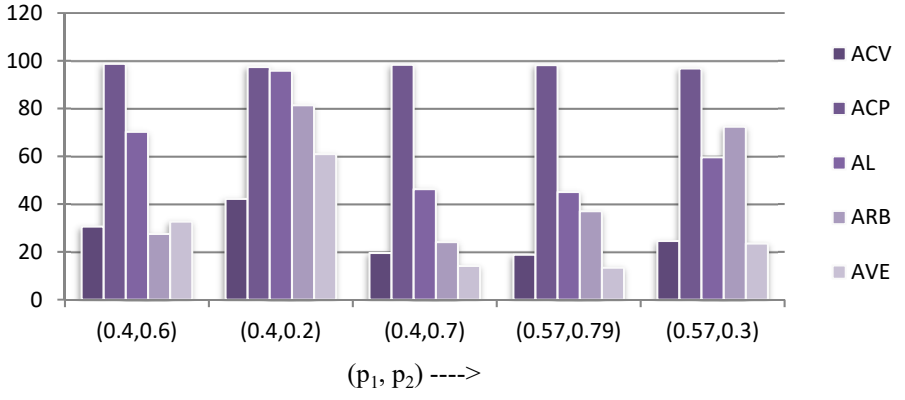


Figure 5.3: Performances of the proposed RRTs for the sample size n=35

Table 5.2 demonstrates how well the privacy of sensitive features may be protected for the proposed models. For this purpose, the response-specific jeopardy measures $J_i(R, R')$ are computed taking different combinations of $(p_1, p_2, w_2, w_3, w_4)$. The overall measure of jeopardy $(\tilde{J}_i)$ is shown in the last column of the mentioned table.
As noted in Section 4,

i) $J_i(R, R') \to 1$ if $p_1 \to p_2$ and the proportion of *"I possess A"* cards tends to the proportion of *"I possess A^c"* cards, which also ensures that the overall measure of jeopardy tends to 1.

In Table 5.2, we have tried to show such a condition taking the combinations of $(p_1, p_2, w_2, w_3, w_4)$values as (0.33,0.327, 0.49,0.2,0.3) and (0.2,0.22,0.27,0.2,0.4).

ii) In both RR devices if the proportion of "Yes" cards tends to the proportion of "No" cards, the measure of jeopardy $\tilde{J}_i$ will tend to 1.

In Table 5.2, the following parameters, satisfying the above conditions, may be taken as follows:

$(p_1, p_2, w_2, w_3, w_4)$: (0.4, 0.45, 0.2,0.3,0.25) , (0.2,0.3,0.29,0.2,0.27) and (0.4,0.6,0.2,0.3,0.25).

In Table 5.3 and Table 5.4, we have shown the response-specific jeopardy measure along with the overall measure of jeopardy following the suggestion in Chaudhuri et al. (2009) for Warner's (1965) RRT and Greenberg et al.'s (1969) RRT. Here, the arithmetic mean $(\bar{J}_i)$ of all the response-specific jeopardy measure is considered as the overall measure of jeopardy. We refer to Chaudhuri et al. (2009) for detailed derivation of the measures. The results in Table 5.2 can be compared easily with Table 5.4. For better comparison, we have taken the same $p_1$and p_2 values which represent the proportions of "I possess A" cards for proposed RRTs (see Section 3) and Greenberg et al.'s RRT. From there, we may conclude that the proposed RRTs perform better than the Greenberg et al.'s RRT in terms of protecting privacy.

Table 5.2: Protection of Privacy for Proposed RRTs

p_1	p_2	w_2	w_3	w_4	$J_i(1,1)$	$J_i(0,0)$	$J_i(1,0)$	$J_i(0,1)$	$\tilde{J}_i$
0.4	0.6	0.2	0.3	0.4	3.7119	0.1776	0.4795	1.375	0.8120
0.4	0.6	0.2	0.3	0.2	4.7619	0.2406	0.6349	1.8045	1.0704
0.4	0.45	0.2	0.3	0.25	2.9593	0.3379	0.8892	1.1245	1
0.33	0.38	0.49	0.2	0.3	1.1270	0.8476	0.8528	1.1201	0.9774
0.33	0.327	0.49	0.2	0.3	0.9984	1.0022	1.0085	0.9922	1.0003
0.57	0.69	0.2	0.45	0.2	7.8635	0.1176	0.6579	1.4056	0.9617
0.6	0.5	0.2	0.45	0.2	4.9100	0.1905	1.2647	0.7395	0.9671
0.2	0.22	0.27	0.2	0.4	0.9905	1.0141	0.9578	1.0488	1.0023
0.2	0.3	0.29	0.2	0.27	1.1201	0.8892	0.7962	1.2509	0.998
0.2	0.3	0.4	0.2	0.3	0.8598	1.2228	0.8006	1.3131	1.0254
0.4	0.6	0.2	0.3	0.25	4.4340	0.2255	0.5935	1.6849	1

Table 5.3: Protection of Privacy for Warner's RRT

p_1	$J_i(1) = \frac{p_1}{1-p_1}$	$J_i(0) = \frac{1-p_1}{p_1}$	$\bar{J}_t$
0.2	0.25	0.4	2.125
0.33	0.49254	2.0303	1.26142
0.4	0.66667	1.5	1.08333
0.51	1.04081	0.96078	1.0008
0.57	1.32558	0.75439	1.03998
0.6	1.5	0.66667	1.08333
0.69	2.22581	0.44927	1.33754

Table 5.4: Protection of Privacy for Greenberg et al.'s RRT

p_1	p_2	$J_i(1,1) = \frac{p_1 p_2}{(1-p_1)(1-p_2)}$	$J_i(0,0) = \frac{(1-p_1)(1-p_2)}{p_1 p_2}$	$J_i(1,0) = \frac{p_1(1-p_2)}{p_2(1-p_1)}$	$J_i(0,1) = \frac{p_2(1-p_1)}{p_1(1-p_2)}$	$\bar{J}_t$
0.4	0.6	1	1	0.4444	2.25	1.1736
0.4	0.45	0.5455	1.8333	0.8149	1.2273	1.1052
0.33	0.38	0.3019	3.3126	0.8036	1.2444	1.4156
0.57	0.69	2.9505	0.3389	0.5955	1.6791	1.3910
0.6	0.5	1.5	0.6667	1.5	0.6667	1.0833
0.2	0.3	0.1071	9.3333	0.5833	1.7143	2.9345

6. Concluding Remarks

In this work, we have attempted to introduce two proposed methods permitting five questions to the respondents. Model 2, i.e. Respondent-specific RRT, is an extension of the proposed Model 1. In the proposed Model 2, respondents are allowed to build their own RR devices. It is anticipated that the participation of respondents in the RR survey will be better than other existing RRTs. Our simulation study gives us satisfactory results in terms of ACP, ACV and AL values. We have also calculated protection of privacy measure of the proposed RRTs, which is close to 1. The findings described in this study will stimulate researchers and survey practitioners to apply the response-specific RRT in real surveys. Respondents will co-operate freely in the survey methods as they are building their own RR devices.

Acknowledgement

The authors gratefully acknowledge the support received from two referees, which enabled them to produce this revised and improved version out of the original submission.

References

Anderson, H., (1975b). Efficiency versus Protection in a General RR model, *Technical Report 10, University of Lund,* Lund, Sweden.

Anderson, H., (1975c). Efficiency versus Protection in RR Designs. *Mimeo notes, University of Lund,* Lund, Sweden.

Anderson, H., (1975a). Efficiency versus Protection in the RR for Estimating Proportions. *Technical Report 9, University of Lund, Lund, Sweden*.

Arnab, R., Mothupi, T., (2015). Randomized Response Techniques: A Case Study of the Risky Behaviors' of Students of a Certain University. *Model Assisted Statistics and Applications*, Vol. 10, pp. 421–430.

Barabesi, L., Diana, G., Perri, P. F., (2013). Design-based distribution function estimation for stigmatized population. *Metrika*, Vol. 76, pp. 919–935.

Boruch, R. F., (1972). Relations Among Statistical Methods for Assuring Confidentiality of Social Research Data. *Social Science Research*, Vol. 1, pp. 403–414.

Chaudhuri, A., (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton: CRC Press.

Chaudhuri, A., Christofides, T. C., Rao, C. R., (2016). Handbook of Statistics, Data Gathering, *Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, Vol. 34, NL: Elsevier.

Chaudhuri, A., Christofides, T. C., Saha, A., (2009). Protection of Privacy in Efficient Application of Randomized Response Techniques. *Statistical Methods and Applications*, Vol. 18, pp. 389–418.

Chong, A. C., Chu, A. M., So, M. K., Chung, R. S., (2019). Asking Sensitive Questions Using the Randomized Response Approach in Public Health Research: An Empirical Study on the Factors of Illegal Waste Disposal. *International Journal of Environmental Research and Public Health*, Vol. 16, pp. 970.

Diana, G., Perii, P. F., (2013). Randomized Response Surveys: A note on some privacy protection measures. *Model Assisted Statistics and Applications*, Vol. 8, pp. 19–28.

Fox, J., Veen, D., Klotzke, K., (2018). Generalized linear mixed models for randomized responses. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* Vol. 15, pp. 1–18.

Greenberg, B. G., Abul-Ela, A. L., Simmons, W. R.,Horvitz, D. G., (1969). The unrelated question randomized response model: theoretical framework. *Journal of American Statistical Association*, Vol. 64, pp. 520–539 .

Horvitz, D. G., Thompson, D. J., (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, Vol. 47, pp. 663–685 .

Hout, A. V., Heijden, P. G., Gilchrist, R., (2007). The logistic regression model with response variables subject to randomized response. *Computational Statistics & Data Analysis,* Vol. 51, pp. 6060–6069.

Kuk, A. Y., (1990). Asking Sensitive Questions Indirectly. *Biometrika*, Vol. 77, pp. 436–438.

Lahiri, D. B., (1951). A method of sample selection providing unbiased ratio estimates. *Bulletin of International Statistical Institute*, Vol. 3, pp. 133–140.

Lanke, J., (1975). On the choice of the unrelated question in Simmons' version of randomized response. *Journal of American Statistical Association*, Vol. 70, pp. 80–83 .

Lanke, J., (1976). On the degree of protection in randomized interviews. *International Statistical Review*, Vol. 44, pp. 197–203.

Leysieffer, R. W., Warner, S. L., (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of American Statistical Association*, Vol. 71, pp. 649–656 .

Midzuno, H., (1952). On the sampling system with probability proportional to the sum of the sizes. *Annals of Institute of Statistical Mathematics*, Vol. 3, pp. 99–107.

Pal, S., Chaudhuri, A., Patra, D., (2020). How privacy may be protected in Optional Randomized Response Surveys. *Statistics in Transition*, Vol. 21, pp. 61–87.

Sen, A. R., (1953). On the estimator of the variance in sampling with varying probabilities. *Journal of Indian Society of Agricultural Statistics*, Vol. 5, pp. 119–127.

Van der Heijden, P. G., Van Gils, G., Bouts, J., Hox, J., (2000). A Comparison of Randomized Response, Computer-Assisted Self Interview, and Face to Face Direct Questioning: Eliciting Sensitive Information in the Context of Welfare and Unemployment Benefit. *Sociological Research & Methods*, Vol. 28, pp. 505–537 .

Warner, S. L., (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of American Statistical Association*, Vol. 60, pp. 63–69.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 71–92, https://doi.org/10.59170/stattrans-2023-052
Received – 08.11.2020; accepted – 17.01.2023

Odd log-logistic generalised Lindley distribution with properties and applications

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

Abstract

In this paper, a new three-parameter lifetime model, called the *odd log-logistic generalised Lindley* distribution, is introduced. Some structural properties of the new distribution including ordinary and incomplete moments, quantile and generating functions and order statistics are obtained. The new density function can be expressed as a linear mixture of exponentiated Lindley densities. Different methods are discussed to estimate the model parameters and a simulation study is carried out to show the performance of the new distribution. The importance and flexibility of the new model are also illustrated empirically by means of two real data sets. Finally, Bayesian analysis and Gibbs sampling are performed based on the two real data sets.

Key words: Lindley distribution, odd log-logistic generalised family, moments, Bayesian analysis, simulation study.

1. Introduction

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

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

[1]Golestan University, Gorgan, Iran. E-mail: vahidranjbar@gmail.com. ORCID: https://orcid.org/0000-0003-3743-0330.
[2]University of Hormozgan, Bandar Abbas, Iran. ORCID: https://orcid.org/0000-0002-5343-8597.
[3]Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran. ORCID: https://orcid.org/0000-0003-4176-9708.
[4]Persian Gulf University, Bushehr, Iran. ORCID: https://orcid.org/0000-0001-6638-2185.

beta generated. Recently, the extended exponentiated-G (EE-G) family has been defined by Alizadeh et al. (2018a).

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

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

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

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

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

Many authors have published various extensions of the Lindley distribution recently. For example, a three-parameter generalization of the Lindley distribution proposed by Zakerzadeh and Dolati (2009), Nadarajah et al. (2011) defined a generalized Lindley distribution, a new generalized Lindley distribution based on the weighted mixture of two gamma distributions was studied by Abouammoh et al. (2015).

Asgharzadeh et al. (2016) and Asgharzadeh et al. (2018) introduced a weighted Lindley distribution and Weibull Lindley distribution, respectively and Alizadeh et al. (2017a), Alizadeh et al. (2017b), Alizadeh et al. (2018b) proposed several generalizations of the Lindley distribution based on the odd log-logistic model. Given the vast amount of papers published recently, we can only mention a few of the most recent contributions: Gomes-Silva et al. (2017), Afify et al. (2019) and Alizadeh et al. (2019).

The problem here is to construct a new extension of the Lindley distribution that may be useful for complex situations. The suggested distribution provides an acceptable flexibility based on the pdf and hazard rate function and it can be applied in actuarial science, finance, bioscience, telecommunications and lifetime data analysis. Alzaatreh et al. (2013) defined a generalization of odd ratio and it called as transformer (T-X) generator, where $W[G(x)] = \frac{G(x)^{\alpha}}{[1-G(x)]^{\beta}} = \frac{G(x)^{\alpha}}{1-\left\{1-[1-G(x)]^{\beta}\right\}}$ is an increasing and continuous function of $G(x)$. One can say

that $W[G(x)] = \frac{G(x)^{\alpha}}{[1-G(x)]^{\beta}}$ for integer α, β is a relative odd ratio of two systems, the first system with α parallel subcomponents and the second with β series subcomponents, which are useful in reliability theory. Motivated by Alzaatreh et al. (2013), we propose a new lifetime distribution called *odd log-logistic generalized Lindley* (OLLG-L) distribution by integrating the log-logistic density function, which yields the cdf

$$F(x) = \frac{\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x}}\right]^{\alpha\beta}}{\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x}}\right]^{\alpha\beta} + \left[1-\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x}}\right]^{\alpha}\right]^{\beta}} \tag{3}$$

where $\alpha, \beta > 0$ are the extra shape parameters. Then, the corresponding pdf of the OLLGL distribution is given by

$$f(x) = \frac{\alpha\beta\lambda^2(1+x)\mathrm{e}^{-\lambda \mathrm{x}}\left[1-(1+\frac{\lambda \mathrm{x}}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x}}\right]^{\alpha\beta-1}\left[1-\left[1-(1+\frac{\lambda \mathrm{x}}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x}}\right]^{\alpha}\right]^{\beta-1}}{(1+\lambda)\left\{\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x}}\right]^{\alpha\beta} + \left[1-\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda \mathrm{x}}\right]^{\alpha}\right]^{\beta}\right\}^2} \tag{4}$$

A random variable X with pdf (4) is denoted by $X \sim \mathrm{OLLGL}(\alpha, \beta, \lambda)$. The OLLGL distribution is more flexible than the Lindley distribution and allows for greater flexibility of the tails.

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

- If $\alpha = 1$, then X reduces to the Odd Log-Logistic Lindley (OLL-L).
- If $\beta = 1$, then X reduces to the Generalized Lindley (GL).
- For $\alpha = \beta = 1$, X is ordinary Lindley.

Plots of the density function for the OLLGL distribution are shown in Figure 1 for several values of parameters. As seen from Figure 1, the density function can take various forms depending on the parameter values. Both unimodal, symmetric, skewed, and monotonically decreasing shapes appear to be possible.

The rest of the paper is organized as follows. In Section 2, main properties of the OLLGL distribution such as moments, parameters estimation and asymptotic properties are obtained. A simulation study is reported in Section 3. In Section 4, the performance and application of the OLLGL distribution are evaluated using a real data set. Bayesian inference and Gibbs sampling procedure for the considered data sets are investigated in Section 5. Finally, some conclusions are stated in Section 6.

2. Main Properties

2.1. Survival and Hazard Rate Functions

The survival function is a function that gives the probability that a patient, device, or other object of interest will survive beyond any given specified time. The survival function

is also known as the survivor function or reliability function. We obtain the survival function corresponding to (3) as

$$S(x;\alpha,\beta,\lambda) = 1 - \frac{\left[1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x}\right]^{\alpha\beta}}{\left[1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x}\right]^{\alpha\beta} + \left[1-\left[1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x}\right]^{\alpha}\right]^{\beta}} \tag{5}$$

In reliability studies, the hazard rate function (hrf) is an important characteristic and fundamental to the design of safe systems in a wide variety of applications. The hrf of the OLLGL distribution takes the form

$$h(x;\alpha,\beta,\lambda) = \frac{\alpha\beta\lambda^2(1+x)e^{-\lambda x}\left[1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x}\right]^{\alpha\beta-1}}{(1+\lambda)\left[1-\left[1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x}\right]^{\alpha}\right]\left\{\left[1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x}\right]^{\alpha\beta} + \left[1-\left[1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x}\right]^{\alpha}\right]^{\beta}\right\}} \tag{6}$$

where $\alpha > 0$, $\beta > 0$ and $\lambda > 0$.

Plots for the hrfs of the OLLGL distribution for several parameter values are displayed in Figure 1. As seen in Figure 1, the hrf of the OLLGL distribution has very flexible shapes such as increasing, decreasing, upside-down, bathtub. It is evident that the OLLGL distribution is more flexible than the Lindley distribution, i.e. the additional parameters $\alpha > 0, \beta > 0$ allow for a high degree of flexibility of the OLLGL distribution. This attractive flexibility makes the hrf of the OLLGL useful for non-monotone empirical hazard behaviour, which is more likely to be observed in real life situations.

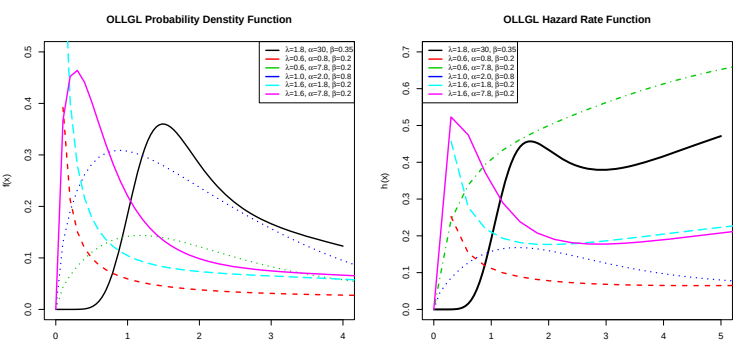


Figure 1: Plots of the density and hazard rate functions for the OLLGL distribution for some selected values.

2.2. Quantile Function

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

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

For $\alpha = \beta$, the closed form for the quantile function can be obtained. Then, we define

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

for $0 < p < 1$. Substituting $Z(p) = -1 - \lambda - \lambda Q(p)$, one can write (8) as

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

Hence, the solution $Z(p)$ is given by

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

where $W_{-1}[.]$ is the negative branch of the Lambert function (Corless (1996)). Inserting (10), we obtain

$$Q(p) = -1 - \frac{1}{\lambda} - \frac{1}{\lambda} W_{-1}\left\{-(1+\lambda) \mathrm{e}^{-1-\lambda} \left[1 - \frac{\mathrm{p}^{\frac{1}{\alpha\beta}}}{(\mathrm{p}^{\frac{1}{\beta}} + (1-\mathrm{p})^{\frac{1}{\beta}})^{\frac{1}{\alpha}}}\right]\right\}. \tag{11}$$

Note that the particular case of (11) for $\alpha = \beta = \gamma = 1$ is derived by Jodr´a (2010).

Now, we propose following two different algorithms for generating random data from the OLLGL distribution for the case $\alpha = \beta$.

(a) The first algorithm is based on generating random data from the Lindley distribution mixturing the exponential and gamma distributions.

Algorithm 1 (Mixture form of the Lindley distribution)

- Generate $U_i \sim \text{Uniform}(0,1), \quad i = 1, \ldots, n;$
- Generate $V_i \sim \text{Exponential}(\lambda), \quad i = 1, \ldots, n;$
- Generate $W_i \sim \text{Gamma}(2, \lambda), \quad i = 1, \ldots, n;$
- If $\frac{U^{\frac{1}{\alpha\beta}}}{(U^{\frac{1}{\beta}} + (1-U)^{\frac{1}{\beta}})^{\frac{1}{\alpha}}} \leq \frac{\lambda}{1+\lambda}$ set $X_i = V_i$, otherwise, set $X_i = W_i, \quad i = 1, \ldots, n.$

(b) The second algorithm is based on generating random data from the inverse cdf in (3) of the OLLGL distribution.

Algorithm 2 (Inverse cdf)

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

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

2.3. Mixture representations for the pdf and cdf

The cdf and pdf can be written as mixture representations and such forms of cdf and pdf can be used to derive some mathematical properties, e.g. moments, moments of residual life and incomplete moments. To this purpose, first let us remind inverse of a power series using the following Remark.

Remark 1 *(Gradshteyn and Ryzhik (2007), page 17)*
Inverse of a power series $\sum_{k=0}^{\infty} b_k x^k$ *is*

$$\frac{1}{\sum_{k=0}^{\infty} b_k x^k} = \sum_{k=0}^{\infty} c_k x^k,$$

where $c_0 = \frac{1}{b_0}$ *and for* $k \geq 1$, *and* $c_k = -\frac{1}{b_0}\sum_{r=1}^{k} c_{k-r} b_r$.

To obtain the mixture representation of the cdf of OLLGL, note that for any $0 < u < 1$,

$$\begin{aligned} u^{\alpha\beta} &= \sum_{i=1}^{\infty}(-1)^i \binom{\alpha\beta}{i}(1-u)^i = \sum_{i=1}^{\infty}\sum_{k=0}^{i}(-1)^{i+k}\binom{\alpha\beta}{i}\binom{i}{k}u^k \\ &= \sum_{k=0}^{\infty}\sum_{i=k}^{\infty}(-1)^{i+k}\binom{\alpha\beta}{i}\binom{i}{k}u^k = \sum_{k=0}^{\infty} a_k u^k, \end{aligned}$$

where $a_k = a_k(\alpha\beta) = \sum_{i=k}^{\infty}(-1)^{i+k}\binom{\alpha\beta}{i}\binom{i}{k}$. By similar argument, we have

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

where $b_k = a_k(\alpha\beta) + \sum_{j=0}^{\infty}(-1)^j \binom{\beta}{j} a_k(\alpha j)$. Now, using Remark 1, we get

$$F(x) = \frac{[1-(1+\frac{\lambda}{1+\lambda}x)e^{-\lambda x}]^{\alpha\beta}}{\sum_{k=0}^{\infty} b_k[1-(1+\frac{\lambda}{1+\lambda}x)e^{-\lambda x}]^k} = \sum_{k=0}^{\infty} c_k[1-(1+\frac{\lambda}{1+\lambda}x)e^{-\lambda x}]^{k+\alpha} = \sum_{k=0}^{\infty} c_k G(x;\lambda)^{k+\alpha\beta}, \tag{12}$$

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

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

The equation (12) can be interpreted as a linear combination of generalized Lindley distribution. Using this equation, the mixture representation of pdf is given by

$$f(x) = \sum_{k=0}^{\infty} (k+\alpha\beta)\, c_k\, g(x;\lambda) G(x;\lambda)^{k+\alpha\beta-1}. \tag{13}$$

2.4. Moments and Moment Generating Function

Some of the most important features and characteristics of a distribution can be studied through moments (e.g., central tendency, dispersion, skewness and kurtosis). Now, we obtain ordinary moments and the moment generating function (mgf) of the OLLGL distribution. Nadarajah et al. (2011) defined the following equation for the ordinary moments as

$$A(a,b,c,\delta) = \int_0^{\infty} x^c (1+x) \left[1 - \left(1 + \frac{bx}{b+1}\right) \mathrm{e}^{-bx}\right]^{a-1} \mathrm{e}^{-\delta x} dx \tag{14}$$

which can be used to produce ordinary moments (μ_r'). Then, we have

$$A(a,b,c,\delta) = \sum_{l=0}^{\infty} \sum_{r=0}^{l} \sum_{s=0}^{r+1} \binom{a-1}{l} \binom{l}{r} \binom{r+1}{s} \frac{(-1)^l b^r \Gamma(s+c+1)}{(1+b)^l (bl+\delta)^{c+s+1}}. \tag{15}$$

From equations (12) and (13), we obtain the ordinary moments of the OLLGL distribution as

$$\mu_r' = E[X^r] = \frac{\lambda^2}{1+\lambda} \sum_{k=0}^{\infty} (k+\alpha\beta)\, c_k A(k+\alpha\beta, \lambda, r, \lambda). \tag{16}$$

We now provide a formula for the conditional moments of the OLLGL distribution. Nadarajah et al. (2011) defined the following equation for the conditional moments

$$L(a,b,c,\delta,t) = \int_t^{\infty} x^c (1+x) \left[1 - \left(1 + \frac{bx}{b+1}\right) \mathrm{e}^{-bx}\right] \mathrm{e}^{-\delta x} dx. \tag{17}$$

Using the generalized binomial expansion, we have

$$L(a,b,c,\delta,t) = \sum_{l=0}^{\infty} \sum_{r=0}^{l} \sum_{s=0}^{r+1} \binom{a-1}{l} \binom{l}{r} \binom{r+1}{s} \frac{(-1)^l b^r \Gamma(s+c+1, (bl+\delta)t)}{(1+b)^l (bl+\delta)^{c+s+1}} \tag{18}$$

where

$$\Gamma(a,x) = \int_x^{\infty} t^{a-1} \mathrm{e}^{-t}\, dt \tag{19}$$

denotes the incomplete gamma function. From equations (13) and (18), we obtain the conditional moments of the OLLGL distribution as

$$\mu_r'(t) = E[X^r | X > t] = \frac{\lambda^2}{1+\lambda} \sum_{k=0}^{\infty} (k+\alpha\beta)\, c_k L(k+\alpha\beta, \lambda, r, \lambda, t). \tag{20}$$

The incomplete moments of the OLLGL distribution can be calculated directly from (20).

The mgf of a random variable provides the basis of an alternative route to analytical results com-

pared with working directly with its pdf and cdf. Using (13) and (15), we obtain

$$M_X(t) = E\left[e^{tX}\right] = \frac{\lambda^2}{1+\lambda}\sum_{k=0}^{\infty}(k+\alpha)\,c_k A(k+\alpha\beta,\lambda,0,\lambda-t).$$

Remark 2 *The central moments (μ_n) and cumulants (κ_n) of X are easily obtained from (16) as*

$$\mu_n = \sum_{k=0}^{n}(-1)^k\binom{n}{k}\mu_1'^k\,\mu'_{n-k} \qquad \text{and} \qquad \kappa_n = \mu'_n - \sum_{k=1}^{n-1}\binom{n-1}{k-1}\kappa_k\,\mu'_{n-k},$$

respectively, where $\kappa_1 = \mu'_1$. Thus, $\kappa_2 = \mu'_2 - \mu'^2_1$, $\kappa_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1$, etc.

Figure 2 shows skewness and kurtosis measures of the OLLGL distribution. The skewness and kurtosis are calculated from the ordinary moments given in (16) for $\lambda = 2$. Figure 2 shows that skewness and kurtosis are very sensitive for the shape parameters and it indicates the importance of the proposed distribution.

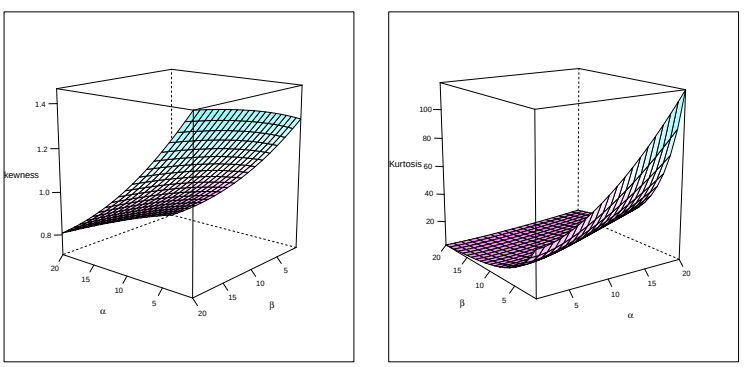


Figure 2: The skewness (left) and kurtosis (right) plots of OLLGL distribution for selected α, β for $\lambda = 2$.

Theorem 1 *If the baseline distribution $G(x)$ has a mgf, then $F(x)$ has also a mgf.*

The proof of this theorem was done by Gleaton and Lynch (2010). Since any moment of the Lindley distribution exists, all moments of the OLLGL distribution can be obtained.

2.5. Asymptotic properties

The asymptotic of cdf, pdf and hrf of the OLLGL distribution as $x \to 0$ are, respectively, given by

$$\begin{aligned} F(x) &\sim (\lambda x)^{\alpha\beta} \quad as \quad x \to 0, \\ f(x) &\sim \alpha\beta\,\lambda^{\alpha\beta}\,x^{\alpha\beta-1} \quad as \quad x \to 0, \\ h(x) &\sim \alpha\beta\,\lambda^{\alpha\beta}\,x^{\alpha\beta-1} \quad as \quad x \to 0. \end{aligned}$$

The asymptotic of cdf, pdf and hrf of the OLLGL distribution as $x \to \infty$ are, respectively, as follows

$$1-F(x) \sim (\frac{\alpha\lambda}{1+\lambda})^{\beta} x^{\beta} \mathrm{e}^{-\lambda\beta \mathrm{x}} \quad \text{as} \quad \mathrm{x} \to \infty,$$
$$f(x) \sim \beta\lambda (\frac{\alpha\lambda}{1+\lambda})^{\beta} x^{\beta} \mathrm{e}^{-\lambda\beta \mathrm{x}} \quad \text{as} \quad \mathrm{x} \to \infty,$$
$$h(x) \sim \beta\lambda \quad as \quad x \to \infty.$$

These equations show the effect of parameters on the tails of the OLLGL distribution.

2.6. Extreme Value

If $\bar{X} = (X_1 + ... + X_n)/n$ denotes the sample mean, then by the usual central limit theorem, $\sqrt{n}(\bar{X} - E(X))/\sqrt{\mathrm{Var}(X)}$ approaches the standard normal distribution as $n \to \infty$. One may be interested in the asymptotic of the extreme values $M_n = \max(X_1, ..., X_n)$ and $m_n = \min(X_1, ..., X_n)$. Let $\tau(x) = \frac{1}{\lambda}$, we obtain following equations for the cdf in (3) as

$$\lim_{t\to 0} \frac{F(t\,x)}{F(t)} = \lim_{t\to 0} \frac{G(t\,x)^{\alpha}}{G(t)^{\alpha}} = \lim_{t\to 0} \frac{\left[1-\left(1+\frac{\lambda\,t\,x}{1+\lambda}\right)\mathrm{e}^{-\lambda\,t\,x}\right]^{\alpha\beta}}{\left[1-\left(1+\frac{\lambda\,t}{1+\lambda}\right)\mathrm{e}^{-\lambda\,t}\right]^{\alpha\beta}} = \lim_{t\to 0} \frac{\left[1-\mathrm{e}^{-\lambda\,t\,x}\right]^{\alpha\beta}}{\left[1-\mathrm{e}^{-\lambda\,t}\right]^{\alpha\beta}}$$
$$= \lim_{t\to 0} \frac{(\lambda\,t\,x)^{\alpha\beta}}{(\lambda\,t)^{\alpha\beta}} = x^{\alpha\beta} \tag{21}$$

and

$$\lim_{t\to\infty} \frac{1-F(t+x\,\tau(t))}{1-F(t)} = \lim_{t\to\infty} (\frac{1-G(t+x\,\tau(t))^{\alpha}}{1-G(t)^{\alpha}})^{\beta} = \mathrm{e}^{-\beta\,x}. \tag{22}$$

Thus, from Leadbetter et al. (2012), there must be norming constants $a_n > 0, \quad b_n, \quad c_n > 0$ and d_n such that

$$Pr[a_n(M_n - b_n) \le x] \to \mathrm{e}^{-\mathrm{e}^{-\beta\,\mathrm{x}}}$$

and

$$Pr[c_n(m_n - d_n) \le x] \to 1 - \mathrm{e}^{-\mathrm{x}^{\alpha\beta}}$$

as $n \to \infty$. The form of the norming constants can also be determined. For instance, using Corollary 1.6.3 in Leadbetter et al. (2012), one can see that $b_n = F^{-1}(1-\frac{1}{n})$ and $a_n = \lambda$, where $F^{-1}(\cdot)$ denotes the inverse function of $F(\cdot)$.

2.7. Maximum likelihood estimation

We determine the maximum likelihood estimates (MLEs) of the parameters of the OLLGL distribution from complete samples. Let $x_1, \ldots, x_n$ be a random sample of size n from the OLLGL(α, β, λ) distribution. The log-likelihood function for the vector of parameters $\theta = (\alpha, \beta, \lambda)^T$ can be written

as

$$\begin{aligned} l(\theta) &= n\log\left(\frac{\alpha\beta\,\lambda^2}{1+\lambda}\right)+\sum_{i=1}^{n}\log(1+x_i)+(\alpha\beta-1)\sum_{i=1}^{n}\log(q_i) \\ &+ (\beta-1)\sum_{i=1}^{n}\log(1-q_i^{\alpha})-2\sum_{i=1}^{n}\log\left[q_i^{\alpha\beta}+(1-q_i^{\alpha})^{\beta}\right] \end{aligned} \tag{23}$$

where $q_i = 1-(1+\frac{\lambda}{1+\lambda}x_i)\mathrm{e}^{-\lambda \mathrm{x_i}}$ is a transformed observation.

The log-likelihood can be maximized either directly by using the SAS (Procedure NLMixed) or the MaxBFGS routine in the matrix programming language Ox (Doomik (2007)) or by solving the nonlinear likelihood equations obtained by differentiating (23). The components of the score vector $U(\theta)$ are given by

$$\begin{aligned} U_{\lambda}(\theta) &= \frac{2n}{\lambda}-\frac{n}{1+\lambda}-\sum_{i=1}^{n}x_i+(\alpha\beta-1)\sum_{i=1}^{n}\frac{q_i^{(\lambda)}}{q_i}+\alpha(1-\beta)\sum_{i=1}^{n}\frac{q_i^{(\lambda)}\,q_i^{\alpha-1}}{1-q_i^{\alpha}} \\ &\quad -2\alpha\beta\sum_{i=1}^{n}q_i^{(\lambda)}\,\frac{q_i^{\alpha\beta-1}-q_i^{\alpha-1}\left[1-q_i^{\alpha}\right]^{\beta-1}}{q_i^{\alpha\beta}+(1-q_i^{\alpha})^{\beta}}, \\ U_{\alpha}(\theta) &= \frac{n}{\alpha}+\beta\sum_{i=1}^{n}\log(q_i)+(1-\beta)\sum_{i=1}^{n}\frac{q_i^{\alpha}\log(q_i)}{1-q_i^{\alpha}} \\ &- 2\beta\sum_{i=1}^{n}\frac{q_i^{\alpha\beta}\log(q_i)-q_i^{\alpha}\left[1-q_i^{\alpha}\right]^{\beta-1}\log(q_i)}{q_i^{\alpha\beta}+(1-q_i^{\alpha})^{\beta}} \end{aligned}$$

and

$$U_{\beta}(\theta) = \frac{n}{\beta}+\alpha\sum_{i=1}^{n}\log(q_i)+\sum_{i=1}^{n}\log(1-q_i^{\alpha})-2\sum_{i=1}^{n}\frac{\alpha\,q_i^{\alpha\beta}\log(q_i)+\left[1-q_i^{\alpha}\right]^{\beta}\log\left[1-q_i^{\alpha}\right]}{q_i^{\alpha\beta}+(1-q_i^{\alpha})^{\beta}}.$$

For interval estimation and hypothesis tests on the model parameters, the 2×2 observed information matrix $J = J(\theta)$ is required.

Under conditions that are fulfilled for parameters in the interior of the parameter space but not on the boundary, the asymptotic distribution of $\sqrt{n}(\widehat{\theta}-\theta)$ is $N_3(0, I(\theta)^{-1})$, where $I(\theta)$ is the expected information matrix. In practice, we can replace $I(\theta)$ by the observed information matrix evaluated at $\widehat{\theta}$ (say $J(\widehat{\theta})$). We can construct approximate confidence intervals and confidence regions for the individual parameters and for the hazard and survival functions based on the multivariate normal $N_3(0, J(\widehat{\theta})^{-1})$ distribution.

Further, the likelihood ratio (LR) statistic can be used for comparing this distribution with some of its special sub-models. We can compute the maximum values of the unrestricted and restricted log-likelihoods to construct the LR statistics for testing some sub-models of the OLLGL distribution. For example, the test of $H_0 : \alpha = \beta = 1$ versus $H_1 : H_0$ *is not true* is equivalent to comparing the OLLGL and Lindley distributions and the LR statistic reduces to

$$w = 2\{\ell(\hat{\alpha},\hat{\beta},\hat{\lambda})-\ell(1,1,\tilde{\lambda})\},$$

where $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\lambda}$ are the MLEs under H and $\tilde{\lambda}$ is the estimate under H_0.

2.8. Least-Square Estimator

Let $x_{(1)}, x_{(2)}, x_{(3)}, \cdots, x_{(n)}$ denote the ordered sample of the random sample of size n from the OLLGL distribution function in (3). The least square estimators (LSEs) of the OLLGL distribution can be obtained by minimizing the following equation

$$\ell(\theta) = \sum_{i=1}^{n} \left\{ \frac{\left[1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x_{(i)}}\right]^{\alpha\beta}}{\left[1-(1+\frac{\lambda x_{(i)}}{1+\lambda})e^{-\lambda x_{(i)}}\right]^{\alpha\beta} + \left[1-\left[1-(1+\frac{\lambda x_{(i)}}{1+\lambda})e^{-\lambda x_{(i)}}\right]^{\alpha}\right]^{\beta}} - \frac{i}{n+1} \right\}^2. \quad (24)$$

The **optim** function of R software can be used to minimize the (24). The partial derivatives of (24) with respect to α, λ and β can be obtained from authors upon request.

3. Simulation

In this section, a simulation study on the model parameters is investigated. We consider MLE and LSE methods for estimating unknown parameters of the OLLGL distribution and compare the efficiency of parameters using these methods. The simulation procedure is as follows:

1. Set the sample size n and the vector of parameters $\theta = (\lambda, \alpha, \beta)$,
2. Generate random observations from the $OLLGL(\lambda, \alpha, \beta)$ distribution with size n,
3. Estimate $\hat{\theta}$ by means of MLE and LSE methods using the generated random observations in Step 2,
4. Repeat Steps 2 and 3 for N times,
5. Compute the mean relative estimates (MREs) and mean square errors (MSEs) using $\hat{\theta}$ and θ with the following equations:

$$MRE = \sum_{j=1}^{N} \frac{\hat{\theta}_{i,j}/\theta_i}{N},$$
$$MSE = \sum_{j=1}^{N} \frac{(\hat{\theta}_{i,j}-\theta_i)^2}{N}, i = 1,2,3.$$

where $\hat{\theta}_{i,j}$ for $i = 1,2,3$ and $j = 1,...,N$, is the estimation of ith element of parameter vector in jth iteration. The simulation results are obtained with software R. The chosen parameters of the simulation study are $\theta = (\lambda = 1.2, \alpha = 2, \beta = 0.2)$, $N = 1000$ and $n = (50, 55, 60, ..., 500)$. We expect that MREs are closer to one when the MSEs are near zero. Figures 3 represents estimated MSEs and MREs from MLE and LSE methods. Based on Figures 3, the MSE of all estimates tends to zero for large n and also as expected, the values of MREs tend to one. It is clear that the estimates of parameters are asymptotically unbiased. In estimation of β and λ, the MLE method approach to nominal values of the MSEs and MREs faster than the LSE method. The LSE method exhibits better performance than the MLE method for the large sample size in estimating α. Therefore, the MLE is a more suitable method than other for estimating parameters of the OLLGL distribution for small a sample size.

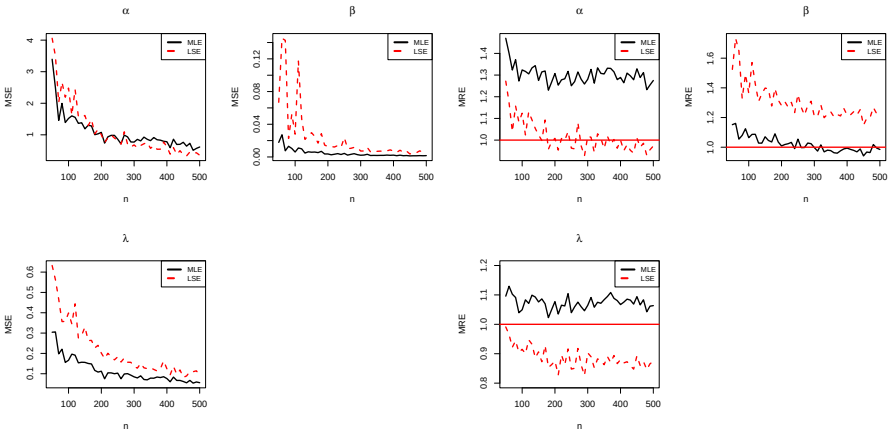


Figure 3: Estimated MREs and MSEs for the selected parameter values.

4. Applications

In this section, we illustrate the fitting performance of the OLLGL distribution using a real data set. For the purpose of comparison, we fitted the following models to show the fitting performance of the OLLGL distribution by means of real data set:

- Lindley Distribution, $L(\lambda)$.
- Power Lindley distribution, $PL(\beta,\lambda)$.
- Generalized Lindley, $GL(\alpha,\lambda)$, (Nadarajah et al. (2011)), with distribution function given by

$$F(x) = \left(1-(1+\frac{\lambda x}{1+\lambda})e^{-\lambda x}\right)^{\alpha}.$$

- Beta Lindley, $BL(\alpha,\beta,\lambda)$, Merovci and Sharma (2014), with distribution function given by

$$F(x) = \int_0^{L(x,\lambda)} t^{\alpha-1}(1-t)^{\beta-1}dt.$$

- Exponentiated power Lindley distribution, Ashour and Eltehiwy (2015), $EPL(\alpha,\beta,\lambda)$, with distribution function given by

$$F(x) = \left(1-(1+\frac{\lambda x^{\beta}}{1+\lambda})e^{-\lambda x^{\beta}}\right)^{\alpha}.$$

- Odd log-logistic Lindley distribution $OLL-L(\alpha,\lambda)$, (Ozel et al. (2017)), with distribution function given by

$$F(x) = \frac{L(x,\lambda)^{\alpha}}{L(x,\lambda)^{\alpha}+(1-L(x,\lambda))^{\alpha}}.$$

- Kumaraswamy Power Lindley, $KPL(\alpha,\beta,\gamma,\lambda)$ (Oluyede et al. (2016)

$$F(x) = 1-[1-PL(x,\beta,\lambda)^{\alpha}]^{\gamma}.$$

- Extended generalized Lindley, $EGL(\alpha,\gamma,\lambda)$, (Ranjbar et al. (2018)),

$$F(x) = \frac{L(x,\lambda)^{\alpha}}{L(x,\lambda)^{\alpha} + 1 - (1 - L(x,\lambda))^{\gamma}}.$$

- New Odd-log logistic Lindley, $NOLLL(\alpha,\beta,\lambda)$, Alizadeh et al. (2018b)

$$F(x) = \frac{L(x,\lambda)^{\alpha}}{L(x,\lambda)\alpha + (1 - L(x,\lambda))^{\beta}}.$$

Estimates of the parameters of OLLGL distribution, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Cramer Von Mises and Anderson-Darling statistics (W^* and A^*) are presented for each data set. We have also considered the Kolmogorov-Smirnov (K-S) statistic and its corresponding p-value and the minimum value of the minus log-likelihood function (-Log(L)) for the sake of comparison. Generally speaking, the smaller values of AIC, BIC, W^* and A^*, the better fit to a data set. Furthermore, the likelihood ratio (LR) tests apply for evaluating the OLLGL distribution with its sub-models. For example, the test of $H_0 : \beta = 1$ against $H_1 : \beta \neq 1$ is equivalent to comparing the OLLGL with GL and the LR test statistic is given by

$$LR = 2\left[l(\hat{\alpha},\hat{\beta},\hat{\lambda}) - l(\hat{\alpha^*},1,\hat{\lambda^*})\right],$$

where $\hat{\alpha^*}$ and $\hat{\lambda^*}$ are the ML estimators under H_0 of α and λ, respectively. All the computations were carried out using the software R.

The data set is given from Murthy (2004). The ML estimates of the parameters and the goodness-of-fit test statistics for the first data set is presented in Table 3 and 4 respectively. As we can see, the smallest values of AIC, BIC, A^*, W^* and $-l$ statistics and the largest p-values belong to the OLLGL distribution. Therefore, the OLLGL distribution outperforms the other competitive considered distribution in the sense of this criteria.

Table 1: Data set.

0.032	0.035	0.104	0.169	0.196	0.260	0.326	0.445	0.449	0.496
0.543	0.544	0.577	0.648	0.666	0.742	0.757	0.808	0.857	0.858
0.882	1.138	1.163	1.256	1.283	1.484	1.897	1.944	2.201	2.365
2.531	2.994	3.118	3.424	4.097	4.100	4.744	5.346	5.479	5.716
5.825	5.847	6.084	6.127	7.241	7.560	8.901	9.000	10.482	11.133

In addition, the profile log-likelihood functions of the OLLGL distribution are plotted in Figure 4. These plots reveal that the likelihood equations of the OLLGL distribution have solutions that are maximizers.

The values of LR test statistics and their corresponding p-values are exhibited in Table 5. From Table 5, we observe that the computed p-values are too small so we reject all the null hypotheses and conclude that the OLLGL fits the data set better than the considered sub-models according to the LR criterion.

We also plotted the fitted pdfs and TTT plots of the considered models for the sake of visual comparison, in figures 5 and 6, respectively. Therefore, the OLLGL distribution can be considered as an appropriate model for fitting the data set.

5. Bayesian estimation

The Bayesian inference procedure has been taken into consideration by many statistical researchers, especially researchers in the field of survival analysis and reliability engineering. In this section, the complete sample data are analysed through a Bayesian point of view. We assume that the parameters α, β and λ of the *OLLGL* distribution have independent prior distributions as

$$\alpha \sim Gamma(a,b), \lambda \sim Gamma(e,f), \beta \sim Gamma(g,h)$$

where a, b, e, f, g and h are positive. Hence, the joint prior density function is formulated as follows:

$$\pi(\alpha,\beta,\lambda) = \frac{b^a f^e h^g}{\Gamma(a)\Gamma(e)\Gamma(g)} \alpha^{a-1}\beta^{h-1}\lambda^{e-1}e^{-(b\alpha+h\beta+f\lambda)}. \quad (25)$$

In the Bayesian estimation, we do not know the actual value of the parameter, which may be adversely affected by loss when we choose an estimator. This loss can be measured by a function of the parameter and corresponding estimator. For the Bayesian discussion, we consider different types of symmetric and asymmetric loss functions such as squared error loss function (*SELF*), weighted squared error loss function (*WSELF*), modified squared error loss function (*MSELF*), precautionary loss function (*PLF*) and *K*-loss function (*KLF*). These loss functions, associated Bayesian estimators and posterior risks are presented in Table 2. For more details see Calabria and Pulcini (1996). Next,

Table 2: Bayes estimator and posterior risk under different loss functions

Loss function	Bayes estimator	Posterior risk
$SELF = (\theta - d)^2$	$E(\theta\|x)$	$Var(\theta\|x)$
$WSELF = \frac{(\theta-d)^2}{\theta}$	$(E(\theta^{-1}\|x))^{-1}$	$E(\theta\|x) - (E(\theta^{-1}\|x))^{-1}$
$MSELF = \left(1 - \frac{d}{\theta}\right)^2$	$\frac{E(\theta^{-1}\|x)}{E(\theta^{-2}\|x)}$	$1 - \frac{E(\theta^{-1}\|x)^2}{E(\theta^{-2}\|x)}$
$PLF = \frac{(\theta-d)^2}{d}$	$\sqrt{E(\theta^2\|x)}$	$2\left(\sqrt{E(\theta^2\|x)} - E(\theta\|x)\right)$
$KLF = \left(\sqrt{\frac{d}{\theta}} - \sqrt{\frac{\theta}{d}}\right)$	$\sqrt{\frac{E(\theta\|x)}{E(\theta^{-1}\|x)}}$	$2\left(\sqrt{E(\theta\|x)E(\theta^{-1}\|x)} - 1\right)$

we provide the posterior probability distributions for a complete data set. Let us define the function φ as

$$\varphi(\alpha,\beta,\lambda) = \alpha^{a-1}\beta^{h-1}\lambda^{e-1}e^{-(b\alpha+h\beta+f\lambda)},\ \alpha > 0,\ \beta > 0,\ \lambda > 0.$$

The joint posterior distribution in terms of a given likelihood function $L(data)$ and joint prior distribution $\pi(\alpha,\beta,\lambda)$ is defined as

$$\pi^*(\alpha,\beta,\lambda|data) \propto \pi(\alpha,\beta,\lambda)L(data). \quad (26)$$

Hence, we get joint posterior density of parameters α, β and λ for complete sample data by combining the likelihood function and joint prior density (25). Therefore, the joint posterior density function is given by

$$\pi^*(\alpha,\beta,\lambda|\underline{x}) = K\varphi(\alpha,\beta,\lambda)L(\underline{x},\xi) \quad (27)$$

where

$$L(\underline{x};\xi)=\prod_{i=1}^{n}\frac{\alpha\beta\lambda^2(1+x)\mathrm{e}^{-\lambda x}\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda x}\right]^{\alpha\beta-1}\left[1-\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda x}\right]^{\alpha}\right]^{\beta-1}}{(1+\lambda)\left\{\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda x}\right]^{\alpha\beta}+\left[1-\left[1-(1+\frac{\lambda x}{1+\lambda})\mathrm{e}^{-\lambda x}\right]^{\alpha}\right]^{\beta}\right\}^2}. \tag{28}$$

and K is given as

$$K^{-1} \quad = \quad \int_0^\infty\int_0^\infty\int_0^\infty \varphi(\alpha,\beta,\lambda)L(\underline{x},\xi)d\alpha d\beta d\lambda.$$

Moreover, the marginal posterior pdf of α, γ and β, assuming that $\Theta=(\alpha,\gamma,\beta)$, can be given

$$\pi(\Theta_i|\underline{x})=\int_0^\infty\int_0^\infty \pi^*(\Theta|\underline{x})\Theta_j\Theta_k, \tag{29}$$

where $i,j,k=1,2,3,i\neq j\neq k$ and also Θ_i is ith member of a vector Θ. It is clear from the equations (27) and (29) that there are no closed-form expressions for the Bayesian estimators under the five loss functions described in Table 2. Because of intractable integrals associated with joint posterior and marginal posterior distributions, we need to use numerical software to solve integral equations numerically via MCMC method. The two most popular MCMC methods are: the Metropolis-Hastings algorithm (Metropolis et al. (1953), Hastings (1970)) and the Gibbs sampling (Geman and Geman (1984)). The Gibbs sampling is a special case of the Metropolis-Hastings algorithm which generates a Markov chain by sampling from the full set of conditional distributions. The Gibbs sampling algorithm can be described generically as follows.

Suppose that the general model $f(\underline{x}|\theta)$ is associated with parameter vector $\theta=(\theta_1,\theta_2,...,\theta_p)$ and observed data $\underline{x}$. Thus, the joint posterior distribution is $\pi(\theta_1,\theta_2,...,\theta_p|\underline{x})$. We also assume that $\theta_0=(\theta_1^{(0)},\theta_2^{(0)},...,\theta_p^{(0)})$ is the initial values vector to start the Gibbs sampler. The Gibbs sampler draws the values for each iteration in p steps by drawing a new value for each parameter from its full conditional given the most recently drawn values of all other parameters. In symbols, the steps for any iteration, say iteration k, are as follows:

- starting with an initial estimate $(\theta_1^{(0)},\theta_2^{(0)},...,\theta_p^{(0)})$
- Draw θ_1^k from $\pi\big(\theta_1|\theta_2^{k-1},\theta_3^{k-1},...,\theta_p^{k-1}\big)$
- Draw θ_2^k from $\pi\big(\theta_2|\theta_1^{k},\theta_3^{k-1},...,\theta_p^{k-1}\big)$; and so on down to
- Draw θ_p^k from $\pi\big(\theta_p|\theta_1^{k},\theta_2^{k},...,\theta_{p-1}^{k}\big)$

As mentioned above, often Bayesian inference requires computing intractable integrals to generate posterior samples. Using Gibbs sampling, one can obtain samples from the joint posterior distribution. In practice, simulations related to Gibbs sampling are conducted through a special software WinBUGS. WinBUGS software was developed in 1997 to simulate data of complex posterior distributions, where analytical or numerical integration techniques cannot be applied. Also, we can use OpenBUGS software, which is an open-source version of WinBUGS. Since there is not any prior information about hyper parameters in (25), one can implement the idea of Congdon (2001) and these

parameters can be chosen as $a = b = c = d = e = f = 0.0001$. Hence, we can use the *MCMC* procedure to extract posterior samples of (27) by means of the Gibbs sampling process in OpenBUGS software.

Bayesian estimators associated with the parameters of the *OLLGL* distribution are computed based on the single chain of 10000 cycles of the Gibbs sampler with a conservative burn-in period of the first 1000 iterations. The corresponding Bayesian point and interval estimation and posterior risk are provided in Tables 6 and 7 for the data set. Table 7 provides 95% credible and *HPD* intervals for each parameter of the *OLLGL* distribution. The convergence of the Gibbs sampler process is verified through graphical inspection (Trace, Autocorrelation and Histogram plots) of the posterior sampled values. It is observed that the Gibbs samples of all the parameter estimates achieved a good stationary phase for both considered data sets. We provide the posterior summary plots in Figures 7, 8 and 9. These plots confirm that the convergence of the Gibbs sampling process is occurred.

6. Conclusion

In this paper, a new distribution which is called odd log-logistic generalized-Lindley (OLLGL) distribution was introduced. The statistical properties of the OLLGL distribution including the hazard function, quantile function, moments, incomplete moments, generating functions, mean deviations and maximum likelihood estimation for the model parameters were given. Simulation studies were conducted to examine the performance of this distribution. We also presented applications of this new distribution for a real-life data set in order to illustrate the usefulness of the distribution. Finally, the Bayesian estimation and the Gibbs sampling procedure for the considered data sets were discussed.

References

Abouammoh, A., Alshangiti, A. M., and Ragab, I., (2015). A new generalized lindley distribution. *Journal of Statistical Computation and Simulation*, 85(18), pp. 3662–3678.

Afify, A. Z., Cordeiro, G. M., Maed, M. E., Alizadeh, M., Al-Mofleh, H., and Nofal, Z. M., (2019). The generalized odd lindley-g family: properties and applications. *Anais da Academia Brasileira de Ci^encias*, 91(3).

Alizadeh, M., Afify, A. Z., Eliwa, M., and Ali, S., (2019). The odd log- logistic lindley-g family of distributions: properties, bayesian and non-bayesian estimation with applications. *Computational Statistics*, 35, pp. 281–308.

Alizadeh, M., Afshari, M., Hosseini, B., , and Ramires, T. G., (2020). Extended exp-g family of distributions: Properties and applications. *Communication in Statistics-Simulation and Computation*, 49 (7), pp. 1730–1745.

Alizadeh, M., Altun, E., Ozel, G., Afshari, M., and Eftekharian, A., (2018b). A new odd log-logistic lindley distribution with properties and applications. *Sankhya*, 81(2), pp. 323–346.

Alizadeh, M., K MirMostafaee, S., Altun, E., Ozel, G., and Khan Ah- madi, M., (2017). The odd log-logistic marshall-olkin power lindley distribution: Properties and applications. *Journal of Statistics and Management Systems*, 20(6), pp. 1065–1093.

Alizadeh, M., Ozel, G., Altun, E., Abdi, M., and Hamedani, G., (2017). The odd log-logistic marshall-olkin lindley model for lifetime data. *Journal of Statistical Theory and Applications*, 16(3), pp. 382–400.

Alzaatreh, A., Lee, C., and Famoye, F., (2013). A new method for generating families of continuous distributions. *Metron*, 71(1), pp. 63–79.

Asgharzadeh, A., Bakouch, H. S., Nadarajah, S., Sharafi, F., et al., (2016). A new weighted lindley distribution with application. *Brazilian Journal of Probability and Statistics*, 30(1), pp. 1–27.

Asgharzadeh, A., Nadarajah, S., and Sharafi, F., (2018). Weibull lindley distribution. *REVSTAT Statistical Journal*, 16(1), pp. 87–113.

Ashour, S. K. and Eltehiwy, M. A., (2015). Exponentiated power lindley distribution. *Journal of advanced research*, 6(6), pp. 895–905.

Calabria, R. and Pulcini, G., (1996). Point estimation under asymmetric loss functions for left-truncated exponential samples. *Communications in Statistics Theory and Methods*, 25(3), pp. 585–600.

Congdon, P., (2001). Bayesian statistical analysis. *Wiley, New York.*

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E., (1996). On the lambertw function. *Advances in Computational mathematics*, 5(1), pp. 329–359.

De Haan, L., Ferreira, A., and Ferreira, A., (2006). Extreme value theory: an introduction, *Springer*, volume 21.

Doomik, J., (2007). Object-Oriented Matrix Programming Using OX. *International Thomson Business Press and Oxford, London*.

Geman, S. and Geman, D., (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*,6, pp. 721–741.

Ghitany, M., Atieh, B., and Nadarajah, S., (2008). Lindley distribution and its application. *Mathematics and computers in simulation*, 78(4), pp. 493–506.

Gleaton, J. U. and Lynch, J. D., (2010). Extended generalized log- logistic families of lifetime distributions with an application. *J. Probab. Stat. Sci*, 8, pp. 1–17.

Gomes-Silva, F. S., Percontini, A., de Brito, E., Ramos, M. W., Ven^ancio, R., and Cordeiro, G. M., (2017). The odd lindley-g family of distributions. *Austrian Journal of Statistics*, 46(1), pp. 65–87.

Gradshteyn, I. and Ryzhik, I., (2007). Table of Integrals, Series, and Products. *Elsevier/Academic Press*.

Hastings,W. K., (1970). Monte Carlo sampling methods using Markov chains and their applications. *Oxford University Press*.

Jodr´ a, P., (2010). Computer generation of random variables with lindley or poisson–lindley distribution via the lambert w function. *Mathematics and Computers in Simulation*, 81(4), pp. 851–859.

Leadbetter, M. R., Lindgren, G., and Rootz´en, H., (2012). Extremes and related properties of random sequences and processes. *Springer Science and Business Media*.

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

Lo, G. S., Ngom, M., Kpanzou, T. A., and Diallo, M., (2018). Weak convergence (iia)-functional and random aspects of the univariate extreme value theory. *arXiv preprint arXiv:1810.01625*.

Merovci, F. and Sharma, V. K., (2014). The beta-lindley distribution: properties and applications. *Journal of Applied Mathematics*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6), pp. 1087–1092.

Murthy, D. P., Xie, M., and Jiang, R., (2004). *Weibull models, volume 505. John Wiley and Sons*.

Nadarajah, S., Bakouch, H. S., and Tahmasbi, R., (2011). A generalized lindley distribution. *Sankhya B*, 73(2), pp.331–359.

Oluyede, B. O., Yang, T., and Makubate, B., (2016). A new class of generalized power lindley distribution with applications to lifetime data. *Asian Journal of Mathematics and Applications*, 2016, pp. 1–34.

Ozel, G., Alizadeh, M., Cakmakyapan, S., Hamedani, G., Ortega, E. M., and Cancho, V. G., (2017). The odd log-logistic lindley poisson model for lifetime data. *Communications in Statistics-Simulation and Computation*, 46(8), pp. 6513–6537.

Ranjbar, V., Alizadeh, M., and Altun, E., (2018). Extended generalized lindley distribution: Properties and applications. *Journal of Mathematical Extension*, 13(1), pp. 117–142.

Zakerzadeh, H. and Dolati, A., (2009). Generalized lindley distribution. *Journal of Mathematical Extension*, 3(2), pp.1–17.

Appendix

Tables and figures of the real data analyses section:

Tables:

Table 3: Parameter ML estimates and theirs standard errors (in parentheses).

Model	α	β	γ	λ
Lindley(λ)	–	–	–	0.5656(0.0585)
GL(α,λ)	0.6223(0.1142)	–	–	0.4351(0.0709)
PL(β,λ)	0.7593(0.0792)	–	–	0.7701(0.1088)
BL(α,β,λ)	0.6605(0.1407)	0.4098(0.5014)	–	0.9475(1.0566)
EPL(α,β,λ)	0.6825(0.2692)	1.2376(0.9557)	–	0.9372(0.6396)
OLLL(α,λ)	0.7099(0.0894)	–	–	0.6317(0.0853)
KPL$(\alpha,\beta,\gamma,\lambda)$	0.7799(0.1003)	1.5335(0.5297)	0.1262(0.0368)	4.3580(1.0558)
EGL(α,γ,λ)	0.6192(0.1068)	0.4135(0.4174)	–	0.3805(0.1489)
NOLLL(α,β,λ)	0.2513(0.1063)	1.4241(0.5008)	–	1.2655(0.3774)
OLLGL(α,β,λ)	0.2575(0.0972)	1.5854(0.5016)	–	5.4620(3.3951)

Table 4: Goodness-of-fit test statistics.

Model	*AIC*	*BIC*	$p-value$	W^*	A^*	$-l$
Lindley(λ)	215.8801	217.7921	0.0128	0.1358	0.7415	106.9412
GL(α,λ)	210.5744	214.3985	0.3338	0.1393	0.7576	103.2872
PL(β,λ)	209.6294	213.4534	0.5108	0.1085	0.6061	102.8147
BL(α,β,λ)	212.1457	217.8818	0.3376	0.1318	0.7167	103.0729
EPL(α,β,λ)	211.5485	217.2846	0.5544	0.0992	0.5667	102.7742
OLLL(α,λ)	209.0254	212.8494	0.4245	0.1212	0.6521	102.5127
KPL$(\alpha,\beta,\gamma,\lambda)$	212.9133	220.5614	0.5858	0.0809	0.4850	215.8257
EGL(α,γ,λ)	212.4044	218.1405	0.4438	0.1281	0.7047	103.2022
NOLLL(α,β,λ)	206.9584	212.6945	0.8271	0.0369	0.2691	100.4792
OLLGL(α,β,λ)	**206.5137**	**212.2498**	**0.8984**	**0.0362**	**0.2569**	**100.2569**

Table 5: The LR test results.

	Hypotheses	LR	p-value
OLLGL versus Lindley	$H_0:\alpha=\beta=1$	13.3663	0.00125
OLLGL versus OLL-L	$H_0:\beta=1$	4.5116	0.03366
OLLGL versus GL	$H_0:\alpha=1$	6.0607	0.01382

Table 6: Bayesian estimates $\widehat{\theta}$ and their posterior risks $r_{\widehat{\theta}}$ of the parameters under different loss functions.

Data	First data set		
Bayesian estimation			
Loss function	$\widehat{\alpha}\ (r_{\widehat{\alpha}})$	$\widehat{\beta}\ (r_{\widehat{\beta}})$	$\widehat{\lambda}\ (r_{\widehat{\lambda}})$
SELF	274.818 (10.648)	0.3393 (0.0021)	6.4192 (0.1030)
WSELF	274.336 (0.4825)	0.3332 (0.0061)	6.4032 (0.0160)
MSELF	273.853 (0.0018)	0.3270 (0.0185)	6.3872 (0.0025)
PLF	275.059 (0.4824)	0.3423 (0.0060)	6.4272 (0.0160)
KLF	274.577 (0.0018)	0.3362 (0.0183)	6.4112 (0.0025)

Table 7: Credible and *HPD* intervals of the parameters α, β and λ.

	Credible interval	HPD interval
α	(266.8, 282.9)	(252.8, 296.9)
β	(0.3081, 0.3687)	(0.2466, 0.4236)
λ	(6.196, 6.638)	(5.819, 7.066)

Figures:

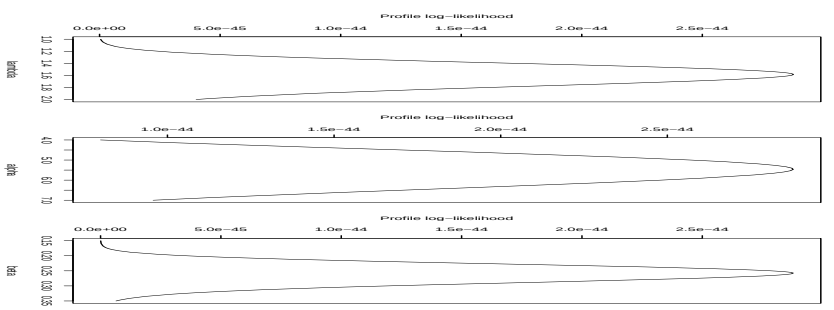


Figure 4: The profile log-likelihood functions of the OLLGL distribution.

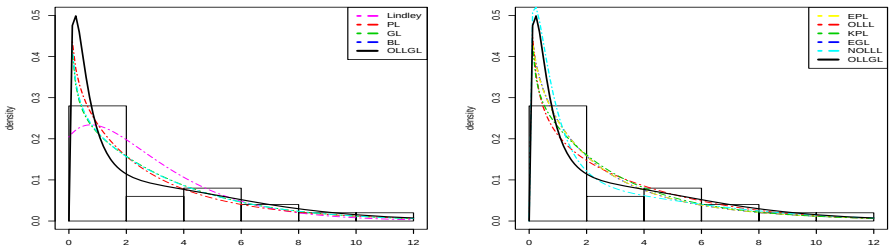


Figure 5: Fitted densities of distributions.

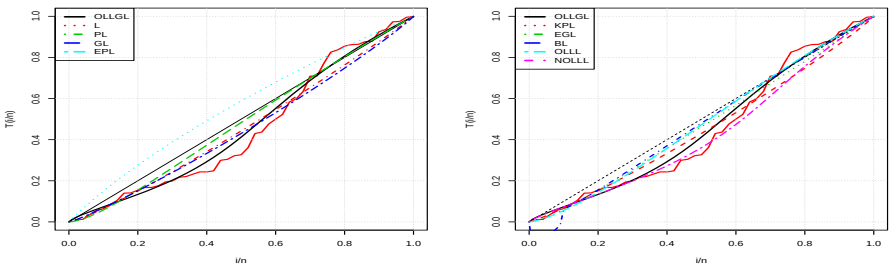


Figure 6: TTT plots of distributions.

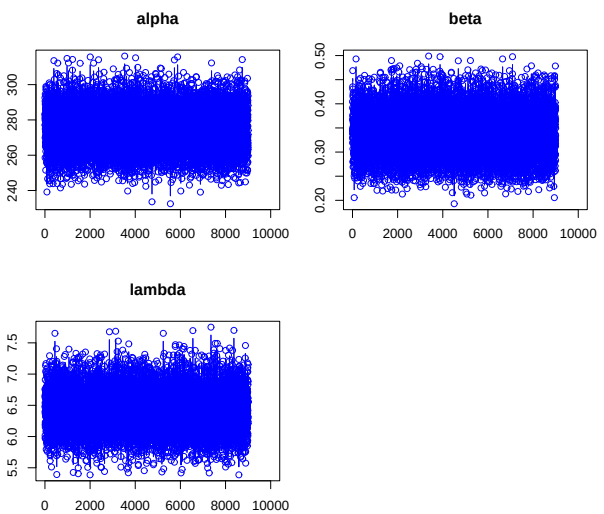


Figure 7: Plots of Bayesian analysis and performance of Gibbs sampling. Trace plots of each parameter of *OLLGL* distribution.

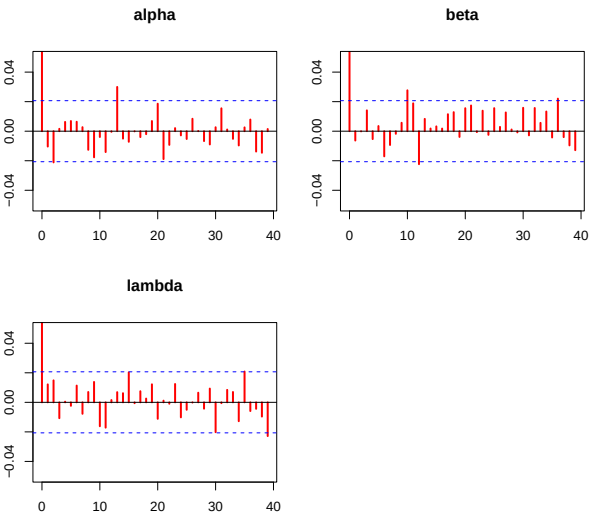


Figure 8: Plots of Bayesian analysis and performance of Gibbs sampling. Autocorrelation plots of each parameter of *OLLGL* distribution.

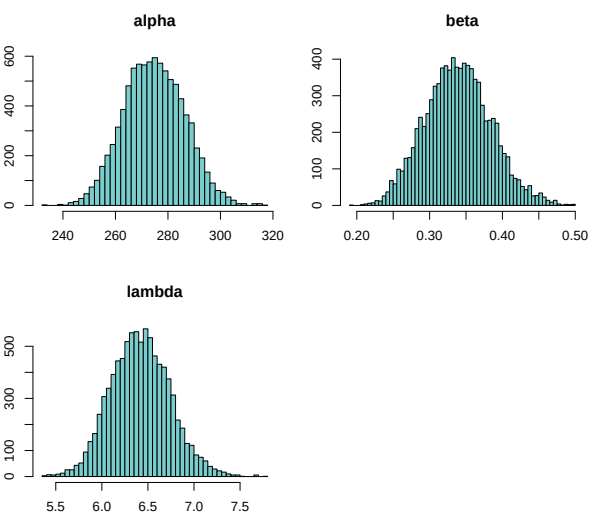


Figure 9: Plots of Bayesian analysis and performance of Gibbs sampling. Histogram plots of each parameter of *OLLGL* distribution.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 93–107, https://doi.org/10.59170/stattrans-2023-053
Received – 06.11.2021; accepted – 21.11.2022

Hyper-parametric Generalized Autoregressive Scores (GASs): an application to the price of United States cooking gas

Rasaki Olawale Olanrewaju[1], **Sodiq Adejare Olanrewaju**[2], **Omodolapo Waliyat Isamot**[3]

Abstract

This paper presents the framework of the Generalized Autoregressive Score (GAS) model with a variety of symmetric conditional densities of different time-varying hyper-parameters. The distinctive trait and goal of the observation-driven GAS model is to use its score and information functions as the compeller of time-variation via hyper-parameters of conditional densities. 10 robust hyper-parametric conditional densities were used as random error drivers for the GAS model with an application to the price of the United States cooking gas in the period between 2005 and 2020. Out of the 10 robust hyper-parametric conditional noises for the GAS model, the Asymmetric Student–t with one tail decay parameter (AST1) outperformed other categories of its variants and other conditional densities subjected to the GAS model, achieving a minimum reduced-error performance of AIC=11943.277 and BIC=12014.525. The hyper-parametric model obtained a location score and scale score of -1.2634 and 0.6636, respectively, while its location information and scale information was 0.2691 and 0.0362, respectively. Furthermore, the GAS model via AST1 proved more efficient than the core volatile conditional heteroscedasticity model of the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) at GARCH (1,1) via a Gaussian distributed noise.

Key words: Asymmetric Student–t, Generalized Autoregressive Score, hyper-parameters, score, information.

1. Introduction

Describing and estimating time-varying variation in stochastic time series has been the process of aperture across all fields of applied statistics and most scientific

[1] Africa Business School (ABS), Mohammed VI Polytechnic University (UM6P), Rabat, Morocco. E-mail: olanrewaju_rasaq@yahoo.com. ORCID: https://orcid.org/0000-0002-2575-9254.

[2] Department of Statistics, University of Ibadan, Ibadan, Oyo State, Nigeria.
E-mail: sodiqadejare19@gmail.com. ORCID: . https://orcid.org/0009-0006-4494-2421.

[3] Department of Epidemiology and Medical Statistics, University of Ibadan, Ibadan, Oyo State, Nigeria. E- mail: omodolapo.isamot@gmail.com.

investigations. Time-varying variation is cognate in modelling parameter selection for strategizing and capturing behavioural dynamics of either multivariate or univariate stochastic time series process with different myriad of possible specifications (Cox, 1981; Creal *et al*., 2013). According to Harvey & Luati (2014), some time-varying parameters of some proposed time series models are not only difficult to estimate (especially the class of stochastic volatility models reviewed by Olanrewaju *et al.* (2020) & Shephard (2005)), but also at times fail to take into consideration the shape of the conditional distribution of the data. These time-varying models in time series are categorized in two classes: parameter-driven models and observation-driven models. In the latter, the time variations of the parameters are used by subjecting the stochastic parameters to be functions of lagged dependent variables as well as synchronous and lagged exogenous variables. A typical example of such a model is the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) by Engle (1982). In the former, parameters are stochastic processes with their associated source of error, such that given the past and concurrent information the parameters cannot be perfectly predictable. Examples of such models are Stochastic Volatility (SV) model by Shephard (2005) and stochastic intensity models by Koopman *et al*. (2008).

In order to strengthen the observation-driven based model, Creal *et al*. (2013) and Harvey (2013) proposed the score function of conditional density functions as the compeller of time-variation in the time series parameters to describe the data. This resulted into score-driven model called Generalized Autoregressive Score (GAS) model, otherwise known as Dynamic Conditional Score (DCS) model. Among the merits of GAS model over other observation-driven models are: it is based on complete density function rather than moments, its likelihood evaluation is free from ambiguity, its driven mechanism is based on score and information functions (Hessian). The model can be extended to long memory, asymmetric and other intricate dynamics. It is flexible enough to be used in all fields in which the use of time-varying parameter models is relevant. It can be subjected to real-value, integer-valued, strictly positive or (0, 1)-bounded observations provided the conditional density (either probability density function or probability mass function) for the score function and Hessian exists and is well-defined (Oh and Patton, 2016). It provides framework for modelling time variation in parametric models when computing the score of a parametric conditional observational density with respect to time-varying parameter. The practical relevance of GAS model includes default and credit risk modeling as affirmed by Lucas & Zhang (2016); stock volatility and correlation modeling as declared by Harvey & Sucarrat (2014); modeling time-varying dependence structures as established by Harvey & Thiele (2016); CDS spread modeling and questions relating to financial stability and systemic risk, modeling high frequency data as confirmed by Janus *et al*. (2014) and spatial econometrics as affirmed by Blasques *et al*. (2016).

The novelty of this article is to extend the driving mechanism of the score function and Hessian of the GAS model via its random noise to some probability density functions like Normal and its variants, Asymmetric Student–t with two tail decay parameters, Asymmetric Student–t with one tail decay parameter, Student-t, location-scale skew-normal distribution, Skew-t distribution, Asymmetric Laplace, Gamma, and exponential. The notion of the mentioned conditional densities to GAS model is to be able to improve its score function and Hessian robustly via each conditional density time-varying hyper-parameters like location, scale, skewness, and shapes indexes. The high frequency financial data to be subjected to the GAS model via the mentioned conditional densities is the price of United State cooking gas. The raw dataset of the price of the United State cooking gas from 2005-2020 will be used as extracted from U.S. Energy Information Administration (EIA).

2. Model Specification

In this section, the general class of observation-driven time-varying parameter model will be formulated. Thereafter, the Generalized Autoregressive Score (GAS) for the time-varying hyper-parameters driven by scale function of conditional likelihood will be formulated to drive the score-function and Hessian. According to Monache and Petrella (2014), time-varying parametric autoregressive model of order "i" can be defined as:

$$x_t = \phi_{0,t} + \phi_{1,t}x_{t-1} + \phi_{2,t}x_{t-2} + \cdots + \phi_{i,t}x_{t-i} + \omega_t \tag{1}$$

where the error term is $\omega_t \sim (0, \sigma_t^2) t = 1,2,\cdots,n$; $\phi_0, \cdots, \phi_i$ are the parameters of the autoregressive model; $x_{t-1}, \cdots, x_{t-i}$ are the past series values (lags).

Olanrewaju & Folorunsho (2018) proposed an updating rule by defining the associated variation of the time-varying hyper-parameters in a vector form to be: $g_t = (\phi_t', \sigma_t^2) \ni \phi_t' = \left(\phi_{0,t}, \phi_{1,t}, \cdots, \phi_{i,t}\right)'$. This implies that equation (1) can be interpreted as the first order of a Markov process with

$$g_{t+1} = \eta + Kg_t + \xi_t, \xi_t \sim (0, \Sigma_t) \tag{2}$$

where η is a vector of constants; K and Σ are the matrices of hyper-parameters (updated location and scale parameters respectively), g_t connotes the time-varying parameters. The Generalized Autoregressive Score (GAS) for the time-varying hyper-parameters driven by scale function of conditional likelihood of g_t given the immediate past of "$t-1$", $g_{t-1} = (\phi_{t-1}', \sigma_{t-1}^2)$

$$g_{t+1/t} = \eta + Kg_{t/t+1} + Zc_t \tag{3}$$

where, $X_{t-1} = \{x_{t-1}, x_{t-2}, \cdots, x_1\}$, $\eta$and K are the same as defined above, where $Zc_t \sim (0, \sigma_t^2) t = 1,2, \cdots, n$ is the error term of the GAS time-varying hyper-parameters with driven mechanism called the score-function.

$$c_t = C_t \nabla_t \tag{4}$$

$$\ni \quad \nabla_t = \frac{\partial[log\, p(x_t/(Z_t;\theta_t)]}{\partial g_{t/t-1}}; \quad C_t = I_{t-1}^{-1} = \left[\frac{\partial[log\, p(x_t/(Z_t;\theta_t)]}{\partial g_{t/t-1} g'_{t/t-1}}\right]^{-1}$$

with I_{t-1}^{-1} being the Information matrix (Hessian), $Z_t = [G_t, X_{t-1}]$ and $G_t = \{g_{t/t-1}, g_{t-1/t-2}, \cdots, g_{1/0}\}$ defined for vector parameters of θ_t; $p(x_t/Z_t;\theta)$ is probability of the past series values (lags) at time "t" given that the error (Z_t) and vector parameters (θ_t) at time. Rewriting equation (1) in matrix form gives

$$x_t = A'\phi_{t/t-1} + \omega_t \ni \omega_t/X_{t-1} \sim (0, g_{t/t-1}), for t = 1, \cdots, n, \tag{5}$$

$\sigma_{t/t-1}^2 = g_{t/t-1}, A' = [1, x_{t-1}, \cdots, x_{t-p}] \& \phi_{t/t-1} = [\phi_{0,t/t-1}, \phi_{1,t/t-1}, \cdots, \phi_{p,t/t-1}]'$
$\omega_t = x_t - A'\phi_{t/t-1},$

$\omega_t = x_t - A'\phi_{t/t-1}, \mu_t = A'\phi_{t/t-1}$

The matrix form of equation (5) will be incorporated into:

Student-t-Distribution as

$$p(x_t;\theta_t) = \frac{\Gamma\left(\frac{v_t+1}{2}\right)}{\Gamma\left(\frac{v_t}{2}\right) g_t \sqrt{\pi v_t}} \left(1 + \frac{(2\phi'_{t/t-1} A + \omega'_t)}{v_t g_{t/t-1}^2}\right)^{-\frac{v_t+1}{2}} - \infty < x_t < +\infty \tag{6}$$

$\omega_t/X_{t-1} \sim NID(0, g_{t/t-1}, v_t)$ for location parameter μ_t, scale parameter g_t, v_t degree of freedom $\theta_t = \{\mu_t, \phi, g_t, v_t\}'$. According to Jones and Faddy (2003), **Asymmetric Student–t with two tail decay parameters** (that is the Student t-distribution, which is both heavy tailed and skew). Then, the density function of this new distribution is

$$p(x_t;a,b) = C_{a,b}^{-1} \left\{1 + \frac{x_t}{(a+b+x_t^2)^{\frac{1}{2}}}\right\}^{a+\frac{1}{2}} \left\{1 - \frac{x_t}{(a+b+x_t^2)^{\frac{1}{2}}}\right\}^{b+\frac{1}{2}} \tag{7}$$

where,

$C_{a,b} = 2^{a+b-1} B(a,b)(a+b)^{\frac{1}{2}}$, where $B(a,b)$ denotes the beta function. When a = b, $p(x_t;b,a)$ reduces to the Student- t-distribution on (2a) degrees of freedom (**Asymmetric Student–t with one tail decay parameter**). When a<b or a>b, $p(x_t;b,a)$ is negatively or positively skewed respectively. In fact, $p(x_t;b,a) = p(-x_t;b,a)$. Note that "a" and "b" are positive real numbers and need not to be integer or half-integer.

Location-Scale Skew-Normal distribution

According to Owen (2008), a random variable X is said to be a location-scale skew-normal distribution, with location at μ, scale at δ and shape parameter α, and denoted $X \sim \theta = SN(\mu, \delta^2, \alpha)$ if its probability density function (pdf) is given by

$$p(x_t\,;\theta_t) = \frac{2}{\delta}\phi\left(\frac{x_t-\mu}{\delta}\right)\Phi\left(\alpha\frac{x_t-\mu}{\delta}\right), x_t \in \mathbb{R}(\alpha, \mu \in \mathbb{R}, \delta \in \mathbb{R}^+),$$

$$\text{Then,} \qquad p(x_t\,;\theta_t) = \frac{2}{\delta}\phi\left(\frac{x_t-A'\phi_{t/t-1}}{\delta}\right)\Phi\left(\alpha\frac{x_t-A'\phi_{t/t-1}}{\delta}\right) \tag{8}$$

$$\ni \omega_t/X_{t-1} \sim N\left(0, g_{t/t-1}\right)$$

Normal Distribution

$$p(x_t\,;\theta_t) = \frac{1}{\sqrt{2\pi g_{t/t-1}^2}}\frac{\left(x_t-A'\phi_{t/t-1}\right)'\left(x_t-A'\phi_{t/t-1}\right)}{2g_{t/t-1}} - \infty < x_t < +\infty \tag{9}$$

$$\ni \omega_t/X_{t-1} \sim N\left(0, g_{t/t-1}\right)$$

Its inverse, that is **Inverse Normal distribution** is

$$p(x_t\,;\theta_t) = \sqrt{\frac{\lambda}{\sqrt{2\pi x^3}}}exp\left[\frac{\lambda\left(x_t-A'\phi_{t/t-1}\right)'\left(x_t-A'\phi_{t/t-1}\right)}{2(A'\phi_{t/t-1})^{2x}}\right] \tag{10}$$

λ is the shape parameter. The inverse normal distribution always works on sided tail.

Skew-t Distribution

To accommodate asymmetry and long tailed data, Hansen (1994) introduced the so-called skewed-t-distribution while maintaining the property of a zero mean and variance equal to one. Skew-t-distribution is derived by introducing a universalization of the Student-t distribution as follows:

$$p(x_t\,;\lambda\,,r) = b\frac{\Gamma\left(\frac{r+1}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\sqrt{\pi(r-2)}}\left(1+\frac{\zeta^2}{r-2}\right)^{-\frac{r+1}{2}} \tag{11}$$

where,

$$\zeta = \begin{cases}(bx_t + a)/(1-\lambda) if x_t < -a/b \\ (bx_t + a)/(1+\lambda) if x_t \geq -a/b\end{cases}$$

such that the constant terms "a" and "b" are defined as: $a = 4\lambda c\frac{r-2}{r-1}$; $b = 1 + 3\lambda^2 - a^2$. In this distribution, $2 < r < \infty$ denotes the degrees of freedom parameter and $-1 < \lambda < 1$ is the asymmetry parameter.

Asymmetric Laplace

A random variable has an Asymmetric Laplace (μ, λ, κ) Distribution (ALD) if its probability density function is

$$p(x_t;\mu,\lambda,\kappa)=\left(\frac{\lambda}{\kappa+\frac{1}{\kappa}}\right)e^{-(x_t-\mu)\lambda S\kappa^S}$$

So, $$p(x_t;\phi_{t/t-1},\lambda,\kappa)=\left(\frac{\lambda}{\kappa+\frac{1}{\kappa}}\right)e^{-(x_t-A'\phi_{t/t-1})\lambda S\kappa^S} \quad (12)$$

where $S=sign(x_t-\mu)$

μ is a location parameter, $\lambda>0$is a scale parameter, and κ is an asymmetry parameter. When $\kappa=1$, $(x_t-\mu)S\kappa^S$simplifies to $|x_t-\mu|$ and the distribution simplifies to the Laplace distribution.

Gamma Distribution

A random variable X is said to be a Gamma distribution if:

$$p(x_t)=\left(\frac{x_t}{\beta}\right)^{\alpha-1}\times\frac{e^{\left(-\frac{x_t}{\beta}\right)}}{\beta\Gamma(\alpha)}x_t\in(0,\infty) \quad (13)$$

where $\Gamma(\alpha)=\int_0^\infty e^t t^{\alpha-1}\partial t$

with scale parameter$\beta>0$ and shape parameter$\alpha>0$.

Exponential distribution

A random variable X is said to be an exponential distribution (λ) if its probability density function is

$$p(x_t;\lambda)=\lambda e^{-\lambda x_t}x_t\in 0,\infty)$$
$$p(x_t;\lambda)=A'\phi_{t/t-1}e^{-A'\phi_{t/t-1}x_t} \quad (14)$$

The autoregressive score and information functions, hyper-parameters and autoregressive coefficients for the distributions specified from equation (6) to equation (14) can be estimated via the specifications made in equation (3), (4) and (5) using Maximum Likelihood (ML) or Reweighted Least Square Algorithm. See Creal *et al.* (2013), Harvey (2014), Olanrewaju & Folorunsho (2018).

3. Numerical Analysis

This section discusses the analyzes and results of the time-varying and time series hyper-parametric Generalized Autoregressive Scores (GASs) of the aforementioned conditional densities. The data to be subjected to the GASs with the random noise densities will be the averge monthly price of cooking gas in the United State from 1:2005

to 12:2020. The raw dataset of the price of the United State cooking gas will be used as extracted from U.S. Energy Information Administration (EIA). The monthly unit of the price is in US Dollar ($).

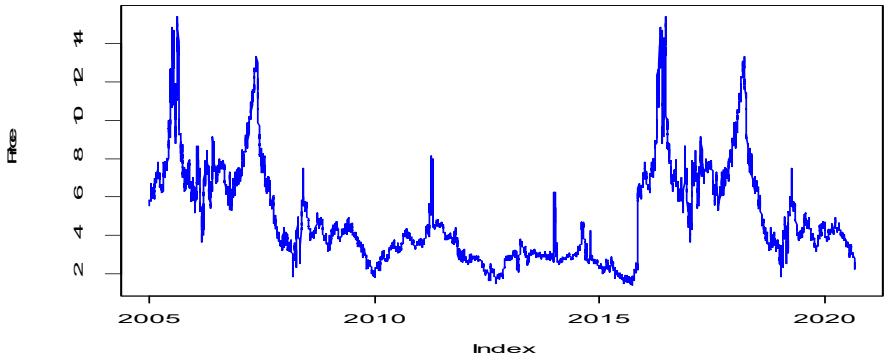


Figure 1: Time Plot of the Price of the Cooking Gas

From Figure 1, it is glaring that the monthly price of cooking gas in ($) was firstly pegged at around 6 ($) before skyrocketing to over 14($) towards ending of 2005 until 2006. It maintained an oscillating price between 12($) and 2($) from 2006 to 2015. The price also skyrocketed again mid-2015 to over 16($), it pendulum between 14($) and 10($) until around 2017 before a continuous drastic to 2($) was experienced. In general, from 2005 to 2020 the price of the cooking experienced a shocky zig-zag fluctuation.

Table 1: Coefficients of Skewness and Kurtosis

D'Agostino Skewness test	Skew. = 1.564	z = 29.540	P-value < 2.2e-16
Anscombe-Glynn kurtosis test	Kurt. = 5.5841	z = 15.6419	P-value < 2.2e-16
Bonett-Seier test for Geary kurtosis	tau = 1.7997	z = 10.6083	p-value < 2.2e-16

Under the hypothesis of normality, that is under the null hypothesis that the price of the cooking gas dataset is not skewed, which is the data should be symmetry (i.e. skewness should be equal to zero). However, since the D'Agostino Skewness coefficient is 29.540 with its P-value < 2.2e-16<0.05, there is sufficient evidence that the price of the cooking gas dataset is skewed with indication that the dataset is not normally distributed (this connotes that we fail to accept the null hypothesis). In a similar vein, since the Anscombe-Glynn kurtosis coefficient of 5.5841 is far greater than three, this suggested that the price of the US cooking gas is affected by heaviness in the tail of normal distribution. In collaboration, since Geary's kurtosis coefficient of

1.7997 ≠ sqrt(2/pi) (0.7979), tailedness of the normal distribution of the price of the cooking gas data is no doubt affected. Consequently, there is a need for hyper-parameters in the conditional densities to modify the lacuna.

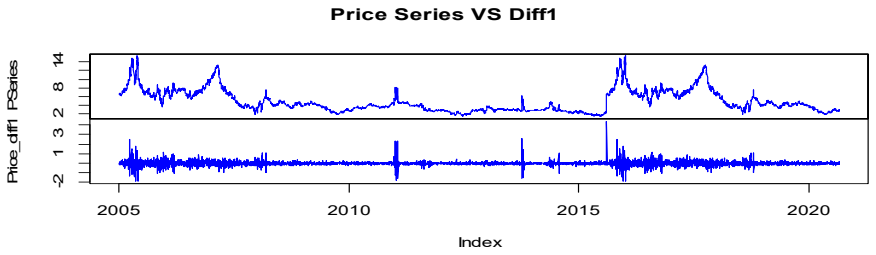


Figure 2: Time Plot versus the First Differencing Plot of the Price of the Cooking Gas

The upper visual time series plot in Figure 2 above is the raw plot of the price of the cooking gas in the US from 2005 to 2020, but was not stationary due to visual characterization of up and down shocks.

Table 2: Test of Stationarity and ARCH Effects for the Price of the US Cooking Gas

Estimates	ADF Test Statistic	Lag	P-value	LM Statistic	LM P-value
Price Series	-1.553	12	0.674	32.84	0.005
First Differencing	-50.715	12	0.01	26.678	0.0002

We tested the stationarity of the price of the US cooking gas via the Augmented Dickey-Fuller Test (ADF). We hypothesized both the price of the cooking series and its first differencing that their Null hypotheses display a unit root, that is both series are nonstationary. The number of lag used for testing is 12. The Test Statistic for the former was -1.553, while the latter gave -50.715. Since the p-value for the latter (first differencing) is 0.01 and the only one less than 0.05. We concluded that there is enough evidence to reject the Null hypothesis, meaning that the first differencing of the price series is the only one that is stationary. We also tested for Autoregressive Conditional Heteroskedasticity (ARCH) in order to ascertain if conditional variance on the information exists at a given point in time for both price of the cooking series and its first differencing. The formulated Null hypothesis for both series was there are no ARCH effects. Since the p-values for both the latter and former are less than 5% level of significance, the Null hypotheses are rejected and it is concluded that both series possessed ARCH effects. The first differencing series of the price of the cooking gas was used to model time-varying hyper-parameters for the GAS model because of its stationarity.

Table 3: Model Adequacy of the Density GAS w.r.t to the Price of United State Cooking Gas.

Specification	ddist	Pdist	qdist	Location-Score	Scale-Score	Skewness-Score	Shape1-Score	Shape-2 Score	Information-Location	Information-Scale	Information-Skewness	Information-Shape	Information-Shape2	AIC	BIC
Normal	-3.2807	0.0000	5.1845	-1.0278	0.3313	-----	-----	-----	0.4275	0.9941	-----	-----	-----	58242.958	58247.334
Inv. Normal	-3.6068	0.0026	1.1101	-0.8511	0.2783	-----	-----	-----	0.1676	0.0140	-----	-----	-----	12958.755	12994.380
Inv. Sqrt. Normal	-3.6068	0.0026	1.1101	-0.8511	0.2783	-----	-----	-----	0.1676	0.0140	-----	-----	-----	12958.755	12994.380
Skewed Normal	-6.2708	0.0050	2.7476	-2.3015	2.4255	0.0000	-----	-----	0.4460	0.0995	0.0000	-----	-----	12304.230	12357.666
Student-t	-3.9999	0.0004	1.5743	-1.2635	0.6636	0.0000	-----	-----	0.2691	0.0362	0.0000			12786.397	12839.833
Student-t Inv.	-3.9999	0.0004	1.5743	-1.2635	0.6636	0.0000	-----	-----	0.2691	0.0362	0.0000	-----	-----	12786.397	12839.833
Student-t Inv.Sqrt.	-3.9999	0.0004	1.5743	-1.2635	0.6636	0.0000	-----	-----	0.2691	0.0362	0.0000	----	-----	12786.397	12839.833
SkewStudent-t	-3.9999	0.0004	1.5743	-1.2635	0.6636	0.0000	0.0000	-----	0.2691	0.0362	0.0000	0.0000	-----	12198.097	12269.346
AST	-3.9999	0.0004	1.5743	-1.2635	0.6636	0.0000	0.0000	0.0000	0.2691	0.0362	0.0000	0.0000	0.0000	11983.633	12072.693
AST Inv.	-3.9999	0.0004	1.5743	-1.2634	0.6636	0.0000	0.0000	0.0000	0.2691	0.0362	0.0000	0.0000	0.0000	11983.633	12072.693
AST Inv. Sqrt.	-3.9999	0.0004	1.5743	-1.2635	0.6636	0.0000	0.0000	0.0000	0.2691	0.0362	0.0000	0.0000	0.0000	11983.633	12072.693
AST1 (Identitiy)	-3.9999	0.0004	1.5743	-1.2634	0.6636	-----	-----	-----	0.2691	0.0362	-----	-----	-----	**11943.277*****	**12014.525*****
ALD(Identity)	-2.8941	0.0001	0.7969	-1.8781	1.4196	-----	-----	-----	0.6879	0.2366	-----	-----	-----	11992.746	12046.182
Gamma (Identity)	-11.626	0.0192	3.4824	-5.0779	12.392	-----	-----	-----	1.0002	0.5002	-----	-----	-----	12153.334	12188.958
Exponential	-1.7321	0.0000	0.0100	5.0779	-----	-----	-----	-----	26.2950	-----	-----	-----	-----	14760.256	14778.068
Negative - Binomial	0.0000	0.0000	0.0000	**4.6542**	-0.1386	-----	-----	-----	41.4354	1.0000	-----	-----	-----	44052.93	44056.91
Skellam	-1.6243	0.1538	0.0000	-0.1655	-0.1129	-----	-----	-----	1.0000	1.0000	-----	-----	-----	12510.243	12545.867

Table 4: Model Adequacy of GARCH (1,1) Model w.r.t to the Price of United State Cooking Gas.

Specification	Estimate	Std. Error	t-value	Pr(>|t|)	Information Criteria			
					AIC	BIC	Shibata	Hannan-Quinn
Omega	0.4642	0.2905	1.5978	0.0101	22235.08	22237.99	22235.94	22236.034
Alpha1	0.9801	0.0684	14.3364	0.00000				
Beta1	0.0000	0.0630	0.0001	0.9999				

Description of each conditional density with respect to GAS or Dynamic Conditioal Correlation (DCC) was explicitly tabled in Table 1 and Table 2 (Table 1A in appendix). DCC is one of the most famous models for multivariate volatility. It uses multivariate GAS to model and analyze volatilities when the framework is based on score-driven time series for time-varying parameters. The model summary includes for each density of GAS includes their long-term value of the time-varying hyper-parameters, their estimated score and Hessian values, their model performance and concerned estimated autoregressive coefficients. Among the ten(10) hyper-parametric conditional noises that were subjected to the GAS model via the application of the price of the cooking gas, Asymmetric Student–t with one tail decay parameter (AST1) outperformed other category of its variants as well as other conditional densities for the GAS/DCC model with the minimum reduced-error performance of AIC=11943.277 and BIC=12014.525. The model hyper-parametric scores for the location-score and scale-score are -1.2634 and 0.6636 respectively. Its location-information and scale- information are 0.2691 and 0.0362 respectively. The concerned estimated coefficients of $kappa_1$, $kappa_2$, $kappa_3$ and $kappa_4$ are the elements of vector η i.e. η_μ, η_ϕ, η_{g_t},η_{v_t}which are 0.1444, 0.7434, -10195 and -0.8836 respectively. Analogously, a_1, a_2, a_3, a_4 are estimates of a_μ, a_ϕ, a_{g_t},b_{v_t} with 0.0000, 0.0000, 0.0000, and 0.0000 respectively, similarly to that of b_1, b_2, b_3, b with 0.9468, 0.5166, 0.1566 and 0.5993 respectively.

In comparison of the GAS or DCC model with the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model, it was affirmed that model GARCH(1,1) was the optimal lag (that is both Autoregressive (AR) and Moving Average (MA) at lag 1 each) for the volatiled price of the United State cooking gas studied over the period of time. It is to be noted that the GARCH model was subjected to different distributional error noises, like Student-t, Gaussian, Skew-Normal, etc., but Gaussian noise gave a robust generalization. The estimated model criteria of AIC, BIC, Shibata and Hannan-Quinn of 22235.08, 22237.99, 22235.94, and 22236.034 respectively for the model performance of the GARCH (1,1) model were far below the model performance of the GAS or DCC model via the random noise of the Asymmetric

Student–t with one tail decay parameter (AST1). The robustness of the GAS model via the Asymmetric Student–t with one tail decay parameter (AST1) might be via the location and scale scores of the noise.

4. Conclusions

This article introduced the possible conditional densities for the Generalized Autoregressive Score (GAS) model with embedded time-varying hyper-parameters. The score and Hessian functions (via location, scale, skewness, and shapes parameters) are of paramount interest due to their capability to curtail the lacuna of heaviness in the tail of normal distribution and possibility of skewed observations. Due to the flexibility of the GAS model to several statistical distributions, an empirical application to financial data of the price of the United State cooking gas was subjected to the GAS model with ten (10) different conditional densities. Each of the conditional density subjected to the GAS model via the application of the price of cooking gas from 2005 to 2020 was driven by the mechanism of time-varying score and Hessian functions of their embedded hyper-parameters. Asymmetric Student–t with one tail decay parameter (AST1) outperformed other category of its variants as well as other re-parameterized distributions used. In addition, the GAS model via Asymmetric Student–t with one tail decay parameter (AST1) random noise outshined the core volatile conditional heteroscedasticity of Generalized Autoregressive Conditional Heteroscedasticity (GARCH) with Gaussian distributed noise. For further studies, the conditional densities of the GAS model might be subjected to a driven mechanism of family of distributions with strictly positive values or integer values.

Acknowledgement

The authors personally extend gratitude to the US Energy Information Administration (EIA).

References

Blasques, F., Koopman, S. J., Łasak, K., Lucas, A., (2016). In-Sample Confidence Bands and Out-of-Sample Forecast Bands for Time-Varying Parameters in Observation-Driven Models. *International Journal of Forecasting*, Vol. 32(3), pp. 875–887, doi: 10.1016/j.ijforecast.2015. 11.018.

Cox, D. R., (1981). Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, Vol. 8(2), pp. 93–115.

Creal, D., Koopman, S.J., Lucas, A., (2013). Generalized Autoregressive Score Models with Applications. *Journal of Applied Econometric*, Vol. 28(5), pp. 777–795. https://doi.org/10.1002/jae.1279.

Janus, P., Koopman, S. J., Lucas, A., (2014). Long Memory Dynamics for Multivariate Dependence under Heavy Tails. *Journal of Empirical Finance*, Vol. 29, pp. 187–206, doi: 10.1016/j.jempfin, 2014.09.07.

Engle, R. F., (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica,* Vol. 50(4), pp. 987–1007.

Hansen, B. E., (1994). Autoregressive conditional density estimation. *International Economic Review*, Vol. 35 (3), pp. 705–729.

Harvey, A. C, Thiele, S. (2016). Testing Against Changing Correlation. *Journal of Empirical Finance*, Vol. 38(B), pp. 575–589. doi:10.1016/j.jempfin.2015.09.003.

Harvey, A. C., Sucarrat, G., (2014). EGARCH Models with Fat Tails, Skewness, and Leverage. *Computational Statistics & Data Analysis*, Vol.76, pp. 320–338, doi: 10.1016/j.csda.2013.09. 022.

Harvey, A. C., Luati, A., (2014). Filtering with Heavy Tails. *Journal of the American Statistical Association*, Vol. 109(507), pp. 1112–1122, doi: 10.1080/ 01621459.2014.887011.

Harvey, A. C., (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. Cambridge University Press.

Koopman, S. J., Lucas, A., Monteiro, A., (2008). The multi-state latent factor intensity model for credit rating transitions. *Journal of Econometrics,* Vol. 142(1), pp. 399–424.

Owen, C. B., (2008). *Parameter Estimation for the Beta Distribution*. Brigham Young University – Provo.

Jones, M. C., Faddy, M. J., (2003). A skew extension of the student t-distribution, with applications. *Journal of Royal Statistics Society B*,. Vol 65, Part 1, pp. 159–174.

Lucas, A., Zhang, X., (2016). Score-Driven Exponentially Weighted Moving Averages and Value-at-Risk Forecasting. *International Journal of Forecasting*, Vol. 32(2), pp. 293–302, doi: 10.1016/ j.ijforecast.2015.09.003.

Oh, D. H., Patton, A. J., (2016). Time-Varying Systemic Risk: Evidence from a Dynamic Copula Model of CDS Spreads. *Journal of Business* & *Economic Statistics,* Vol. 36(2), pp. 181–195, doi: 10.1080/07350015.2016.1177535.

Olanrewaju, R. O., Ojo J. F., Adekola, L. O., (2020). Bayesian latent autoregressive stochastic volatility: an application of naira to eleven exchangeable currencies rates. *Open Journal of Mathematical Sciences*, Vol.4 (1), pp. 386–396, doi: 10.30538/oms2020.0128.

Olanrewaju, R. O., Folorunsho, S. A., (2018). Generalized autoregressive score (GAS) functions under Gaussian and Student-t distributions. *International Journal of Statistics and Applied Mathematics*, Vol. 3(5), Part A, pp. 56–61.

Monache, D. D., Petrella, I., (2014). *Adaptive Models and Heavy Tails*. School of Economics and Finance, Working Paper No. 720, ISSN 1473-0278.

Shephard, N., (2005). *Stochastic Volatility: Selected Readings*. Oxford University Press, Oxford.

APPENDICES

Appendix 1.

Table 1A: Coefficients of the Hyper-Parameterization of the Generalized Autoregressive Scores (GASs)

a_3	a_4	a_5	b_1	b_2	b_3	b_4	b_5	Location	Scale	Skewness	Shape1	Shape2
-----	-----	-----	0.9789	0.8686	-----	-----	-----	5.4877	0.0334	-----	-----	-----
			(0.0009)	(0.0042)								
-----	-----	-----	0.5000	0.4999	-----	-----	-----	5.1279	5.9665	-----	-----	-----
			(4.4854)	(4.4967)								
-----	-----	-----	0.5000	0.4999	-----	-----	-----	5.1279	5.9665	-----	-----	-----
			(4.4854)	(4.4967)								
0.000001	-----	-----	0.8807	0.5166	0.5166	-----	-----	5.2107	2.2423	**1.**5000	-----	-----
(0.0000)			(0.0000)	(0.0000)	(0.000)							
0.000001	-----	-----	0.8807	0.5662	0.56620	-----	-----	4.7451	3.7160	5.0892	-----	-----
(0.0000)			(2.2661)	(0.0000)	(3.7787)							
0.000001	-----	-----	0.8807	0.5662	0.5662	-----	-----	4.7451	3.7160	-----	5.0892	-----
(0.0000)			(2.2661)	(3.7788)	(3.7788)						Realized=1.85	
0.000001	-----	-----	0.8807	0.5662	0.5662	-----	-----	4.7451	3.7160	-----	5.0892	-----
(0.0000)			(2.2661)	(0.0000)	(3.7788)							
0.000001	0.000001		0.7648	0.5166	0.5165	0.5166		4.9908	2.3230	**1.**4999	5.7047	
(0.0000)	(0.0000)	-----	(0.0000)	(2.0346)	(0.3261)	(1.0512)						4.0001
0.000001	0.0000	0.0000	0.8999	0.9138	0.9138	0.9138	0.9138	2.7768	4.3048	0.1259	4.0343	Realized=1.85
(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.5077)	(1.3458)	(0.0000)	(0.0000)					
0.000001	0.000001	**0.**000001	0.8999	0.9138	0.9138	0.9138	0.9138	2.7768	4.3047	0.1258	4.0342	4.0000
(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.5077)	(1.3457)	(2.5291)	(3.4201)					
0.000001	0.000001	0.000001	0.8999	0.9138	0.9138	0.9138	0.9138	2.7768	4.3047	0.1258	4.0342	4.0000
(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.5077)	(1.3457)	(1.1827)	(3.7102)					
0.000001	**0.000001**		**0.9468**	**0.5166**	**0.5166**	**0.5993**		**2.7203**	**4.6539**	**0.1161**	**8.5669**	-----
(0.0000)	**(0.0000)**	-----	**(0.0129)**	**(0.0008)**	**(0.0050)**	**(0.01431)**	-----					
0.000001			0.8476	0.5331	0.5496							
(0.0000)	----	-----	(0.0009)	(5.3925)	(0.0000)	-----	-----	2.7800	1.4536	0.3758	-----	-----
			3.3109	3.3109								
-----	-----	-----	(0.0000)	(0.0000)	-----	-----	-----	-----	5.1271	0.9998	-----	-----
			0.90000					0.1950				
-----	-----	-----	(0.1478)	-----	-----	-----	-----		-----	-----	-----	-----
			0.90000	0.90000				0.8593				
-----	-----	-----	(0.0000)	(0.0000)	-----	-----	-----		31.3185	-----	-----	-----
			0.5829	0.5131								
			(0.0002)	(0.0002)				5.1288	5.1333			

Appendix 2.

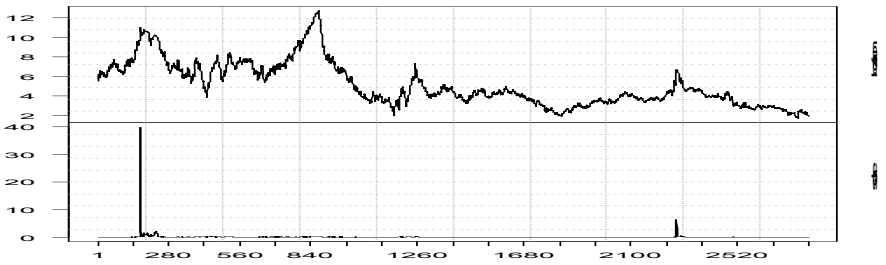


Figure 1A: Graphical Plot of the AST1 Location and Scale Parameters.

Appendix 3.

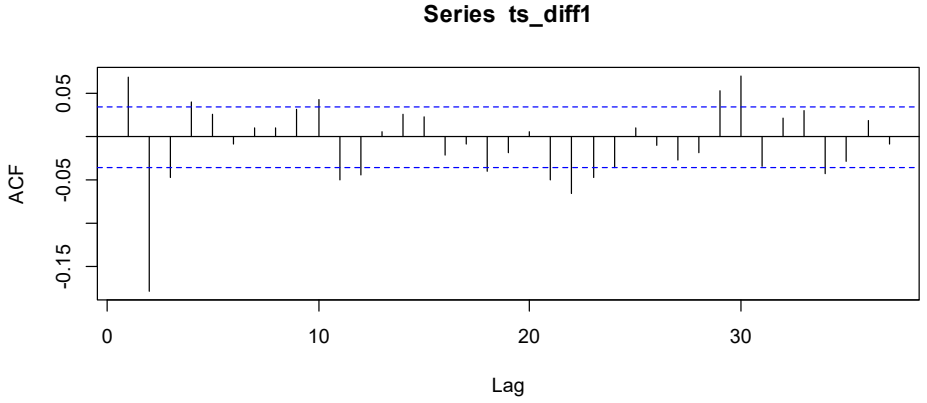


Figure A2: ACF Graph of the First Differencing of the Price of the US Cooking GAS

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 109–122, https://doi.org/10.59170/stattrans-2023-054
Received – 12.04.2022; accepted – 10.02.2023

Testing the annual rainfall dispersion in Chaiyaphum, Thailand, by using confidence intervals for the coefficient of variation of an inverse gamma distribution

Wararit Panichkitkosolkul[1]

Abstract

In Thailand, droughts are regular natural disasters that happen nearly every year due to several factors such as precipitation deficiency, human activity, and the global warming. Since annual rainfall amount fits an inverse gamma (IG) distribution, we wanted to try testing annual rainfall dispersion via the coefficient of variation (CV). Herein, we propose two statistics for testing the CV of an IG distribution based on the Score and Wald methods. We evaluated their performances by means of the Monte Carlo simulations conducted under several shape parameter values for an IG distribution based on empirical type I error rates and powers of the tests. The simulation results reveal that the Wald-method test statistic performed better than the Score-method one in terms of the attained nominal significance level, and is thus recommended for analysis in similar scenarios. Furthermore, the efficacy of the proposed test statistics was illustrated by applying them to the annual rainfall amounts in Chaiyaphum, Thailand.

Key words: statistical test, measure of dispersion, continuous distribution, simulation, meteorology.

1. Introduction

Since damage from natural disasters has increased due to anomalous global climate changes, researchers have become interested in studying their occurrences. Thailand has been divided into six geographical regions by the National Research Council: north, northeast, central, east, west, and south; many of them are prone to droughts but they most often occur in the central northeastern part of Thailand. Thailand is one of the most drought-affected countries in the Asia-Pacific region and is marred by frequent droughts (Khadka et al., 2021). Drought in Thailand directly affects agriculture and water resources, which has a significant impact on the country's economy since most of the country is agrarian.

The north-eastern of Thailand is one of the highly drought-prone regions of the country (Prabnakorn et al., 2018). Chaiyaphum, one of the north-eastern provinces of Thailand, is faced with drought every year due to long periods of little rain causing a severe shortage of water for both consumption and farming (Srichaiwong et al., 2020). Figure 1 shows the map of Chaiyaphum from the Google Maps (2023). In July 2019, parts of Chaiyaphum were faced with a severe drought, and the water volume in the Chulabhorn Dam decreased

[1]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathumthani 12120, Thailand. E-mail: wararit@mathstat.sci.tu.ac.th. ORCID: https://orcid.org/0000-0001-8315-8185.

to its lowest level in 30 years (only 25% of its capacity) (Pattayamail, 2022). Moreover, in January 2020, eight hospitals in Chaiyaphum were impacted by the drought, leading to the Chaiyaphum Provincial Public Health Office drilling artesian wells to reserve water for medical services and sufficient staff consumption for at least three days while also requesting citizens to help by saving water (Nationthailand, 2022).

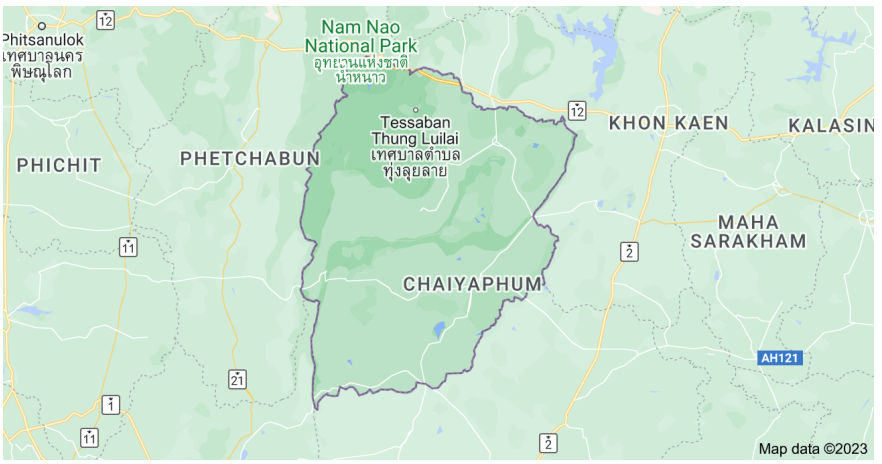


Figure 1: The map of Chaiyaphum, Thailand

Droughts take place whenever there are prolonged periods of rainfall deficiency for one season or more (Eartheclipse, 2022). The major cause of meteorological drought is a deficit of rainfall (Wichitarapongsakun et al., 2016). Since the rainfall amount varies greatly depending on the region and season, the coefficient of variation (CV) can be used to represent rainfall dispersion in different regions. The CV is a unit-free measure of variability relative to the population mean (Albatineh et al., 2017). It is defined as the ratio of the population standard deviation σ to the population mean μ namely $\theta = \sigma/\mu$, where $\mu \neq 0$. It has been more widely used than the standard deviation for comparing the variations of several variables obtained by different units.

The estimator of the CV has been widely applied in many fields of science, including the medical sciences, engineering, economics and others. For example, the applicability of the CV method for analyzing synaptic plasticity was studied by Faber and Korn (1991). Reed et al. (2002) used the CV in assessing the variability of quantitative assays. Kang et al. (2007) applied the CV for monitoring variability in statistical process control. Pang et al. (2008) proposed a simulation-based approach to the study of CV of dividend yields. The improved estimators of CV in a finite population were introduced by Archana and Rao (2011). Calif and Soubdhan (2016) used the CV to measure the spatial and temporal correlation of global solar radiation. Singh and Mishra (2019) proposed an improved estimation method for the population coefficient of variation, which uses information on a single auxiliary variable. Thangjai and Niwitpong (2020) proposed confidence interval estimation for the ratio of CV of two log-normal distributions constructed using the Bayesian approach.

The inverse gamma (IG) distribution is a two-parameter family of continuous distributions on the positive real line based on the reciprocal of a variable (Abid and Al-Hassany, 2016). Milevsky and Posner (1998) studied the IG distribution and pointed out estimation by method of moments. It is often used as a conjugate prior distribution in Bayesian statistics (Zhang and Zhang, 2022). There have been several research papers published on applying the IG distribution. For example, Gelman (2006) applied the IG distribution as a prior distribution for variance parameters in hierarchical models. Rasheed and Sultan (2015) proposed the Bayesian estimator for the scale parameter of IG distribution using Linex loss function and squared error loss function with non-informative prior. Abid and Al-Hassany (2016) studied some issues related to the inverted gamma distribution, which is the reciprocal of the gamma distribution. Llera and Beckmann (2016) introduced five different algorithms based on the method of moments, maximum likelihood, and Bayesian methodology to estimate the parameters of an IG distribution. Glen and Leemis (2017) applied the IG distribution to survival studies. Ramírez-Espinosa and Lopez-Martinez (2019) proposed the utility of the IG distribution in modeling composite fading channels. Yoo et al. (2019) provided empirical evidence that the IG distribution is an excellent alternative for the lognormal and gamma distributions which are often used to model shadowing. Furthermore, the confidence intervals for the ratio of the CVs of the IG distributions were introduced by Kaewprasert et al. (2023).

The literature on testing the CV for the IG distribution is limited. However, there are many methods available for estimating the confidence interval for a population CV for the IG distribution. Kaewprasert et al. (2020) presented three confidence intervals for the CV of an IG distribution using the Score method, the Wald method and the percentile bootstrap confidence interval. These confidence intervals for the CV can be used to test the hypothesis for the CV.

The objective of this paper is to propose some methods for testing the CV for the IG distribution and identify the appropriate methods for practitioners. Two confidence intervals proposed by Kaewprasert et al. (2020) are considered in order to test the CV. A simulation study was conducted to compare the performance of these methods. Based on the simulation results, test statistic with high power that attained a nominal significance level is recommended for practitioners.

The rest of this paper is organized as follows. The point estimation of parameters in an IG distribution is reviewed in Section 2. In Section 3, we present the proposed methods for testing the CV of the IG distribution. The simulation study and results are discussed in Section 4. Section 5 shows the application of the proposed statistical tests to real data is shown using the annual rainfall amounts in Chaiyaphum, Thailand. Discussion and conclusions are presented in the final section.

2. Point estimation of parameters in an inverse gamma distribution

In this section, we explain the point estimation of parameters in an IG distribution. Let $X = (x_1, \ldots, x_n)$ be a random sample from the IG distribution with the shape parameter α and scale parameter β, denoted as $IG(\alpha, \beta)$. The probability density function of X (Rivera et al., 2021) is given by

$$f(x;\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{-\alpha-1}\exp\left(-\frac{\beta}{x}\right),\ x>0,\ \alpha>0,\ \beta>0. \tag{1}$$

The population mean and variance of X are defined as $E(X)=\beta/(\alpha-1)$, for $\alpha>1$ and $Var(X)=\beta^2/\left[(\alpha-1)^2(\alpha-2)\right]$, for $\alpha>2$. Therefore, the CV of X can be expressed as $CV(X)=\theta=1/\sqrt{\alpha-2}$.

Since α is an unknown parameter, it is required to be estimated. We consider the maximum likelihood estimators (MLEs) for α and β. Thus, the log-likelihood function of α and β is given by

$$\ln L(\alpha,\beta)=-\sum_{i=1}^{n}\left(\frac{\beta}{X_i}\right)-(\alpha+1)\sum_{i=1}^{n}\ln(X_i)-n\ln\Gamma(\alpha)+n\alpha\ln(\beta).$$

Taking partial derivatives of the above equation with respect to α and β, respectively, the Score function is derived as

$$U(\alpha,\beta)=\begin{bmatrix}\sum_{i=1}^{n}\ln(X_i)-nln(\alpha)+\frac{n}{2\alpha}-n\ln(\beta)\\ -\sum_{i=1}^{n}X_i^{-1}+\frac{n\alpha}{\beta}\end{bmatrix}.$$

Then, the MLEs can be conducted for α and β, respectively,

$$\hat{\alpha}=\frac{1}{2\left[\frac{\sum_{i=1}^{n}\ln(X_i)}{n}+\ln\left(\frac{\sum_{i=1}^{n}X_i^{-1}}{n}\right)\right]},\ \text{and}\ \hat{\beta}=\frac{n\hat{\alpha}}{\sum_{i=1}^{n}X_i^{-1}}.$$

Also, the estimator of CV is given by $\hat{\theta}=1/\sqrt{\hat{\alpha}-2}$.

3. Methods for testing the coefficient of variation of the inverse gamma distribution

Let $X_1,...,X_n$ be an independent and identically distributed random sample of size n from the IG distribution with the shape parameter α and scale parameter β. We want to test for the population CV. The null and alternative hypotheses are defined as follows:

$$H_0:\theta=\theta_0\ \text{ versus }\ H_1:\theta\neq\theta_0.$$

In this section, we discuss two test statistics for the CV based on the Score method and the Wald method.

3.1. Score method

Let α and β be the parameter of interest and the nuisance parameters, respectively. In general, the Score statistic (Rao, 1948, 2005) is denoted as

$$W_1 = U^T(\alpha_0, \hat{\beta}_0) I^{-1}(\alpha_0, \hat{\beta}_0) U(\alpha_0, \hat{\beta}_0),$$

where $\hat{\beta}_0$ is the MLE for β, under the null hypothesis $H'_0 : \alpha = \alpha_0$, $U(\alpha_0, \hat{\beta}_0)$ is the vector of the Score function and $I(\alpha_0, \hat{\beta}_0)$ is the matrix of the Fisher information; see e.g., Kay (1993). Here, it is easy to derive that the Score function under H'_0 is

$$U(\alpha_0, \hat{\beta}_0) = \begin{bmatrix} -\sum_{i=1}^{n} \ln(X_i) + \frac{n}{2\alpha_0} - n \ln\left(n / \sum_{i=1}^{n} X_i^{-1} \right) \\ 0 \end{bmatrix}.$$

The inverse of the matrix of the Fisher information can be derived as follows:

$$I^{-1}(\alpha_0, \hat{\beta}_0) = \begin{bmatrix} \frac{2\alpha_0^2}{n} & -\frac{2\alpha_0^2}{\sum_{i=1}^{n} X_i^{-1}} \\ -\frac{2\alpha_0^2}{\sum_{i=1}^{n} X_i^{-1}} & \frac{n\alpha_0(2\alpha_0 - 1)}{\left(\sum_{i=1}^{n} X_i^{-1} \right)^2} \end{bmatrix}.$$

Using the property of the Score function, we can see that the pivotal

$$Z_{score} = \sqrt{\frac{2\alpha_0^2}{n}} \left[-\sum_{i=1}^{n} \ln(X_i) + \frac{n}{2\alpha_0} + n \ln\left(n / \sum_{i=1}^{n} X_i^{-1} \right) \right] \tag{2}$$

converges in distribution to the standard normal distribution. Since the variance of $\hat{\alpha}$ is $\frac{2\alpha_0^2}{n}$, it is approximated by substituting $\hat{\alpha}$ in its variance. Under H'_0, the statistic in (2) is given as

$$Z_{score} \cong \sqrt{\frac{2\hat{\alpha}^2}{n}} \left[-\sum_{i=1}^{n} \ln(X_i) + \frac{n}{2\hat{\alpha}} + n \ln\left(n / \sum_{i=1}^{n} X_i^{-1} \right) \right].$$

From the probability statement, $1 - \gamma = P(-Z_{1-\gamma/2} \leq Z_{score} \leq Z_{1-\gamma/2})$, it can be simply written as $1 - \gamma = P(l_s \leq \theta \leq u_s)$. Therefore, the $(1-\gamma)100\%$ confidence interval for θ based on the Score method is given by

$$CI_S = [l_s, u_s] = \left[\frac{1}{\sqrt{\frac{n}{2\left(z_1 - Z_{\gamma/2}\sqrt{\frac{n}{2\hat{\alpha}^2}}\right)}} - 2}, \frac{1}{\sqrt{\frac{n}{2\left(z_1 + Z_{\gamma/2}\sqrt{\frac{n}{2\hat{\alpha}^2}}\right)}} - 2} \right],$$

where $z_1 = \sum_{i=1}^{n} \ln(X_i) - n\ln\left(n/\sum_{i=1}^{n} X_i^{-1}\right)$ and $Z_{\gamma/2}$ is the $\gamma/2$-upper quantile of the standard normal distribution. Therefore, we will reject the null hypothesis, $H_0 : \theta = \theta_0$, if

$$\theta_0 < \frac{1}{\sqrt{\frac{n}{2\left(z_1 - Z_{\gamma/2}\sqrt{\frac{n}{2\hat{\alpha}^2}}\right)}} - 2} \text{ or } \theta_0 > \frac{1}{\sqrt{\frac{n}{2\left(z_1 + Z_{\gamma/2}\sqrt{\frac{n}{2\hat{\alpha}^2}}\right)}} - 2}.$$

3.2. Wald method

The Wald statistic is an asymptotic statistic derived from the property of the MLE (Gaffke et al., 2002). The general form of the Wald statistic under the null hypothesis $H_0' : \alpha = \alpha_0$ is defined as

$$W_2 = (\hat{\alpha} - \alpha_0)^T \left[I^{\alpha\alpha}(\hat{\alpha}, \hat{\beta}) \right]^{-1} (\hat{\alpha} - \alpha_0),$$

where $I^{\alpha\alpha}(\hat{\alpha}, \hat{\beta})$ is the estimated variance of $\hat{\alpha}$ obtained from the first row and the first column of $I^{-1}(\hat{\alpha}, \hat{\beta})$. Using the information of partial derivatives from the previous subsection, the inverse matrix is given by

$$I^{-1}(\hat{\alpha}, \hat{\beta}) = \begin{bmatrix} \frac{2\hat{\alpha}^2}{n} & -\frac{2\hat{\alpha}^2}{\sum_{i=1}^{n} X_i^{-1}} \\ -\frac{2\hat{\alpha}^2}{\sum_{i=1}^{n} X_i^{-1}} & \frac{n\hat{\alpha}(2\hat{\alpha}-1)}{\left(\sum_{i=1}^{n} X_i^{-1}\right)^2} \end{bmatrix},$$

where $I^{\alpha\alpha}(\hat{\alpha}, \hat{\beta}) = \frac{2\hat{\alpha}^2}{n}$. Therefore, under H_0', we obtain the Wald statistic

$$Z_{wald} \cong \sqrt{\frac{n}{2\hat{\alpha}^2}}(\hat{\alpha} - \alpha), \tag{3}$$

which has the limiting distribution of a standard normal distribution. Thus, the $(1-\gamma)100\%$ confidence interval for θ based on the Wald method is given by

$$CI_W = [l_w, u_w] = \left[\frac{1}{\sqrt{\hat{\alpha} - 2 + Z_{\gamma/2}\sqrt{\frac{2\hat{\alpha}^2}{n}}}}, \frac{1}{\sqrt{\hat{\alpha} - 2 - Z_{\gamma/2}\sqrt{\frac{2\hat{\alpha}^2}{n}}}} \right],$$

where $Z_{\gamma/2}$ is the $\gamma/2$-upper quantile of the standard normal distribution. Therefore, we will reject the null hypothesis, $H_0 : \theta = \theta_0$, if

$$\theta_0 < \frac{1}{\sqrt{\hat{\alpha} - 2 + Z_{\gamma/2}\sqrt{\frac{2\hat{\alpha}^2}{n}}}} \text{ or } \theta_0 > \frac{1}{\sqrt{\hat{\alpha} - 2 - Z_{\gamma/2}\sqrt{\frac{2\hat{\alpha}^2}{n}}}}.$$

4. Simulation Study and Results

In this study, two statistical methods for testing the population CV in an IG distribution are considered. Since a theoretical comparison is not possible, a Monte Carlo simulation was conducted using the R version 4.1.3 statistical software (Ihaka and Gentleman, 1996) to compare the performance of the test statistics. The methods were compared in terms of their attainment of empirical type I error rates and the powers of their performance. We count the number of times for each test that the null hypothesis was rejected when H_0 was true, to obtain the empirical type I error rates. In addition, the number of times for each test, that the null hypothesis was rejected when H_0 was not true, was counted to obtain the power of the test. The simulation results are presented for the significance level γ= 0.05, since a) γ= 0.05 is widely used to compare the power of the test and b) similar conclusions were obtained for other values of γ.

To observe the behaviour of small, moderate and large sample sizes, we used n = 25, 50, 75, 100 and 200. Each Monte Carlo experiment consisted of 10,000 replications. The data were generated from an IG distribution with β =1 and α was adjusted to obtain the required coefficient of variation θ. We set θ = 0.10, 0.15, 0.20 and 0.30.

As can be seen in the simulation results shown in Tables 1-4, the empirical type I error rates of the Wald method were close to the nominal significance level of 0.05 for all sample sizes while those of the Score method were close to the nominal significance level of 0.05 for larger sample sizes. Note that the Score method had a high empirical type I error rate when sample sizes were small. The Score method performed well in terms of the power of the test for $\theta < \theta_0$. On the other hand, the Wald method performed better for $\theta > \theta_0$. We observed a general pattern; when the sample size increases, the power of the test also increases and the empirical type I error rate approaches 0.05. Also, the power increases as the value of the CV departs from the hypothesized value of the CV. It was observed that for large sample sizes, the performance of the test statistics did not differ greatly in the sense of power and the attainment of the nominal significance level of the test. However, a significant difference was observed for small sample sizes.

Table 1. Empirical type I error rates (bold numeric) and powers of tests for IG(102,1), θ =0.10.

n	Method	θ_0								
		0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14
25	Score	0.6696	0.2744	0.0610	0.0413	**0.1155**	0.2711	0.4674	0.6806	0.8383
	Wald	0.9417	0.7357	0.3994	0.1433	**0.0431**	0.0483	0.1055	0.2268	0.3772
50	Score	0.9853	0.7963	0.3257	0.0583	**0.0820**	0.2996	0.6087	0.8623	0.9722
	Wald	0.9977	0.9455	0.6460	0.2192	**0.0432**	0.0962	0.2933	0.5913	0.8365
75	Score	0.9996	0.9561	0.5740	0.1078	**0.0680**	0.3319	0.7339	0.9436	0.9948
	Wald	1.0000	0.9883	0.7927	0.2888	**0.0420**	0.1454	0.4792	0.8088	0.9681
100	Score	1.0000	0.9911	0.7512	0.1785	**0.0688**	0.3806	0.8169	0.9816	0.9995
	Wald	1.0000	0.9982	0.8867	0.3592	**0.0469**	0.1976	0.6333	0.9316	0.9967
200	Score	1.0000	1.0000	0.9799	0.4410	**0.0595**	0.5595	0.9690	0.9999	1.0000
	Wald	1.0000	1.0000	0.9919	0.5863	**0.0496**	0.4089	0.9279	0.9992	1.0000

Table 2. Empirical type I error rates (bold numeric) and powers of tests for IG(46.44, 1), θ =0.15.

n	Method	θ_0								
		0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19
25	Score	0.1695	0.0592	0.0323	0.0542	**0.1131**	0.1966	0.3222	0.4602	0.5987
	Wald	0.6244	0.3933	0.1982	0.0933	**0.0398**	0.0350	0.0559	0.1014	0.1703
50	Score	0.6296	0.3168	0.0985	0.0401	**0.0794**	0.1973	0.3845	0.5925	0.7685
	Wald	0.8673	0.6352	0.3253	0.1241	**0.0476**	0.0580	0.1421	0.2815	0.4585
75	Score	0.8726	0.5564	0.2101	0.0549	**0.0702**	0.2085	0.4603	0.7160	0.8807
	Wald	0.9595	0.7910	0.4546	0.1564	**0.0446**	0.0814	0.2292	0.4591	0.6954
100	Score	0.9612	0.7378	0.3319	0.0750	**0.0651**	0.2276	0.5131	0.7890	0.9465
	Wald	0.9884	0.8882	0.5468	0.1856	**0.0465**	0.1005	0.3030	0.5954	0.8423
200	Score	0.9996	0.9757	0.6977	0.1775	**0.0556**	0.3102	0.7294	0.9599	0.9979
	Wald	0.9998	0.9909	0.8137	0.2988	**0.0501**	0.1901	0.5836	0.9112	0.9923

Table 3. Empirical type I error rates (bold numeric) and powers of tests for IG(27, 1), θ =0.20.

n	Method	θ_0								
		0.16	0.17	0.18	0.19	0.20	0.21	0.22	0.23	0.24
25	Score	0.0531	0.0332	0.0383	0.0683	**0.1146**	0.1693	0.2406	0.3386	0.4419
	Wald	0.3825	0.2363	0.1356	0.0727	**0.0421**	0.0319	0.0382	0.0617	0.0920
50	Score	0.2990	0.1337	0.0525	0.0418	**0.0782**	0.1537	0.2766	0.4017	0.5590
	Wald	0.6224	0.4029	0.2099	0.0950	**0.0446**	0.0501	0.0924	0.1523	0.2578
75	Score	0.5438	0.2697	0.0988	0.0428	**0.0677**	0.1550	0.3047	0.4879	0.6761
	Wald	0.7710	0.5278	0.2700	0.1097	**0.0497**	0.0571	0.1323	0.2445	0.4195
100	Score	0.7210	0.4130	0.1611	0.0571	**0.0616**	0.1630	0.3386	0.5600	0.7552
	Wald	0.8735	0.6327	0.3462	0.1282	**0.0485**	0.0642	0.1731	0.3494	0.5535
200	Score	0.9734	0.7889	0.4010	0.1028	**0.0545**	0.2039	0.4865	0.7950	0.9462
	Wald	0.9889	0.8803	0.5573	0.1895	**0.0495**	0.1178	0.3402	0.6589	0.8870

Table 4. Empirical type I error rates (bold numeric) and powers of tests for IG(13.11, 1), θ =0.30.

n	Method	θ_0								
		0.26	0.27	0.28	0.29	0.30	0.31	0.32	0.33	0.34
25	Score	0.0324	0.0431	0.0594	0.0784	**0.1094**	0.1444	0.1834	0.2214	0.2787
	Wald	0.1879	0.1306	0.0915	0.0557	**0.0429**	0.0328	0.0308	0.0345	0.0457
50	Score	0.0858	0.0535	0.0427	0.0534	**0.0736**	0.1136	0.1653	0.2352	0.3164
	Wald	0.2997	0.2058	0.1178	0.0717	**0.0481**	0.0410	0.0513	0.0759	0.1034
75	Score	0.1832	0.0934	0.0534	0.0426	**0.0623**	0.1103	0.1775	0.2544	0.3573
	Wald	0.4103	0.2550	0.1497	0.0818	**0.0511**	0.0474	0.0676	0.1048	0.1649
100	Score	0.2806	0.1494	0.0678	0.0442	**0.0603**	0.1024	0.1809	0.2853	0.4150
	Wald	0.5001	0.3153	0.1690	0.0878	**0.0497**	0.0474	0.0816	0.1320	0.2227
200	Score	0.6430	0.3718	0.1668	0.0652	**0.0517**	0.1115	0.2307	0.4030	0.5970
	Wald	0.7682	0.5246	0.2832	0.1242	**0.0513**	0.0618	0.1361	0.2686	0.4456

5. An Empirical Application

To illustrate the applicability of the two statistical methods for testing the CV introduced in the previous section, we used annual rainfall data in millimetres obtained from the Hydrology Irrigation Center for the Upper Northeastern Region, the Royal Irrigation Department, Thailand (http://hydro-3.rid.go.th). The annual rainfall amounts were measured at the Irrigation Station, Mueang District, Chaiyaphum, Thailand from 1998 to 2021. The descriptive statistics are as follows: sample size = 23, mean = 1088.44 mm, standard deviation (SD) = 245.79 mm, CV = 0.226, coefficient of skewness = 0.886, and kurtosis = 0.946. The distribution of the annual rainfall amount is right-skewed and it has heavy-tailed data distribution. The histogram, density plot, Box and Whisker plot, and inverse gamma quantile-quantile (Q-Q) plot shown in Figure 1 confirm that the fitted distribution for the annual rainfall amounts is not symmetric.

Table 5 reports the Akaike information criterion (AIC) (Akaike, 1974) results to check the fitting of the distribution for the annual rainfall amounts in Chaiyaphum. The AIC is defined as $AIC = -2\ln L + 2k$, where L is the likelihood function and k is the number of parameters. The results show that the annual rainfall amounts in Chaiyaphum follow an IG distribution because the AIC value for this distribution was the smallest. The annual rainfall amounts in Chaiyaphum had an IG distribution with shape parameter $\hat{\alpha}$= 26.8951 and scale parameter $\hat{\beta}$= 23588.810, while the MLE for the CV is $\hat{\theta}$= 0.2148.

Our interest was in testing the population CV of the annual rainfall amounts in Chaiyaphum. Suppose the researcher wanted to test the claim that a population CV equals 0.25. The null and alternative hypotheses are respectively given as follows:

$$H_0 : \theta = 0.25 \text{ versus } H_1 : \theta \neq 0.25.$$

The lower and upper critical values of both test statistics were shown in Table 6. The null hypothesis H_0 was not rejected since $0.1390 \leq \theta_0 \leq 0.2843$ and $0.1721 \leq \theta_0 \leq 0.3634$ us-

ing test statistics based on the Score and Wald methods, respectively. We conclude that the population CV of the annual rainfall amounts in Chaiyaphum does not differ from 0.25 at the 0.05 significance level.

Table 5. Results of AIC for the annual rainfall amounts in Chaiyaphum, Thailand.

Normal	Cauchy	Exponential	Weibull	Gamma	Inverse Gamma
321.4549	328.0113	369.6550	323.7112	319.2704	**318.2490**

Table 6. Critical values of test statistics based on the Score and Wald methods for the significance level of 0.05

Method	Critical values	
	Lower	Upper
Score	0.1390	0.2843
Wald	0.1721	0.3634

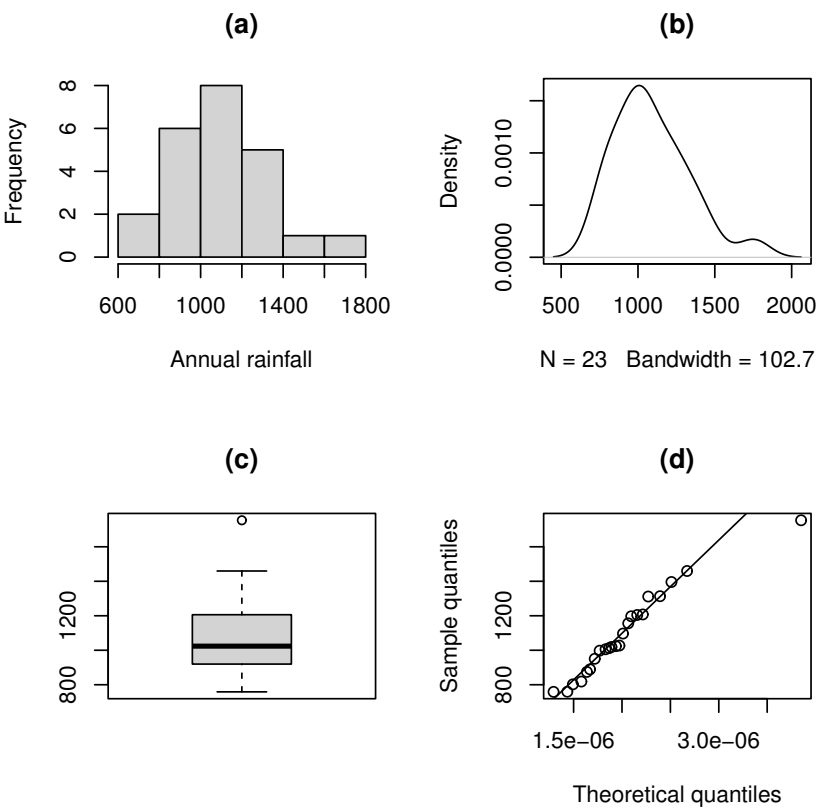


Figure 2: (a) histogram (b) density plot (c) Box and Whisker plot (d) inverse gamma Q-Q plot of the annual rainfall amounts in Chaiyaphum, Thailand

6. Conclusions and Discussion

The aim of this study is to identify potential methods that can be recommended to practitioners for testing the population CV in an IG distribution. A general pattern was observed (as expected); as the sample size increased, the power of the test also increased and the empirical type I error rates approached 0.05. Moreover, the power increased as the value of CV departed from the hypothesized value of the CV. It can be observed that for large sample sizes, the performance of both methods did not differ greatly in terms of the power and attaining the nominal size of the test. However, a significant difference was observed for small sample sizes.

In this study, two statistical methods for testing the population CV in an IG distribution were derived. Based on the simulation results, it is evident that the Wald method performed better than the Score method in terms of the empirical type I error rate. The Score method performed well in the sense of the power of the test when the population CV was smaller than the hypothesized value of the CV. On the other hand, the Wald method performed better when the population CV was greater than the hypothesized value of the CV. In summary, we would recommend the Wald method for testing since its empirical type I error rate is close to the nominal significance level. Furthermore, Kaewprasert et al. (2020) concluded that the best method for estimating confidence interval for the CV of the IG distribution was the Wald method. The conclusions of this study were consistent with the study of Kaewprasert et al. (2020). In addition, the researchers can apply the proposed methods for testing the population CV in an IG distribution with other data sets fitted well to an IG distribution. For example, the IG distribution has been used for the hitting time distribution of a Wiener process. Future research could focus on the one-tailed hypothesis testing.

Acknowledgements

The author would like to thank editor and the reviewers for the valuable comments and suggestions to improve this paper.

References

Abid, S. H., Al-Hassany, S. A., (2016). On the inverted gamma distributions. *International Journal of Systems Science and Applied Mathematics*, Vol. 1, pp. 16–22.

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

Albatineh, A. N., Boubakari, I. and Kibra, B. M. G., (2017). New confidence interval estimator of the signal-to-noise ratio based on asymptotic sampling distribution. *Communication in Statistics-Theory and Methods*, Vol. 46, pp. 574–590.

Archana, V., Rao, K. A., (2011). Improved estimators of coefficient of variation in a finite population, *Statistics in Transition new series*, Vol. 12, pp. 357–380.

Calif, R., Soubdhan, T., (2016). On the use of the coefficient of variation to measure spatial and temporal correlation of global solar radiation. *Renewable Energy*, Vol. 88, pp. 192–199.

Eartheclipse, (2022). *What is a Drought?*. [online]. Retrieved from https://eartheclipse.com/natural-disaster/causes-and-effects-of-drought.html, 2 April 2022.

Faber, D. S., Korn, H., (1991). Applicability of the coefficient of variation method for analyzing synaptic plasticity. *Biophysical Journal*, Vol. 60, pp. 1288–1294.

Gaffke, N., Heiligers, B. and Offinger, H., (2002). On the asymptotic null-distribution of the Wald statistic at singular parameter points. *Statistics & Decisions*, Vol. 20, pp. 379–398.

Gelman, A., (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, Vol. 1, pp. 515–533.

Glen, A. G., Leemis, L. M., eds. (2017). Computational Probability Applications. *International Series in Operations Research & Management Science*, Vol. 247, Cham: Springer.

Google Maps, (2023). Chaiyaphum, Available from: https://www.google.co.th/maps/@16.0082727, 101.33255,9z?hl=en [Accessed 8 February 2023].

Ihaka, R., Gentleman, R., (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, Vol. 5, 299–314.

Kaewprasert, T., Niwitpong, S. A. and Niwitpong, S., (2020). Confidence interval for coefficient of variation of inverse gamma distributions, In: HUYNH, V. N., ENTANI, T., Jeenanunta, C., Inuiguchi, M., Yenradee, P. (Eds.). *Lecture Notes in Artificial Intelligence: Vol.12482. Integrated Uncertainty in Knowledge Modelling and Decision Making*, pp. 407–418, Cham: Springer.

Kaewprasert, T., Niwitpong, S. A. and Niwitpong, S., (2023). Confidence intervals for the ratio of the coefficients of variation of inverse-gamma distributions. *Applied Science and Engineering Progress*, Vol. 16, doi.org/10.14416/ j.asep.2021.12.002.

Kang, C. W., Lee, M. S., Seong, Y. J. and Hawkins, D. M., (2007). A control chart for the coefficient of variation. *Journal of Quality Technology*, Vol. 39, pp. 151–158.

Kay, S. M., (1993). Fundamentals of statistical signal processing: Estimation theory. Englewood Cliffs: Prentice-Hall.

Khadka, D., Babel, M. S., Shrestha, S., Virdis and S. G. P., Collins, M., (2021). Multivariate and multi-temporal analysis of meteorological drought in the northeast of Thailand. *Weather and Climate Extremes*, Vol. 34, https://doi. org/10.1016/j.wace.2021.100399.

Llera, S., Beckmann, C. F., (2016). Estimating an inverse gamma distribution. *Technical report*, Radbound University Nijmegen, Donders Institute for Brain Cognition and Behavior. ar.Xiv: 1605.01019v2.

Milevsky, M. A., Posner, S. E., (1998). Asian options, the sum of log-normals and the reciprocal gamma distribution. *The Journal of Financial and Quantitative Analysis*, Vol. 33, pp. 203–218.

Nationthailand, (2022). *Nine hospitals in provinces suffer impact from drought*, [online]. Retrieved from https://www.nationthailand.com/in-focus/30380811, 5 April 2022.

Pang, W. K., Yu, B. W. T., Troutt, M. D. and Hou, S. H., (2008). A simulation-based approach to the study of coefficient of variation of dividend yields. *European Journal of Operational Research*, Vol. 189, pp. 559–569.

Pattayamail, (2022). *Drought situation in Chaiyaphum reaches crisis point*, [online]. Retrieved from https://www.pattayamail.com/thailandnews/drought-situation-in-chaiyaphum-reaches-crisis-point-260702, 5 April 2022.

Prabnakorn, S., Maskey, S., Suryadi, F. X. and de Fraiture, C., (2018). Rice yield in response to climate trends and drought index in the Mun River Basin, Thailand. *Science of the Total Environment*, Vol. 621, pp. 108–119.

Rao, C. R., (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 44, pp. 50–57.

Rao, C. R., (2005). Score Test: Historical Review and Recent Developments. In: Advances in Ranking and Selection, Multiple Comparisons, and Reliability. Boston: Birkhäuser, pp. 3–20.

RamÍRez-Espinosa, P. Lopez-Martinez, F. J., (2019). On the utility of the inverse gamma distribution in modeling composite fading channels. In: 2019 IEEE Global Communications Conference (GLOBECOM). Waikoloa, Hawaii, USA, 9–13 December 2019, New York: Curran Associates.

Rasheed, H. A., Sultan, A. J., (2015). Bayesian estimation of the scale parameter for inverse gamma distribution under Linex loss function. *International Journal of Advanced Research*, Vol. 3, pp. 369–375.

Reed, G. F., Lynn, F. and Meade, B. D. (2002). Use of coefficient of variation in assessing variability of quantitative assays. *Clinical and Diagnostic Laboratory Immunology*, Vol. 9, pp. 1235–1239.

Rivera, P. A., CalderÍN-Ojeda, E., Gallardo D. I. and Gómez H. W., (2021). A compound class of the inverse gamma and power series distributions. *Symmetry*, Vol. 13, https://doi.org/10.3390/sym13081328.

Singh, R., Mishra, M., (2019). Estimating population coefficient of variation using a single auxiliary variable in simple random sampling. *Statistics in Transition new series*, Vol. 20, pp. 89–111.

Srichaiwong, P., Ardwichai, S., Tungchuvong, L. and Kenpahoom, S., (2020). The live weir innovation at Chi river watershed, Chaiyaphum province, Thailand. *Bioscience Biotechnology Research Communications*, Vol. 13, pp. 103–107.

Thangjai, W., Niwitpong, S., (2020). Comparing particulate matter dispersion in Thailand using the Bayesian confidence intervals for ratio of coefficients of variation, *Statistics in Transition new series*, Vol. 21, pp. 41–60.

Wichitarapongsakun, P., Sarin, C., Klomjek, P. and Chuenchooklin, S. (2016). Rainfall prediction and meteorological drought analysis in the Sakae Krang River basin of Thailand. *Agriculture and Natural Resources*, Vol. 50, pp. 490–498.

Yoo, S. K., Cotton, S. L., Zhang, L. and Sofotasios, P. C., (2019). The inverse gamma distribution: a new shadowing model, In: 2019 8th Asia-Pacific Conference on Antennas and Propagation (APCAP), Incheon, Korea (South), 4–7 August 2019, pp. 475–476.

Zhang, L., Zhang, Y. Y., (2022). The Bayesian posterior and marginal densities of the hierarchical gamma–gamma, gamma–inverse gamma, inverse gamma–gamma, and inverse gamma–inverse gamma models with conjugate priors. *Mathematics*, Vol. 10, https://doi.org/10.3390/math10214005.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 123–138, https://doi.org/10.59170/stattrans-2023-055
Received – 09.01.2021; accepted – 17.05.2022

Advances in estimation by the item sum technique in two move successive sampling

Kumari Priyanka[1], **Pidugu Trisandhya**[2]

Abstract

The present article proposes an estimator using the Item Sum Technique (IST) for the estimation of dynamic sensitive population mean using non-sensitive auxiliary information in the two-move successive sampling. Properties of the proposed IST estimator have been analysed. Possible allocation designs for allocating long-list and short-list samples pertaining to the IST have been elaborated. The comparison between various allocation designs has been carried out. Theoretical considerations have been integrated with numerical as well as simulation studies to show the working version of the proposed IST estimators in the two-move successive sampling.

Key words: Sensitive variable, Successive moves, Population mean, Variance, Mean squared error, Optimum matching fraction.

1. Introduction

In many social surveys, data gathering on sensitive issues such as incidence of domestic violence, drug addiction, eve teasing, negligence of government rules, duration of suffering from AIDS, use of harmful pesticides in agriculture, sexual behaviour, etc., are a challenging task in the present scenario. Hence, in such circumstances, many respondents either refuse to participate or give false or evasive responses in social surveys. Therefore, to overcome mis-reporting on sensitive issues and to protect respondents confidentiality, the Randomized Response (RR) technique, the Scrambled Response (SR) technique, Item Count Technique (ICT), etc., may be used.

The RR technique was first initiated by Warner (1965) which was followed by Horvitz *et al.* (1967), Greenberg *et al.* (1971), Franklin (1989), Arcos *et al.* (2015), etc. However, SR technique was introduced by Pollock and Bek (1976) and was further explored by Eichhorn and Hayre (1983), Diana and Perri (2010, 2011), Perri and Diana (2013), etc. The ICT is used in surveys that require the study of qualitative sensitive variable and was introduced by Miller (1984). Subsequently the literature addressing ICT was enhanced by Droitcour *et al.* (1991), Wimbush and Dalton (1997), LaBrie and Earleywine (2000), Rayburn *et al.* (2003), and Tsuchiya *et al.* (2007), Holbrook and Krosnick (2010)etc. For estimating quantitative sensitive variable, the concept of ICT was generalized by Chaudhuri and

[1]Department of Mathematics, Shivaji College(University of Delhi), New Delhi-110027, India.
E-mail: priyanka.ism@gmail.com. ORCID:https://orcid.org/0000-0002-6241-1708.
[2]Department of Applied Sciences, Bharati Vidyapeeth's College of Engineering, New Delhi-110063, India.
E-mail: trisandhya.09@gmail.com. ORCID:https://orcid.org//0000-0002-6671-6541.

Christofides (2013). Trappmann *et al.* (2014) named this technique IST. Perri *et al.* (2018) discussed the possibility of optimal sample size allocation in IST.

As the issues are sensitive, single time survey will not be sufficient, one need to monitor continuously over a period of time. So, to observe the situation at different point of time, a statistical tool generally recommended in the literature is 'Successive Sampling'. In order to address the dynamic sensitive variable Arnab and Singh (2013), Yu *et al.* (2015), Priyanka *et al.* (2018, 2019), Priyanka and Trisandhya (2019), etc. added valuable literature. To handle sensitive issues all these researchers dealt with SR technique or RR technique on two move successive sampling. As the IST is now emerging as an alternative technique to deal with sensitive issues, in the present article an attempt has been made to apply IST in successive sampling framework to estimate a sensitive population mean. The concept of linear estimators has been used under IST set-up on successive move which is a methodological advancement to the theory. Hence, IST class of estimator has been proposed and studied under general allocation design advocated by Trappmann *et al.* (2014) as well as optimum allocation design suggested by Perri *et al.* (2018). Properties of the proposed class of estimator have been discussed in detail. Empirical as well as simulation studies have been incorporated to justify the requirement and application of the proposed estimator using a natural population. The proposed estimator has also been compared with respect to direct version of the estimator to show the amount of loss incurred due to sensitivity management of the variable under study by IST.

2. Outline of the Item Sum Technique(IST)

A promising indirect questioning technique, called Item Count Technique [Miller (1984)] is proved to be a very useful technique to estimate the qualitative sensitive variable. In this technique each respondent is provided with a list of items describing behaviors and asked to count and report in how many he or she is engaged in and not in which ones. For example, a random subsample (say subsample A) is provided with three-item list that includes the socially disproved behaviour item; the remaining respondents (say subsample B) are given an identical (two-item) list from which the disapproved item has been removed. By comparing responses from the two subsamples, an estimate of sensitive behavior has been obtained.

This method of estimating qualitative sensitive characteristics was generalized by Chaudhuri and Christofides (2013) to work for estimation of quantitative sensitive variates. Later Trappmann *et al.*(2014) explored it and named the technique as Item Sum Technique, which is described as follows.

Two random sub-samples (say s_1 and s_2) are drawn from a random sample (say s). The respondent belonging to sub-sample s_1 is given a long list (LL) of items containing sensitive question and a number of non-sensitive questions. However, the respondents in sub-sample s_2 are confronted with a short list (SL) of items containing only the same non-sensitive questions present in the long list (LL). In both the sub-samples, the respondents are asked to respond only the total score of all the items given to them, without revealing the individual scores for the items. Finally, the mean difference of answers between the samples s_1 and s_2 is used as an unbiased estimator of the population mean of sensitive variable. The pivotal

point in IST is how to split a single sample(s) into two-samples (s_1 and s_2). Trappmann *et al.* (2014), allocated equal number of units in each sub-sample irrespective of the variation of items in the *LL* and *SL*. Let us name this allocation design 'General Allocation' for further use. However, Perri *et al.* (2018) advocated the concept of 'Optimum allocation design' for allocating units in two sub-samples (s_1 and s_2) instead of assigning equal number of units to both the sub-samples. In the next section the IST set-up is modified to be applied in successive sampling to estimate population mean of dynamic sensitive variable.

3. Survey Design

Let us consider a finite population U of size N units for sampling over two successive moves. Let $y_1(y_2)$ denote quantitative sensitive variable at first (second) move respectively. Similarly, let x and t be non-sensitive auxiliary variables available at both the moves. Let $\bar{Y}_1, \bar{Y}_2, \bar{X}$, and $\bar{T}$ be the population mean of y_1, y_2, x, and t respectively. The aim is to estimate the population mean of sensitive variable y_2 at current move under IST set-up for two move successive sampling. The sampling design under IST frame work is as follows.

At the first move, a sample s_n of size n is drawn using simple random sampling without replacement(SRSWOR). Two independent samples are drawn at the second move by considering the partial overlap case, one is matched sample s_m of size $m = n(1-\mu) = n\lambda$ drawn as sub sample from the sample s_n and other is the sample s_u which is of size $u = n - m = n\mu$ drawn afresh at the second move. Let s_{u^*} denote the left out units from s_n after drawing the sub sample s_m. Moreover all the available samples s_{u^*}, s_m and s_u are split in to two sub samples called *LL* sample and *SL* sample respectively for embedding the IST set-up in two move successive sampling, which is given in Table 1:

Table 1: *LL* and *SL* Sample on two moves

Move	**Sample**	$LL-$ **Sample**	$SL-$ **Sample**
I	s_{u^*}	$s_{u_1^*}$	$s_{u_2^*}$
	s_m	s_{m_1}	s_{m_2}
II	s_m	s_{m_1}	s_{m_2}
	s_u	s_{u_1}	s_{u_2}

Note: s_m denotes matched sample and s_u denotes unmatched sample at current (second) move

The response received and the corresponding IST estimate on two moves under IST set-up are presented in Table 2.

Table 2: Response received and IST estimate

Move	Sample size	Response received	IST estimate
I	u^*	$z_{1i} = \begin{cases} y_{1i}+t_i & \text{if } i\varepsilon s_{u_1^*} \\ t_i & \text{if } i\varepsilon s_{u_2^*} \end{cases}$	$\hat{\bar{y}}_{1u^*} = \bar{z}_{1u_1^*} - \bar{t}_{u_2^*}$
	m	$z_{1i} = \begin{cases} y_{1i}+t_i & \text{if } i\varepsilon s_{m_1} \\ t_i & \text{if } i\varepsilon s_{m_2} \end{cases}$	$\hat{\bar{y}}_{1m} = \bar{z}_{1m_1} - \bar{t}_{m_2}$
II	m	$z_{2i} = \begin{cases} y_{2i}+t_i & \text{if } i\varepsilon s_{m_1} \\ t_i & \text{if } i\varepsilon s_{m_2} \end{cases}$	$\hat{\bar{y}}_{2m} = \bar{z}_{2m_1} - \bar{t}_{m_2}$
	u	$z_{2i} = \begin{cases} y_{2i}+t_i & \text{if } i\varepsilon s_{u_1} \\ t_i & \text{if } i\varepsilon s_{u_2} \end{cases}$	$\hat{\bar{y}}_{2u} = \bar{z}_{2u_1} - \bar{t}_{u_2}$

Note: z_{ji} ; $j = 1, 2$ denote the observed response at first and second move respectively and $\bar{z}_{1j}$; $j \in \{u_1^*, m_1\}, \bar{z}_{2j}$; $j \in \{m_1, u_1\}$ and $\bar{t}_k$; $k \in \{u_2^*, m_2, u_2\}$ are the sample means based on sample size j and k.

4. Proposed class of IST Estimators

Inspired by the classic work of Patterson (1950), who considered a general linear unbiased estimator of population mean at the current move, we intend to propose an estimator for estimation of sensitive population mean at the current move in IST set-up using all the information available at the current move.

In sampling theory, the role of additional auxiliary variable is well known and its availability and use in estimation procedures can do wonders and enhance the results to a great extent. Hence, in IST set-up, the availability of additional auxiliary variable has been embedded and class of IST estimator has been proposed to estimate sensitive population mean at current move as under:

$$\mathbb{T} = \zeta_1 \hat{\bar{y}}^*_{1u^*} + \zeta_2 \hat{\bar{y}}^*_{1m} + \zeta_3 \hat{\bar{y}}^*_{2m} + \zeta_4 \hat{\bar{y}}^*_{2u}, \tag{1}$$

where, the constants ζ_j ; $j = 1, 2, 3,$ and 4 are to be chosen suitably and $\hat{\bar{y}}^*_{2u} = \hat{\bar{z}}^*_{2u} - \hat{\bar{t}}^*_u$ with $\hat{\bar{z}}^*_{2u} = g_1(\bar{z}_{2u_1}, \bar{x}_{u_1})$, and $\hat{\bar{t}}^*_u = h_1(\bar{t}_{u_2}, \bar{x}_{u_2})$,
$\hat{\bar{y}}^*_{2m} = \hat{\bar{z}}^*_{2m} - \hat{\bar{t}}^*_m$ with $\hat{\bar{z}}^*_{2m} = g_2(\bar{z}_{2m_1}, \bar{x}_{m_1})$, and $\hat{\bar{t}}^*_m = h_2(\bar{t}_{m_2}, \bar{x}_{m_2})$,
$\hat{\bar{y}}^*_{1m} = \hat{\bar{z}}^*_{1m} - \hat{\bar{t}}^*_m$ with $\hat{\bar{z}}^*_{1m} = g_3(\bar{z}_{1m_1}, \bar{x}_{m_1})$,
$\hat{\bar{y}}^*_{1u^*} = \hat{\bar{z}}^*_{1u^*} - \hat{\bar{t}}^*_{u^*}$ with $\hat{\bar{z}}^*_{1u^*} = g_4(\bar{z}_{1u_1^*}, \bar{x}_{u_1^*})$, and $\hat{\bar{t}}^*_{u^*} = h_3(\bar{t}_{u_2^*}, \bar{x}_{u_2^*})$.

Following Srivastava and Jhajj (1980) and Tracy *et al.* (1996), $g_1(\bar{z}_{2u_1}, \bar{x}_{u_1})$ is assumed as a function of $\bar{z}_{2u_1}$ and $\bar{x}_{u_1}$ such that:

(i) The point $(\bar{z}_{2u_1}, \bar{x}_{u_1})$ assumes the value in a closed convex subset $\mathbb{R}^2$ of two dimensional real space containing the point $(\bar{Z}_2, \bar{X})$.

(ii) The function g_1 $(\bar{z}_{2u_1}, \bar{x}_{u_1})$ is continuous and bounded in $\mathbb{R}^2$.

(iii) $g_1(\bar{Z}_2, \bar{X}) = \bar{Z}_2$.

(iv) The first and second order partial derivatives of $g_1\,(\bar{z}_{2u_1},\,\bar{x}_{u_1})$ exist and are continuous and bounded in $\mathbb{R}^2$.

The similar regularity conditions holds for $g_2, g_3, g_4,\ h_1,\ h_2$, and h_3 respectively as that of g_1.

5. Properties of proposed class of IST estimator

Since the proposed IST estimator $\mathbb{T}$ has to be linear Unbiased Estimator, therefore, following Garcia and Artes (2002), we consider the following assumptions:

$$E(\hat{\bar{y}}^*_{2u}) = E(\hat{\bar{y}}^*_{2m}) \cong \bar{Y}_2, \tag{2}$$

$$E(\hat{\bar{y}}^*_{1u^*}) = E(\hat{\bar{y}}^*_{1m}) \cong \bar{Y}_1. \tag{3}$$

Now, using the results in equations (2) and (3) into the expression for the proposed estimator in equation (1), we have

$$E(\mathbb{T}) = (\zeta_1 + \zeta_2)\bar{Y}_1 + (\zeta_3 + \zeta_4)\bar{Y}_2. \tag{4}$$

In order to satisfy the assumption in equation (2), we have the following conditions:

$$\zeta_1 + \zeta_2 = 0 \text{ and } \zeta_3 + \zeta_4 = 1. \tag{5}$$

Now, using the conditions in equation (5), the final structure of unbiased IST estimator for estimating the sensitive population mean at current move is given as:

$$\mathbb{T} = \zeta_1(\hat{\bar{y}}^*_{1u^*} - \hat{\bar{y}}^*_{1m}) + \zeta_3\hat{\bar{y}}^*_{2m} + (1-\zeta_3)\hat{\bar{y}}^*_{2u}. \tag{6}$$

Following Mukhopadhyay (2014), as the estimators $\hat{\bar{y}}^*_{1u^*}$, and $\hat{\bar{y}}^*_{1m}$ are based on two independent samples u^* and m respectively, so $\mathrm{Cov}(\hat{\bar{y}}^*_{1u^*},\ \hat{\bar{y}}^*_{1m}) = 0$. Similarly

$$\mathrm{Cov}(\hat{\bar{y}}^*_{1u^*},\ \hat{\bar{y}}^*_{2m}) = \mathrm{Cov}(\hat{\bar{y}}^*_{1u^*},\ \hat{\bar{y}}^*_{2u}) = \mathrm{Cov}(\hat{\bar{y}}^*_{1m},\ \hat{\bar{y}}^*_{2u}) = \mathrm{Cov}(\hat{\bar{y}}^*_{2m},\ \hat{\bar{y}}^*_{2u}) = 0. \tag{7}$$

Also, $\hat{\bar{z}}^*_{1u^*}$ is based on *LL* sample and $\hat{t}^*_{u^*}$ is based on corresponding *SL* sample, therefore $\mathrm{Cov}(\hat{\bar{z}}^*_{1u^*},\ \hat{t}^*_{u^*}) = 0$. Similarly,

$$\mathrm{Cov}(\hat{\bar{z}}^*_{2u},\ \hat{t}^*_{u}) = \mathrm{Cov}(\hat{\bar{z}}^*_{1m},\ \hat{t}^*_{u}) = \mathrm{Cov}(\hat{\bar{z}}^*_{2m},\ \hat{t}^*_{u}) = 0.$$

Properties of the proposed IST estimator are discussed under above conditions, and the following assumptions:
$\bar{z}_{2u_1} = \bar{Z}_2\,(1+e_0),\ \bar{x}_{u_1} = \bar{X}\,(1+e_1),\ \bar{t}_{u_2} = \bar{T}\,(1+e_2),\ \bar{x}_{u_2} = \bar{X}\,(1+e_3),\ \bar{z}_{2m_1} = \bar{Z}_2\,(1+e_4),\ \bar{x}_{m_1} = \bar{X}\,(1+e_5),\ \bar{t}_{m_2} = \bar{T}\,(1+e_6),\ \bar{x}_{m_2} = \bar{X}\,(1+e_7),\ \bar{z}_{1m_1} = \bar{Z}_1\,(1+e_8),\ \bar{z}_{1u^*_1} = \bar{Z}_1\,(1+e_9),\ \ \bar{x}_{u^*_1} = \bar{X}\,(1+e_{10}),\ \ \bar{t}_{u^*_2} = \bar{T}\,(1+e_{11}),\ \ \bar{x}_{u^*_2} = \bar{X}\,(1+e_{12})$, such that, $E(e_i) = 0;\ |e_i| < 1$ where, $i = 0,\ 1,\ 2,\ 3,\ \ldots,\ 12$.

5.1. General and optimum allocations on two moves

As discussed in Section (2), under the general allocation design, the following allocation will be applicable to *LL* and *SL* samples on two moves:
$u_1^* = u_2^* = u_1 = u_2 = \frac{u}{2}$, $m_1 = m_2 = \frac{m}{2}$.

Following Perri *et al.* (2018) and applying the optimum allocation design to allocate *LL* and *SL* samples on two moves, the following assumptions will be applicable:
$u_1^* = u\frac{S_{z_1}}{S_{z_1}+S_t} = u\beta_1$ (say), $u_2^* = u(\frac{S_t}{S_{z_1}+S_t}) = u\beta_2$ (say), $m_1 = m(\frac{S_{z_2}}{S_{z_2}+S_t}) = m\beta_3$ (say), $m_2 = m(\frac{S_t}{S_{z_2}+S_t}) = m\beta_4$ (say), $u_1 = u\beta_3$ and $u_2 = u\beta_4$.

Hence, utilizing the two allocation designs, we further discuss the properties of the proposed IST estimator.

Theorem 5.1 *The variance of the estimator* $\mathbb{T}$ *under general allocation design as well as optimum allocation design is obtained and given as*

$$[V(\mathbb{T})]_i = \frac{1}{n}[(\zeta_1^i)^2(\{\frac{\lambda^i+\mu_f^i}{\lambda^i\mu_f^i}\}s_1^i) + (\zeta_3^i)^2(\{\frac{1}{\lambda^i}\}s_2^i) + (1-\zeta_3^i)^2(\{\frac{1}{\mu_f^i}\}s_2^i) - 2\zeta_1^i\zeta_3^i(\{\frac{1}{\lambda^i}\}s_3^i)],$$

where, $i = \begin{cases} g & \text{for general allocation design} \\ o & \text{for optimum allocation design} \end{cases}$,
$s_1^g = 2S_{z_1}^2 - 2\rho_{z_1x}^2S_{z_1}^2 + 2S_t^2 - 2\rho_{tx}^2S_t^2$, $s_2^g = 2S_{z_2}^2 - 2\rho_{z_2x}^2S_{z_2}^2 + 2S_t^2 - 2\rho_{tx}^2S_t^2$, $s_3^g = 2S_{z_1}S_{z_2}(\rho_{z_1z_2} - \rho_{z_1x}\rho_{z_2x}) + 2S_t^2(1-\rho_{tx}^2)$, $s_1^o = \frac{S_{z_1}^2-\rho_{z_1x}^2S_{z_1}^2}{\beta_1} + \frac{S_t^2-\rho_{tx}^2S_t^2}{\beta_2}$, $s_2^o = \frac{S_{z_2}^2-\rho_{z_2x}^2S_{z_2}^2}{\beta_3} + \frac{S_t^2-\rho_{tx}^2S_t^2}{\beta_4}$,
$s_3^o = \frac{S_{z_1}S_{z_2}(\rho_{z_1z_2}-\rho_{z_1x}\rho_{z_2x})}{\beta_3} + \frac{S_t^2(1-\rho_{tx}^2)}{\beta_4}$.

Proof 5.1 *The variance of* $\mathbb{T}$ *is given by*

$$\begin{aligned}[V(\mathbb{T})]_i =& (\zeta_1^i)^2[V(\hat{\bar{y}}_{1u^*}^*) + V(\hat{\bar{y}}_{1m}^*)]_i + (\zeta_3^i)^2V(\hat{\bar{y}}_{2m}^*)_i + (1-\zeta_3^i)^2V(\hat{\bar{y}}_{2u}^*)_i - \\ & 2\zeta_1^iCov(\hat{\bar{y}}_{1u^*}^*, \hat{\bar{y}}_{1m}^*) + 2\zeta_1^i\zeta_3^i[Cov(\hat{\bar{y}}_{1u^*}^*, \hat{\bar{y}}_{2m}^*) - Cov(\hat{\bar{y}}_{1m}^*, \hat{\bar{y}}_{2m}^*)]_i + \\ & 2\zeta_1^i(1-\zeta_3^i)[Cov(\hat{\bar{y}}_{1u^*}^*, \hat{\bar{y}}_{2u}^*) - Cov(\hat{\bar{y}}_{1m}^*, \hat{\bar{y}}_{2u}^*)] + 2\zeta_3^i(1-\zeta_3^i)Cov(\hat{\bar{y}}_{2m}^*, \hat{\bar{y}}_{2u}^*). \quad (8)\end{aligned}$$

Using equation (7), we have

$$\begin{aligned}[V(\mathbb{T})]_i =& (\zeta_1^i)^2[V(\hat{\bar{y}}_{1u^*}^*) + V(\hat{\bar{y}}_{1m}^*)]_i + (\zeta_3^i)^2V(\hat{\bar{y}}_{2m}^*)_i + (1-\zeta_3^i)^2V(\hat{\bar{y}}_{2u}^*)_i - \\ & 2\zeta_3^i\zeta_1^i[Cov(\hat{\bar{y}}_{1m}^*, \hat{\bar{y}}_{2m}^*)]_i, \quad (9)\end{aligned}$$

where for large population size, the variance of $\hat{\bar{y}}_{1u^*}^*$ *is computed below*

$$V(\hat{\bar{y}}_{1u^*}^*) = V(\hat{\bar{z}}_{1u^*}^*) + V(\hat{\bar{t}}_{u^*}^*) - Cov(\hat{\bar{z}}_{1u^*}^*, \hat{\bar{t}}_{u^*}^*). \quad (10)$$

For this expanding $\hat{\bar{z}}_{1u^*}^*$ *about the point* $G = (\bar{Z}_1, \bar{X})$ *using Taylor's series expansion,*

retaining terms up to the first order approximations, we have

$$\begin{aligned}\hat{\bar{z}}^{*}_{1u^{*}} =& g_4[\bar{Z}_1 + (\bar{z}_{1u^{*}_1} - \bar{Z}_1),\ \bar{X} + (\bar{x}_{u^{*}_1} - \bar{X})] \\ =& \bar{Z}_1 + (\bar{z}_{1u^{*}_1} - \bar{Z}_1)G_1 + (\bar{x}_{u^{*}_1} - \bar{X})G_2 + [(\bar{z}_{1u^{*}_1} - \bar{Z}_1)^2 G_{11} + (\bar{x}_{u^{*}_1} - \bar{X})^2 G_{22} \\ &+ 2(\bar{z}_{1u^{*}_1} - \bar{Z}_1)(\bar{x}_{u^{*}_1} - \bar{X})G_{12} + \ldots].\end{aligned}$$

Expressing above equation in terms of $e_i^{'}s$ and retaining terms up to the first order approximations we have

$$\hat{\bar{z}}^{*}_{1u^{*}_1} - \bar{Z}_1 = (\bar{Z}_1 e_9 G_1 + \bar{X} e_{10} G_2 + [\bar{Z}_1^2 e_9^2 G_{11} + \bar{X}^2 e_{10}^2 G_{22} + \bar{Z}_1 \bar{X} e_9 e_{10} G_{12}]), \tag{11}$$

where,

$$G_1 = \frac{\partial g_4}{\partial \bar{z}_{1u^{*}_1}}|_G = 1,\ G_2 = \frac{\partial g_4}{\partial \bar{x}_{u^{*}_1}}|_G,\ G_{11} = \frac{1}{2}\frac{\partial^2 g_4}{\partial \bar{z}^2_{1u^{*}_1}}|_G = 0,\ G_{22} = \frac{1}{2}\frac{\partial^2 g_4}{\partial \bar{x}^2_{u^{*}_1}}|_G,$$

and $G_{12} = \frac{1}{2}\frac{\partial^2 g_4}{\partial \bar{z}_{1u^{*}_1} \partial \bar{x}_{u^{*}_1}}|_G.$

Squaring both sides of equation (11) and further retaining terms up to the first order approximation, we have

$$\left(\hat{\bar{z}}^{*}_{1u^{*}_1} - \bar{Z}_1\right)^2 = (\bar{Z}_1^2 e_9^2 + \bar{X}^2 e_{10}^2 G_2^2 + 2\bar{Z}_1 \bar{X} e_9 e_{10} G_2). \tag{12}$$

Taking expectations on both sides of the above equation, the variance of $\hat{\bar{z}}^{*}_{1u^{*}_1}$ is obtained as

$$V(\hat{\bar{z}}^{*}_{1u^{*}_1}) = \frac{1}{u_1}\left[S^2_{z_1} - S^2_{z_1}\rho^2_{z_1 x}\right],$$

similarly,

$$V(\hat{\bar{t}}^{*}_{u^{*}_2}) = \frac{1}{u_2}\left[S^2_t - S^2_t \rho^2_{tx}\right].$$

Following similar procedure for $V(\hat{\bar{y}}^{*}_{1m})_i$, $V(\hat{\bar{y}}^{*}_{2m})_i$, $V(\hat{\bar{y}}^{*}_{2u})_i$ and $Cov(\hat{\bar{y}}^{*}_{1m}, \hat{\bar{y}}^{*}_{2m})_i$ and substituting in equation (9), we have the expression of the variance of $[\mathbb{T}]_i$ under general allocation design and optimum allocation design as described in Theorem 5.1.

6. Constants under IST Allocation Designs

It is observed that $[V(\mathbb{T})]_i$ is a function of unknown constants ζ_1^i and ζ_3^i. Hence, they are minimized with respect to ζ_1^i and ζ_3^i respectively to obtain the optimum value of constants. The optimum values obtained are given in Table 3.

Table 3: Optimum Value of Constants

General Allocation Design	Optimum Allocation Design
$\zeta_1^g = \frac{s_2^g s_3^g \lambda^g \mu_f^g}{s_1^g s_2^g - (s_3^g)^2 (\mu_f^g)^2}$,	$\zeta_1^o = \frac{s_2^o s_3^o \lambda^o \mu_f^o}{s_1^o s_2^o - (s_3^o)^2 (\mu_f^o)^2}$,
$\zeta_3^g = \frac{s_1^g s_2^g \lambda^g}{s_1^g s_2^g - (s_3^g)^2 (\mu_f^g)^2}$	$\zeta_3^o = \frac{s_1^o s_2^o \lambda^o}{s_1^o s_2^o - (s_3^o)^2 (\mu_f^o)^2}$

Substituting the above optimum values of ζ_1^i and ζ_3^i in the expression of $[V(\mathbb{T})]_i$ respectively, we get the minimum variance of the proposed IST estimator as presented in Table 4.

Table 4: Optimum Variance

General Allocation Design	Optimum Allocation Design
$[V(\mathbb{T})_{opt.}]_g = \left(\frac{1}{n}\right)\left[\frac{s_2^g (s_1^g s_2^g - (s_3^g)^2 \mu_f^g)}{s_1^g s_2^g - (s_3^g)^2 (\mu_f^g)^2}\right]$	$[V(\mathbb{T})_{opt.}]_o = \left(\frac{1}{n}\right)\left[\frac{s_2^o (s_1^o s_2^o - (s_3^o)^2 \mu_f^o)}{s_1^o s_2^o - (s_3^o)^2 (\mu_f^o)^2}\right]$

6.1. Optimum Replacement policy and Minimum Variance

In surveys repeated over time, the objective is to obtain efficient estimates with minimum cost of the survey. This is technically achieved by maintaining a high overlap between two successive moves. However, the best strategy would be to minimize the variance of the estimator in order to determine the optimum value of μ or λ. Hence, $[\mathbb{V}(\mathbb{T})_{opt.}]_i$ is further minimized with respect to μ_f^i respectively, and the obtained optimum values of μ_f^i say $\hat{\mu}_f^i$ are as:

$$\hat{\mu}_f^i = min\left\{\frac{I_2^i + \sqrt{(I_2^i)^2 - I_1^i I_3^i}}{I_1^i},\ \frac{I_2^i - \sqrt{(I_2^i)^2 - I_1^i I_3^i}}{I_1^i}\right\}\ \varepsilon\ [0,\ 1], \tag{13}$$

where,

$$i = \begin{cases} g & \text{for general allocation design} \\ o & \text{for optimum allocation design} \end{cases},$$

$I_1^g = (s_3^g)^4 s_2^g$, $I_2^g = (s_3^g)^2 s_1^g (s_2^g)^2$, $I_3^g = s_1^g (s_2^g)^2 (s_3^g)^2$, $I_1^o = (s_3^o)^4 s_2^o$, $I_2^o = (s_3^o)^2 s_1^o (s_2^o)^2$ and $I_3^o = s_1^o (s_2^o)^2 (s_3^o)^2$.

Substituting the optimum values of $\hat{\mu}^i$ in $[V(\mathbb{T})_{opt.}]_i$, we have the minimum variance of the proposed IST estimator as presented in Table 5.

Table 5: Optimum Variance in terms of Optimum μ

Estimator	General Allocation Design	Optimum Allocation Design
$\mathbb{T}$	$[V(\mathbb{T})_{opt.^*}]_g = (\frac{1}{n})\left[\frac{s_2^g(s_1^g s_2^g-(s_3^g)^2\hat{\mu}_f^g)}{s_1^g s_2^g-(s_3^g)^2(\hat{\mu}_f^g)^2}\right]$	$[V(\mathbb{T})_{opt.^*}]_o = (\frac{1}{n})\left[\frac{s_2^o(s_1^o s_2^o-(s_3^o)^2\hat{\mu}_f^o)}{s_1^o s_2^o-(s_3^o)^2(\hat{\mu}_f^o)^2}\right]$

7. Comparison

To judge the efficiency of the proposed class of IST estimators $\mathbb{T}$, the IST estimator τ has been considered where no additional auxiliary is used at any move, which is given as

$$\tau = \kappa_1\hat{\bar{y}}_{1u^*} + \kappa_2\hat{\bar{y}}_{1m} + \kappa_3\hat{\bar{y}}_{2m} + \kappa_4\hat{\bar{y}}_{2u}, \tag{14}$$

where, κ_j ; $j = 1,\ 2,\ 3$ and 4 are suitably chosen constants.

The minimum variance of the IST estimator τ has been computed and is given as

$$[V(\tau)_{opt.}]_i = \left(\frac{1}{n}\right)\left[\frac{\nu_2^i(\nu_1^i\nu_2^i-(\nu_3^i)^2\mu_1^i)}{\nu_1^i\nu_2^i-(\nu_3^i)^2(\mu_1^i)^2}\right] \tag{15}$$

with

$$\hat{\mu}_1^i = min\left\{\frac{I_{12}^i+\sqrt{(I_{12}^i)^2-I_{11}^iI_{13}^i}}{I_{11}^i},\ \frac{I_{12}^i-\sqrt{(I_{12}^i)^2-I_{11}^iI_{13}^i}}{I_{11}^i}\right\}\ \varepsilon\ [0,\ 1], \tag{16}$$

where, $I_{11}^i = (\nu_3^i)^4\nu_2^i$, $I_{12}^i = (\nu_3^i)^2\nu_1^i(\nu_2^i)^2$, $I_{13}^i = \nu_1^i(\nu_2^i)^2(\nu_3^i)^2$, $\nu_1^g = 2S_{z_1}^2 + 2S_t^2$,
$\nu_2^g = 2S_{z_2}^2 + 2S_t^2$, $\nu_3^g = 2\rho_{z_1z_2}S_{z_1}S_{z_2} + 2S_t^2$, $\nu_1^o = \frac{S_{z_1}^2}{\beta_1} + \frac{S_t^2}{\beta_2}$, $\nu_2^o = \frac{S_{z_2}^2}{\beta_3} + \frac{S_t^2}{\beta_4}$,
$\nu_3^o = \frac{\rho_{z_1z_2}S_{z_1}S_{z_2}}{\beta_3} + \frac{S_t^2}{\beta_4}$.

8. Performance of IST estimator

In this section, we check the percent relative efficiency of the IST class of estimators $\mathbb{T}$ with respect to the IST estimator τ which are the linear combination of the estimators based on all available samples considering the availability and non-availability of additional non-sensitive auxiliary variable respectively. The percent relative efficiency has been computed under both the general allocation design as well as optimum allocation design as under:

$$\mathbb{E}^i = \frac{[V(\tau)_{opt.^*}]_i}{[V(\mathbb{T})_{opt.^*}]_i} \times 100, \tag{17}$$

where, $i = \begin{cases} g & for\ general\ allocation\ design \\ o & for\ optimum\ allocation\ design \end{cases}$.

8.1. Simulation Study

To validate the theoretical results, simulation studies have been carried out using Monte Carlo Simulation by MATLAB. The simulation is performed by examining $5,000$ different samples at two moves and the process is repeated for varying sample sizes.

Population Source:[Free access to data by Statistical Abstracts of United States]
A real population consisting of $N = 51$ states has been considered for evaluation of the performance of proposed estimators. The variables considered under IST set-up for two moves are assumed as:

y_1 :Rate of abortions in the year 2005
y_2 :Rate of abortions in the year 2008
t :Rate of residents in the year 2004
x :Rate of residents in the year 2000.

From the above considered variables, it is obvious that the rate of abortions is sensitive in nature however the rate of residents is non-sensitive in nature. Therefore, the data are suitable to be used to test the performance of the proposed IST estimators.

The simulated percent relative efficiencies of τ with respect to $\mathbb{T}$ have been computed under general as well as optimum allocation design denoted by $\mathbb{E}^{si}$; $i \in \{g, o\}$.

The simulation results are presented in Table 6 and Figure 1.

Table 6: Simulation Results

	u/n	$\mathbb{E}^{sg}$	$\mathbb{E}^{so}$
$n = 24, u = 3$	0.125	688.4432	716.2598
$n = 24, u = 5$	0.208	643.9376	696.7527
$n = 24, u = 7$	0.291	605.1396	681.0255
$n = 24, u = 10$	0.416	556.1936	664.5010
$n = 24, u = 12$	0.5	527.6227	657.7056
$n = 24, u = 15$	0.625	493.9802	656.8039
$n = 24, u = 17$	0.708	476.4308	662.1325
$n = 24, u = 20$	0.833	461.5208	683.8706
$n = 24, u = 22$	0.916	475.8523	710.6313

8.2. Numerical Illustration

To judge the performance of the proposed estimator, numerical illustration has been done for the real data considered in Section (8.1). The percent relative efficiency of the proposed estimator has been computed under general as well as optimum allocation designs denoted by E^i; $i \in \{g, o\}$.

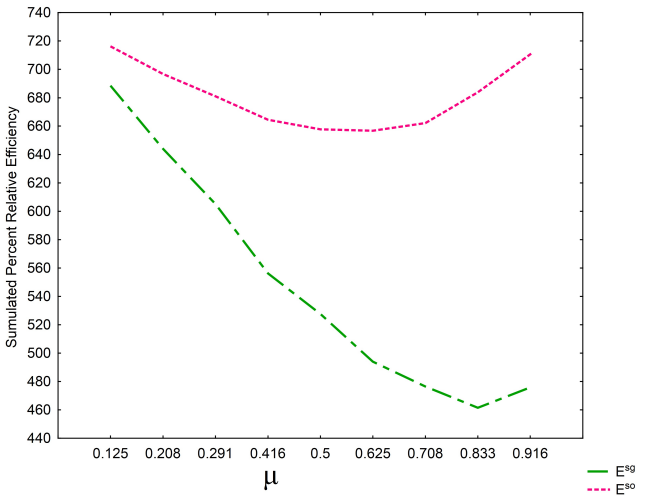


Figure 1: Graphical Representation of Simulation Results

Therefore, Table 7 represents the results obtained on performing the empirical calculation on the considered data in Section (8.1).

Table 7: Optimum value of $\mu^{'}s$ and Percent relative efficiencies

$\hat{\mu}_1^g$	$\hat{\mu}_1^o$	$\hat{\mu}_f^g$	$\hat{\mu}_f^o$	E^g	E^o
0.8461	0.8584	0.6585	0.6712	656.6323	726.9219

9. Direct Method

The direct method of estimation is compared with the IST embedded method in order to observe the amount of loss in the precision of estimators that result due to application of IST. Some loss in precision is expected but application of direct method may not represent the true facts as the variable under consideration is sensitive in nature. As a result, privacy protection becomes an important issue for which the respondents need to be convinced. The direct version of the class of estimators $\mathbb{T}$ denoted by $\mathbb{T}_d$ is discussed as:

$$\mathbb{T}_d = \zeta_{d1}\hat{\bar{y}}^*_{1du^*} + \zeta_{d2}\hat{\bar{y}}^*_{1dm} + \zeta_{d3}\hat{\bar{y}}^*_{2dm} + \zeta_{d4}\hat{\bar{y}}^*_{2du}, \tag{18}$$

where, the constants ζ_{dj} ; $dj = 1,\ 2,\ 3$ and 4 are to be suitably chosen. Now, for computing the variance, we have the following steps as

$$E(\mathbb{T}_d) = (\zeta_{d1} + \zeta_{d2})\bar{Y}_1 + (\zeta_{d3} + \zeta_{d4})\bar{Y}_2 \tag{19}$$

with

$$\zeta_{d1} + \zeta_{d2} = 0 \text{ and } \zeta_{d3} + \zeta_{d4} = 1.$$

Following similar conditions as in Section (4), the optimum variance of the direct method is obtained and is given as

$$V(\mathbb{T}_d)_{opt.^*} = \left(\frac{1}{n}\right)\left[\frac{C_2(C_1C_2 - C_3^2\hat{\mu}_d)}{C_1C_2 - C_3^2(\hat{\mu}_d)^2}\right] \tag{20}$$

with

$$\hat{\mu}_d = min\left\{\frac{F_2 + \sqrt{F_2^2 - F_1F_3}}{F_1},\ \frac{F_2 - \sqrt{F_2^2 - F_1F_3}}{F_1}\right\}\ \varepsilon\ [0,\ 1] \tag{21}$$

$$\text{where, } F_1 = C_3^4C_2,\ F_2 = C_3^2C_1C_2^2,\ F_3 = C_1C_2^2C_3^2,\ C_1 = S_{y_1}^2 - \rho_{y_1x}^2S_{y_1}^2,$$
$$C_2 = S_{y_2}^2 - \rho_{y_2x}^2S_{y_2}^2,\ C_3 = S_{y_1}S_{y_2}(\rho_{y_1y_2} - \rho_{y_1x}\rho_{y_2x}).$$

To examine the performance of the direct method we compare the estimator $\mathbb{T}_d$ with respect to the proposed IST class of estimator $\mathbb{T}$ under both general allocation design as well as optimum allocation design as

$$E_d^i = \frac{V(\mathbb{T}_d)_{opt.^*}}{[V(\mathbb{T})_{opt.^*}]_i} \times\ 100. \tag{22}$$

The numerical comparison has been done on the data considered in Section (8.1) and the results are presented in Table 8.

Table 8: Direct Method comparison with IST under general allocation design and optimum allocation design

$\hat{\mu}_d$	E_d^g	E_d^o
0.6000	33.4446	41.9547

10. Discussion of Results

The following interpretations can be drawn from empirical and simulation results:

1. The minimum variance unbiased estimation is feasible under IST set-up to estimate sensitive population mean on successive moves.

2. The simulation results in Table 6 and Figure 1 show that for all $\mu \in [0, 1]$, the percent relative efficiencies exist for both the allocation designs. As $E^{so} > E^{sg} \; \forall \; \mu$, this indicate that the IST estimator under optimum allocation design is more efficient than the general allocation design. The percent relative efficiencies exit for all considered variations in sample sizes. Also, both E^{sg} and $E^{so} > 0$, this indicates that the IST class of estimators $\mathbb{T}$ is better than that of IST estimator τ.

3. From Table 7, it is observed that the optimum fraction of fresh sample to be drawn afresh at current occasion exists for both the IST class of estimators under both allocation designs. Further, it is observed that IST estimator $\mathbb{T}$ is coming out to be more efficient than IST estimator τ under both the allocation designs. However, the estimators under the optimum allocation design are proved to be more efficient than that of the general allocation design.

4. Table 8 indicates that $E_d^i < 100 \; \forall \; i \in \{g, \; o\}$, which means there is a loss in precision when the IST estimator is compared with the direct method under both the allocation designs. The loss is incurred due to the usage of IST in the estimator $\mathbb{T}$. However, dealing with sensitive issues when the direct method is used, may result in false response or even no response.

11. Concluding Remarks

From the interpretation of results, it is concluded that IST is an alternative technique to deal with sensitive issues in successive sampling. In IST setup, the estimator utilizing additional auxiliary variable is proved to be more efficient than the estimator in which no additional auxiliary variable is used. Out of the two allocation designs for allocating *LL* and *SL* samples, the IST class of estimators using optimum allocation design is coming out to be more efficient than the estimator using general allocation deign. Therefore, the IST estimators with optimum allocation designs may be recommended for their practical use by survey practitioners.

Acknowledgements

Authors are thankful to the honourable reviewer for deeply reading the manuscript and giving valuable suggestions leading to improvement in the quality of content and presentation over the original version of the manuscript. The first authors is thankful to Science and Engineering Research Board, New Delhi, India for providing the financial assistance to carry out the present work. Authors sincerely acknowledged the free access to data by statistical abstracts of United states.

References

Arcos, A., Rueda, M. and Singh, S. A., (2015). Generalized approach to randomized response for quantitative variables. *Quality and Quantity*, 49, pp. 1239 –1256.

Arnab R., Singh S., (2013). Estimation of mean of sensitive characteristics for successive sampling. *Communication in Statistics-Theory and Methods*, 42, 2499–2524.

Chaudhuri,A., Christofides,T. C., (2013). Indirect questioning in sample surveys. *Springer-Verlag, Berlin, Heidelberg, De.*

Diana, G., Perri, P. F., (2010). New scrambled response models for estimating the mean of a sensitive quantitative character. *Journal of Applied Statistics*, 37(11), pp. 1875–1890.

Diana, G., Perri, P. F., (2011). A class of estimators for quantitative sensitive data. *Statistical Papers,* 52(3), pp. 633–650.

Eichhorn, B. H., Hayre, L. S., (1983). Scrambled randomized response method for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7(4), pp. 307–316.

Franklin, L. A., (1989). A comparison of estimators for randomized response sampling with continuous distribution from dichotomous populations. *Communications in Statistics-Theory and Methods*, 18, pp. 489–505.

Garcia, A. V., Artes, E. M., (2002). Improvement on estimating of current population ratio in successive sampling. *Brazilian Journal of Probability and Statistics*, 16(2), pp. 107–122.

Greenberg, B. G., Kubler, R. R., Horvitz, D. G., (1971). Application of the randomized technique in obtaining quantitative data. *Journal of the American Statistical Association*, 66, pp. 243–250.

Holbrook, A. L., Krosnick, J. A., (2010). Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique. *Public Opinion Quarterly*, 74, pp. 37–67.

Horvitz, D. G., Shah, B. V., Simmons, W. R., (1967). The unrelated question randomized response model. *Journal of the American Statistical Association*, pp. 65–72.

LaBrie, J. W., Earleywine, M., (2000). Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-count Technique. *Journal of Sex Research*, 37, pp. 321–326.

Miller, J. D., (1984). A New Survey Technique For Studying Deviant Behavior.The George Washington University Ph.D. thesis.

Mukhopadhyay, P., (2014). Theory and Methods of Survey Sampling. *PHI Learning Private Limited*, Delhi.

Patterson, H. D., (1950). Sampling on successive occasions with partial replacement units. *Journal of royal statistical society, Wiley, series B(Mthodoligical)* 12(2), pp. 241–255.

Perri, P. F., Diana, G., (2013). Scrambled response models Based on auxiliary variables. In: Torelli, N., Pesarin, F., Bar-Hen, A. (eds.). *Advances in Theoretical and Applied Statistics, Springer, Berlin*, pp. 281–291.

Perri, P. F., Rueda, M. M. G., Cobo, B. R., (2018). Multiple sensitive estimation and optimal sample size allocation in the item sum technique. *Biometrical Journal*, 60, pp. 155–173.

Pollock, K. H., Bek, Y., (1976). A comparison of three randomized response models for quantitative data. *Journal of American Statistical Association*, 71, pp. 884–886.

Priyanka, K., Trisandhya, P. and Mittal, R., (2018). Dealing sensitive characters on successive occasions through a general class of estimators using scrambled response techniques. *Metron*, 76, pp. 203–230.

Priyanka, K., Trisandhya, P., (2019). A Composite Class of Estimators using Scrambled Response Mechanism for Sensitive Population mean in Successive Sampling. *Communications in statistics- Theory and Methods*, 48(4), pp. 1009–1032.

Priyanka, K., Trisandhya, P. and Mittal, R., (2019). Scrambled Response Techniques in Two Wave Rotation Sampling for Estimating Population Mean of Sensitive Characteristics with Case Study. *Journal of Indian Society of Agricultural Statistics*, 73(1), pp. 41–52.

Rayburn, N. R., M. Earleywine, and Davison, G. C., (2003). Base Rates of Hate Crime Victimization among College Students. *Journal of Interpersonal Violence*, 18, pp. 1209–1221.

Srivastava, S. K., Jhajj H. S., (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankya series C,* 42, pp. 87–96.

Tracy, D. S., Singh, H. P. and Singh, R., (1996). An alternative to the ratio-cum-product estimator in sample surveys. *Journal of Statistical Planning and Inference*, 53, pp. 375–397.

Trappmann, M., Krumpal, I., Kirchner, A., and Jann, B., (2014). Item sum: A new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology*, 2, pp. 58–77.

Tsuchiya, T., Hirai, Y. and Ono, S., (2007). A Study of the Properties of the Item Count Technique. *Public Opinion Quarterly*, 71, pp. 253–272.

Warner, S. L., (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, pp. 63–69.

Wimbush, J. C., Dalton, D. R., (1997). Base Rate for Employee Theft: Convergence of Multiple Methods. *Journal of Applied Psychology*, 82, pp. 756–763.

Yu, B., Jin, Z., Tian, J. and Gao, G., (2015). Estimation of sensitive proportion by randomized response data in successive sampling. *Computational and Mathematical Methods in Medicine*, pp 6, DOI: 10.1155/2015/172918.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 139–151, https://doi.org/10.59170/stattrans-2023-056
Received – 28.11.2022; accepted – 12.05.2023

Analysis of social and economic conditions of microenterprises based on taxonomy methods

Aneta Ptak-Chmielewska[1], **Agnieszka Chłoń-Domińczak**[2]

Abstract

The situation of microenterprises on the market is difficult as they are faced with barriers and competitiveness imposed by larger units. We used data from the REGIOGMINA project to assess the situation of microenterprises. The REGIOGMINA project was implemented by a consortium of Kujawsko–Pomorskie Voivodship, the SGH Warsaw School of Economics and the Nicolaus Copernicus University in Toruń in the scope of the GOSPOSTRATEG initiative. This data set was complemented by data provided by the Local Data Bank of Statistics Poland. The analysis describes the situation in Kujawsko-Pomorskie Voivodship. The information from both sources illustrates the situation of microenterprises at a local level (gmina) in the years 2019–2020. A cluster analysis based on taxonomy methods was performed. The aim of the research is to expand the knowledge and contribute to a better understanding of the social and economic problems that microenterprises are confronted with at a local level. The study covers the period before the COVID-19 pandemic and the first year following its outbreak, which makes it possible to assess the effects that the measures taken against the pandemic had on the situation of microenterprises at a local level.

Key words: microenterprises, taxonomy methods, COVID-19.

1. Introduction

The COVID-19 pandemic brought unprecedented challenges for SMEs around the world, including disruption of their activities and financial situation, as well as other difficulties. SMEs faced demand disruptions related to lockdowns, but also logistical challenges related to the break of global value chains. The severity of the challenges depended also on the type of industry in which SMEs operated. The support provided to SMEs under the COVID Anti-Crisis shield provided some support to these

[1] Warsaw School of Economics, Poland. E-mail: aptak@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-9896-4240.

[2] Warsaw School of Economics, Poland. E-mail: achlon@sgh.waw.pl. ORCID: https://orcid.org/0000-0001-5068-0276.

companies. However, their capacity to weather the crisis, frequently with limited resources, differed from large companies.

During the COVID-19 pandemic, Polish enterprises could benefit from various solutions under the "Anti-crisis Shield", comprising six stages of support, including one directed towards the tourist sector and one focused on selected sectors that were more affected by the lockdown measures. The majority of Polish enterprises applied for this support.

The COVID-19 pandemic was followed by the war in Ukraine and the energy crisis that has further affected the situation of SMEs. Therefore, it is important to provide continuous monitoring of the situation of SMEs not only at the national or regional level, but also at the local level, given the very diverse socioeconomic conditions depending on the locality of enterprises.

In our paper, we analyse the performance of micro-enterprises at the local level in the Kujawsko-Pomorskie region, using the unique set of administrative data collected from the Social Insurance Institution and tax authorities. We assess to what extent characteristics of microenterprises at the local level changed and which factors are the main drivers of the observed heterogeneity in the first two years of the COVID-19 pandemic and resulting lockdown measures. Furthermore, we provide evidence on the potential of the administrative data in monitoring the situation of enterprises. The use of administrative data enables timely and cost-effective collection of the information related to the situation of enterprises and how the enterprises changed over time. The main goal of this paper is to provide an example of the potential design of the monitoring tools that support analysis of the situation of SMEs in the medium and long term that can support evidence-based policy design focused on the development of this sector during and after the observed shocks and crises.

2. Literature review

Small and medium-sized enterprises are the foundation of the European economy. In 2020, slightly more than 21 million micro, small and medium-sized SMEs were active in the EU- 27, accounting for 99.8% of all enterprises in the EU-27 non-financial business sector (NFBS). Of this total, 93% were micro-SMEs. SMEs were generating 53% of the total value added produced by the EU- 27 NFBS, and employed 65% of workers in the NFBS (European Commission, 2021: 1). The pandemic had a major impact on EU-27 SMEs in 2020. Many SMEs faced financial difficulties, mainly resulting from large declines in sales. Other key challenges faced by many SMEs in 2020 included supply disruptions, an upsurge in late payments and operating at a loss. SMEs implemented a wide range of mitigation measures. In particular, SMEs used the different support programmes implemented by national governments, especially to pay

their wages, overcome cash flow issues, and reduce working hours and/or staffing (European Commission, 2021).

The pandemic shock for SMEs was of unprecedented magnitude. As summarised by Juergensen et al. (2020), during the peak of the COVID-19 pandemic, 41% of UK SMEs ceased operations, half of SMEs in Germany expected decline of revenues exceeding 10%, and 70% of SMEs in Italy were directly affected by the crisis. One of the major challenges faced by the sector during the pandemic was financial challenge (Cepel et al., 2020; Juergensen et al., 2020; Zutshi et al., 2021). The liquidity issues and financial constraints observed in the short term translated also into medium- and long-term responses, related, for instance, to upgrading digital infrastructure to enable online sales, and reorganisation of global value chains, which will also require adjustments of SMEs involved in the GVCs (Juergensen et al., 2020). The COVID pandemic also shifted the perception of business risk. Evidence from Czechia and Slovakia shows that while before COVID entrepreneurs perceived personnel risk as the most important, after the pandemic market risk and financial risk was faced by more than a half of SMEs (Cepel et al., 2020).

The role of the SME sector is also important in Poland. Despite the COVID pandemic, the number of microenterprises in 2020 increased by 2.1%, while this was the lowest rate of growth noted in recent years (Statistics Poland, 2021). SMEs benefited from the government Anti-Crisis Shield support. According to the evaluation of the Polish Economy Institute (Dębkowska et al., 2021), one of the main instruments used by SMEs was Anti-Crisis Shield 1.0. In connection with this support, microenterprises received PLN 18.9 billion in support. Regarding the support coordinated by Bank Gospodarstwa Krajowego (BGK), including the de-minimis guarantee, 60% of all beneficiaries were microenterprises. They also received support from the Labour Fund, in the form of low-interest loans (1.9 million claims for PLN 9.3 billion). Last but not least, microentrepreneurs were able to use the levy from social security contributions. In total, 70% of microenterprises used this levy, while 61% claimed stand-by benefit, and a similar percentage used support under the Anti-Crisis Shield.

The broad support received during the COVID-19 pandemic helped preserve the SME sector in 2020. This also applies to the Kujawsko-Pomorskie region. In 2020, there were more than 113 thousand microenterprises in this region, representing 55 companies per 1000 population (which is below the average in Poland of 59). The number of people employed in this sector per 1000 population was 103, compared to 113 in Poland, which is also below the average. Microenterprises in this region generated 3.6% of total revenues in 2020 and 4.0% of wages. The Kujawsko-Pomorskie region also noted one of the highest gross turnover profitability indicators of microenterprises (20.1% compared to 14% in Poland) (Statistics Poland, 2021).

3. Data and research methods

To assess the situation of microenterprises, we used two types of data sources using administrative data. The first one is data from the Social Insurance Institution administrative register. It includes information on enterprises – payers of social security contributions collected in the Kujawsko-Pomorskie region. The second is data from the tax authorities on the revenues and taxes of enterprises that were paid in the same region. This unique database was collected as part of the project REGIOGMINA. The project is implemented by a consortium led by the regional government of Kujawsko-Pomorskie Voivodeship with the SGH Warsaw School of Economics and Nicolaus Copernicus University in Toruń, financed by the National Centre for Research and Development. We also used context data from the Statistics Poland Local Data Bank.

The database compiled from both sources provides information about the situation of microenterprises on the gmina (municipality) level. The analysis covers the Kujawsko-Pomorskie region, which was involved in this innovative project, which, for instance, developed the proposal of the Regional Entrepreneurship Observatory, to monitor the situation of SMEs using administrative data. We focused on data from the years 2019 and 2020. 2019 was the reference year before the Covid-19 pandemic with limited observed turbulence. 2020 was the first year of full lock-down caused by the Covid-19 pandemic. A comparison of changes between those two years in the situation of microenterprises defines the impact of the pandemic on their situation and general economic situation on a local (gmina) level.

In the dataset, we identify the size of the companies based on the number of workers covered by social security contributions. Thus, we treat companies that employ nine or less workers, for whom they paid social security contributions, as microenterprises.

Variables on gmina level that were used in the analysis include average revenues from microenterprises based on tax information, the number of microenterprises per 10 thousand population in the gmina, type of gmina (urban, rural, rural-urban), gmina revenues per capita, gmina debt per capita (information from Ministry of Finance), share of population of working age, and the unemployment rate.

Table 1: Variables used in the analysis

Variable (label)	Description	Source
rev_pc_micro	average revenues - microenterprises (taxes) per capita in million PLN	Tax Office
micro_per10k	number of microenterprises per 10 000 population	Social Insurance Institution

Table 1: Variables used in the analysis (cont.)

Variable (label)	Description	Source
type_region	type of local region (gmina)	Statistics Poland
pop_prod_share	share of population of working age in percent	Statistics Poland
unemp_prod_share	unemployment rate in percent	Statistics Poland
rev_pc	local region revenues per capita in PLN	Statistics Poland
debt_pc	Debt per capita in ths PLN	Ministry of Finance

Source: own work.

For grouping of variables and revealing hidden factors the Factor Analysis method was applied. In this method principal component calculation method was used and varimax rotation. According to the principal component method, it is correlation matrix used its eigenvalues with eigen-vectors to calculate the coefficients for linear combination of variables. Linear combination of the variables and the coefficients values provides the information on hidden factors. The coefficients provide the input from each variable into the factor with a sign informing on the correlation between variable and the factor itself. This information provides the interpretation of hidden factors (Panek and Zwierzchowski, 2013).

For clustering, the k-means method was used. The k-means method is the most frequently used method. Euclidean distance is used as the default distance measure. The number of clusters is determined for the start, and next cluster seeds are chosen at random. Each observation (i=1,..., n) is classified in the group with the nearest cluster seed measured by Euclidean distance. For all clusters (j=1,..., k) the new cluster centres are calculated as the arithmetic mean of all observations belonging to the group. Those steps are repeated until there are no other moves between groups. The error function is calculated at each step – the sum of the quadratic distance intergroup calculated from group centres (Frątczak ed., 2009):

$$\mathrm{F} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{d}\left(O_i, M_j\right) \tag{1}$$

where d is Euclidean distance, O_i – is centre for i-th group, M_j – observation j=1,…k.

In practice, this process is convergent after a few iterations, but in general, as this algorithm does not have to be convergent, the maximum number of iterations is pre-defined.

4. Empirical results

An empirical analysis was conducted in three steps. In the first step, the factor analysis made it possible to identify the hidden factors and to verify their impact on

gminas. Next, the gminas were grouped into clusters, and finally in the last step the profile of each cluster was specified and described. Profiles were compared between clusters and between two periods, 2019 and 2020, before and the pandemic periods. Descriptive statistics for variables are presented in Table 2.

Table 2: Descriptive statistics for variables

a) 2019

Variable	Mean	Std. dev.	Min	Max	Q1	Median	Q3
expenses_pc	5358.29	676.89	4054.68	7765.99	4872.14	5254.40	5721.99
pop_prod_share	61.70	1.43	53.87	65.39	61.15	61.77	62.51
unemp_prod_share	3.89	1.39	1.19	8.32	2.93	3.88	4.68
rev_pc	5331.42	534.78	4183.56	7701.28	4961.73	5245.48	5628.84
debt_pc	1338.57	1190.64	0.00	7006.69	566.94	1140.48	1657.17
micro_per10ths	149.08	215.72	0.00	2232.48	95.28	124.33	159.02
rev_pc_micro (mln)	181.40	441.64	6.37	3899.69	48.71	80.74	145.65

b) 2020

Variable	Mean	Std. dev.	Min	Max	Q1	Median	Q3
expenses_pc	5643.41	659.03	4650.36	8568.85	5157.25	5528.00	6064.12
pop_prod_share	61.30	1.49	53.10	64.99	60.67	61.31	62.21
unemp_prod_share	4.34	1.43	1.44	8.46	3.45	4.34	5.26
rev_pc	5881.24	570.61	4764.34	7794.14	5416.90	5780.42	6260.01
debt_pc	5643.41	659.03	4650.36	8568.85	5157.25	5528.00	6064.12
micro_per10ths	96.05	126.70	0.00	1307.86	0.00	107.17	136.15
rev_pc_micro (mln)	172.68	346.22	10.94	3046.78	53.51	82.77	147.75

Source: own calculations in SAS 9.4.

4.1. Factor analysis

Factor analysis is performed using principal components with varimax orthogonal rotation. Factor analysis can be based on the principal component method to find factor weights. Using eigenvalues and eigenvectors, linear combinations of variables are calculated with coefficients driven by eigenvector components. Those linear combinations give the highest possible proportion of variance explained. The first few combinations are used for factors with the highest eigenvalue (highest proportion of explained variance). Orthogonal rotation (for example varimax) is applied to give better understanding and interpretability of results. The high coefficient (factor weight) is

high correlation between the variable and factor. The final combination can be interpreted as a hidden factor based on factor weights.

Factor Analysis was done using the principal component method and varimax rotation. Factors were selected based on minimum eigenvalue criteria above 1, which means variance above average.

Three factors were selected as satisfying the minimum eigenvalue criteria (see Table 2). Based on weights, the resultant factors can be interpreted as:

Factor 1. the factor most correlated with regions' revenues and region debt per capita but debt is significant only for 2020.

Factor 2. the factor strongly determined by two variables: share of population of working age and average revenues of microenterprises based on tax information, but the correlation in two periods, 2019 and 2020, is the opposite.

Factor 3. the factor defined mainly by the number of microenterprises per ten thousand population. Additionally, in 2020, there was a strong correlation with the unemployment rate.

The unemployment rate was significant in defining factors in 2020, the first pandemic year.

Table 3: Factors after Varimax rotation

a) 2019

Specification	Factor 1	Factor 2	Factor 3
pop_prod_share	0.18670	-0.77530	-0.10166
unemp_prod_share	-0.27000	-0.20258	-0.39967
rev_pc	0.93080	0.07174	0.10684
debt_pc	0.53898	0.65328	-0.01519
micro_per10ths	-0.05261	-0.01930	0.93633
rev_pc_micro	0.36139	0.79679	0.07793

b) 2020

Specification	Factor 1	Factor 2	Factor 3
pop_prod_share	0.10220	0.87344	0.12819
unemp_prod_share	-0.16087	0.20864	-0.72188
rev_pc	0.86714	0.00304	0.20070
debt_pc	0.97001	-0.08564	0.10297
micro_per10ths	0.09385	0.12731	0.80837
rev_pc_micro	0.28583	-0.75014	0.26444

Source: own calculations in SAS 9.4.

4.2. Cluster analysis

Clustering was done using the k-means method to group gminas into four clusters. Before clustering, the data were standardised.

In 2019 (Table 3), the last year before the pandemic, we received two clusters counting for 92 and 46 regions (gmina) respectively and two clusters outliers counting for only 4 and 2 regions. The most frequent cluster counting for 92 municipalities is the cluster with the majority of rural gminas. In the subsequent cluster, counting for 46 regions, there was a quite equal proportion of urban, urban-rural and rural municipalities. The cluster with only four regions is the cluster of cities: Bydgoszcz, Grudziądz, Toruń and Włocławek. The outlying cluster with only two regions is the cluster containing the Tuchola urban-rural gmina and Grudziądz urban gminas. The profile of those outlying regions was very different to other clusters.

In 2020 (table 3), the first year of the pandemic with high restrictions on economic and business activity resulted in a very frequent cluster of 121 regions with diversified structure and cluster counting for 18 gminas with a high proportion of rural gminas. Additionally, two outlying clusters were created, counting for four gminas and one gmina. The cluster with only four gminas is the cluster of cities: Bydgoszcz, Grudziądz, Toruń and Włocławek, and this is exactly the same situation as in the previous year, 2019. The outlying cluster with only one gmina is the urban gmina Grudziądz. The profile of those outlying regions was very different to other clusters.

Table 4: Types of regions in k-means clusters

a) 2019

Type of region	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
urban	4	2	11	0	17
urban-rural	0	17	17	1	35
rural	0	73	18	1	92

b) 2020

Type of region	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
urban	13	0	0	4	17
urban-rural	32	0	3	0	35
rural	76	1	15	0	92

Source: own calculations in SAS 9.4.

For comparison of profiles between clusters and between the years 2019 and 2020, the mean values of variables in clusters were calculated and used (Table 4, Figures 1 and 2).

In 2019, the year with the biggest cluster (92 municipalities) with a profile of rural and mixed gminas, the revenues per capita were medium, with low gmina debt per capita. With a medium level of the average number of microenterprises per ten thousand population, those gminas have a very low level of average tax income from microenterprises.

The next cluster is the group of 46 municipalities with a mixed profile. In this group, like the biggest cluster, the revenues per capita were at the average level, with a low level of gmina debt per capita. Those levels were generally lower comparing to the biggest cluster. Having a low number of microenterprises per ten thousand population, those regions also have a low level of average revenues from taxes from microenterprises.

The cluster with four cities only: Bydgoszcz, Grudziądz, Toruń and Włocławek, is the group with lowest level of unemployment. At the same time, those cities have the highest level of revenues per capita but a very high level of debt per capita, almost equal to revenues . In those cities, there are on average only 108.5 microenterprises per ten thousand population, but average revenues from microenterprises are more than ten times higher comparing to other gminas.

The last cluster with only two regions, the Tuchola urban-rural gmina and Grudziądz rural gmina, is outlying, with their profile regions based on ten times the average number of microenterprises per ten thousand population. On the other hand, the revenues from those microenterprises are only two times higher comparing to other dominant gminas (with the exception of the cluster with four cities).

In 2020, a majority of gminas were grouped into one cluster counting for 121 municipalities. This is the cluster with the lowest level of average revenues per capita with almost the same average level of debt per capita. The average number of microenterprises per ten thousand population was only 81 enterprises, which is much lower comparing to 2019 in comparable clusters. At the same time, those regions have the highest average unemployment rate.

The cluster with 18 gminas has a high level of average revenues per capita with almost an equal level of average debt per capita.

The cluster of four cities had the same composition as in 2019: Bydgoszcz, Grudziądz, Toruń and Włocławek. The profile of those cities is almost the same as in the year before the pandemic. Those cities have the lowest unemployment rate, but slightly higher comparing to 2019. Those cities have the highest revenues but also the highest debt per capita, higher even than revenues. In this cluster there are on average 157 microenterprises per ten thousand population, which is 50% more comparing to 2019, but average revenues from those enterprises are lower comparing to the year before the pandemic.

In the last outlying cluster, there is only one rural gmina, Grudziądz, because of a very high average number of enterprises per 10 thousand population.

Changes are visible between the years 2019 and 2020. In 2020, there was higher unemployment. In the biggest cluster, the unemployment rate is 4.46%. For comparison, unemployment for the two biggest clusters in 2019 was on average 3.79% and 4.21%. In 2020, there was a significant drop in the number of microenterprises.

Table 5: Mean values for variables in clusters

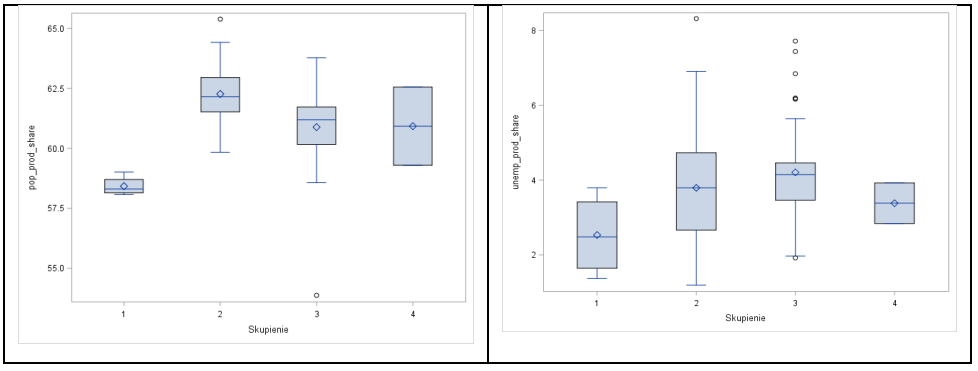
a) 2019

Variable	Cluster 1 (n=4)	Cluster 2 (n=92)	Cluster 3 (n=46)	Cluster 4 (n=2)
pop_prod_share	58.42	62.27	60.89	60.93
unemp_prod_share	2.53	3.79	4.21	3.38
rev_pc	6643.46	5503.33	4877.09	5248.83
debt_pc	6812.92	1282.18	972.23	1409.68
micro_per10ths	108.55	136.39	106.76	1787.08
rev_pc_micro	2397.92	106.29	136.41	238.77

b) 2020

Variable	Cluster 1 (n=121)	Cluster 2 (n=1)	Cluster 3 (n=18)	Cluster 4 (n=4)
pop_prod_share	61.26	62.30	62.28	57.81
unemp_prod_share	4.46	3.55	3.83	3.16
rev_pc	5729.34	6147.93	6634.23	7021.07
debt_pc	5438.61	6109.79	6671.68	7094.96
micro_per10ths	81.14	1307.86	115.41	157.04
rev_pc_micro	118.08	342.95	139.98	1929.02

Source: own calculations in SAS 9.4.



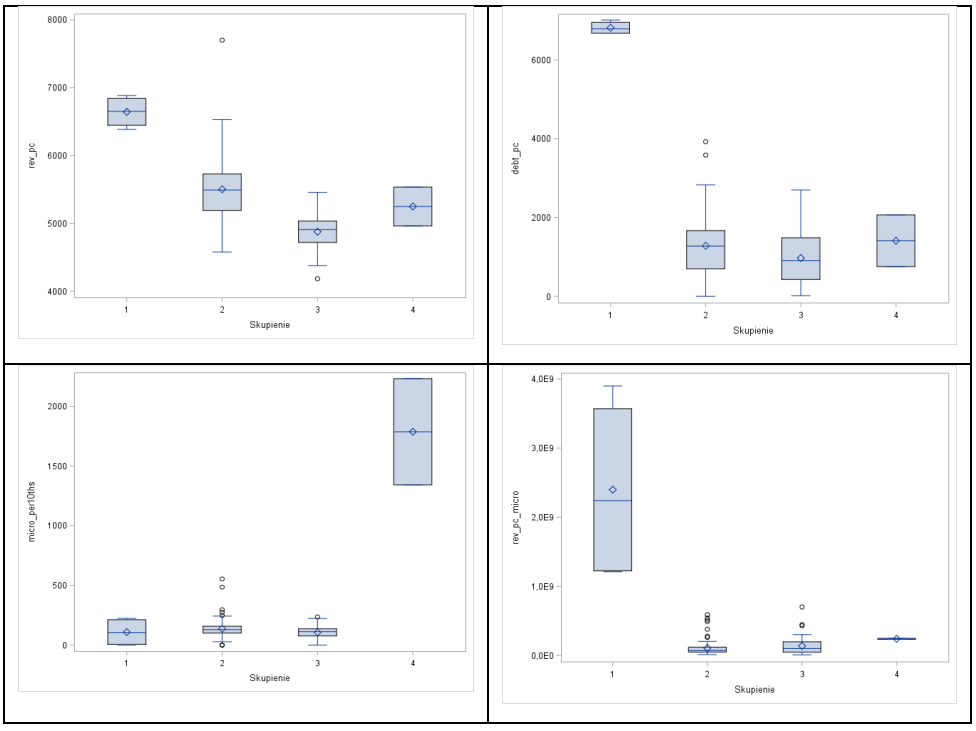
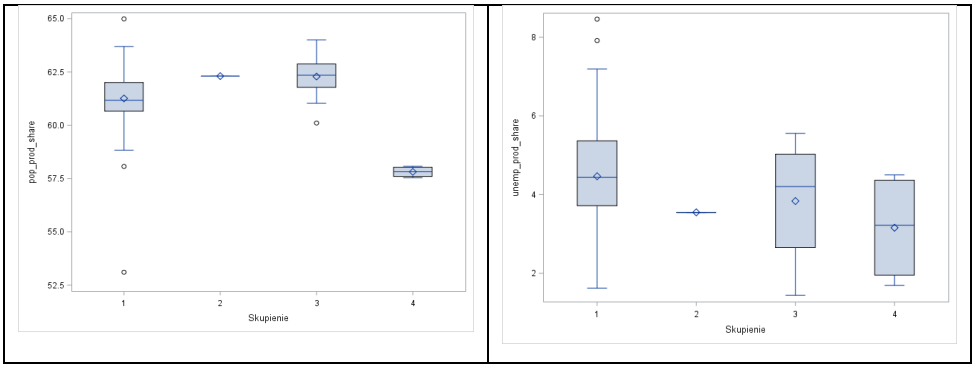


Figure 1: Box-and-whisker plots for variables in clusters – 2019

Source: own calculations in SAS 9.4.



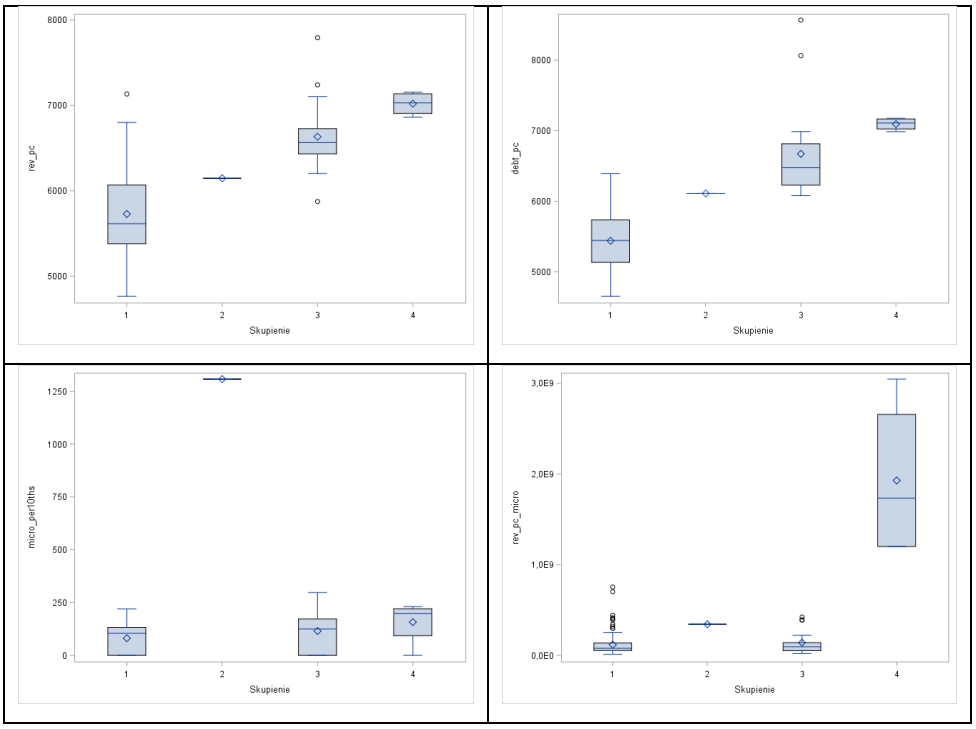


Figure 2: Box-and-whisker plots for variables in clusters – 2020

Source: own calculations in SAS 9.4.

5. Summary and conclusions

In this article, we used a unique data set of the situation of SMEs in the Kujawsko-Pomorskie region to assess the changes of the characteristics of the microenterprise sector at the local level in Poland between 2019 and 2020, that is during the first years of the COVID pandemic. The SME sector in Poland received various types of support from public funds, mainly in the form of stand-by benefits, social security contribution exemptions, and support and guarantees from the Anti-Crisis Shields.

Our analysis shows that there are visible changes in the microenterprise sector and the economic conditions under which they operated. In the largest clusters of gminas, there is a drop in the number of microenterprises per 10 000 population. There is also a significant decline in average revenues reported to tax authorities. This data is consistent with other national statistics, but also observations at the European level, which show that the drop in revenues and financial situation became one of the most important risks faced by the SME sector. We also conclude that the extent to which microenterprises were affected by the pandemic crisis depends on the type of gmina.

Our analysis confirmed that administrative data are a valuable source of information on the situation of SMEs at the local level and potentially a very good source for monitoring the situation of SMEs not only in the Kujawsko-Pomorskie region, but also in other Polish regions.

References

Cepel, M., Gavurova, B., Dvorsky, J. and Belas, J., (2020). *The Impact of the COVID-19 Crisis on the Perception of Business Risk in the SME Segment*, 13, pp. 248–263. https://doi.org/10.14254/2071-8330.2020/13-3/16

Dębkowska, K., Kłosiewicz-Górecka, U., Szymańska, A., Ważniewski, P., Zybertowicz, K., (2021). *Tarcza Antykryzysowa. Koło ratunkowe dla firm i gospodarki*?

European Commission, (2021). Annual Report on European SMEs. *Digitalisation of SMEs* (Issue July).

Frątczak, E. red., (2009). Wielowymiarowa analiza statystyczna. Teoria i przykłady zastosowań z systemem SAS, Warszawa, *Szkoła Główna Handlowa w Warszawie*.

Juergensen, J., Guimón, J. and Narula, R., (2020). European SMEs Amidst the COVID-19 Crisis: Assessing Impact and Policy Responses. *Journal of Industrial and Business Economics*, 47(3), pp. 499–510. https://doi.org/10.1007/s40812-020-00169-4

Panek, T., Zwierzchowski J., (2013). Statystyczne metody wielowymiarowej analizy porównawczej. Teoria i zastosowania. *Oficyna wydawnicza SGH*, Warszawa.

Statistics Poland, (2021). *Activity of Enterprises with up to 9 Persons Employed in 2020*, pp. 23–70.

Zutshi, A., Mendy, J., Sharma, G. D., Thomas, A. and Sarker, T., (2021). From Challenges to Creativity : *Enhancing SMEs' Resilience in the Context of COVID-19*, pp. 1–16.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 153–168, https://doi.org/10.59170/stattrans-2023-057
Received – 01.12.2022; accepted – 10.07.2023

Elimination of characteristics concerning the performance of open-ended equity funds using PCA

Dorota Żebrowska-Suchodolska[1]

Abstract

From an investor's point of view, the appropriate selection of a fund is an important issue. When making such a choice, many elements should be considered. These include not only the fund's rate of return or its risk, but also the comparison of the fund's results with an appropriate benchmark. The aim of the research was to apply principal component analysis (PCA) to reduce the dimension of the indicators that help the investor in selecting a fund. The subject of the study was 15 equity funds that had been on the Polish market for many years. The research showed that it is possible to reduce the primary variables to two dimensions.

Key words: PCA, investment funds, decision-making, tau-Kendall correlation coefficient, investment efficiency.

1. Introduction

The making of an investment decision in the case of mutual funds takes place both at the level of the managers, who have to decide on a specific investment goal, and of the investor. From the point of view of an investor intending to entrust his or her financial funds to investment funds, an important issue is the appropriate choice of fund (Soongswang and Sanohdontree, 2011). Making such a decision is not always obvious, as there are many elements that should be considered when making such a choice. These are, for example, the fund's rate of return, its risk, but also the comparison of the fund's results with the appropriate benchmark. In practice, it is very difficult for an investor to assess a fund in terms of many factors (Kozup et al. 2008), which raises the question of which variables are the most important to guide such a choice. So, there is a need for dimensionality reduction of the variables. This is enabled by principal component analysis (PCA).

[1] Warsaw University of Life Sciences – SGGW, Poland. E-mail: dorota_zebrowska_suchodolska@sggw.edu.pl. ORCID: https://orcid.org/0000-0003-1230-6413.

The aim of the research is therefore to try to look for the main factors determining the choice of an appropriate investment fund in terms of its performance and risk. The research concerns the period from March 12, 2020 to February 23, 2022 and includes fifteen equity funds that have been operating on the Polish market for several years. The period adopted for the research was characterized by an upward trend in the value of participation units. The reaction to the pandemic took place just before the pandemic period (Żebrowska-Suchodolska and Piekunko-Mantiuk, 2022). Therefore, this period was adopted to search for the variables determining fund selection. The subject of the study was 15 equity funds that have been in the market for many years. Principal component analysis (PCA) was used as the research method. The research carried out fits into the issues of investment decision-making and investment efficiency. They also give concrete indications about the choice of appropriate measures for investors. Due to the interconnection of markets with each other, research results may be the basis for making decisions in other markets.

The work is organized as follows. The Chapter 2 contains a literature review. The Chapter 3 presents the characteristics of equity funds against the background of all investment funds in Poland and Chapter 4 presents the methods used for the research, the results of which are presented in Chapter 5. The work ends with the conclusions in Chapter 6.

2. Review of the literature

Investment funds are often assessed in terms of their rate of return and associated risk (Sorros 2003). Risk can be understood here in many ways, whether in a negative, neutral or value at risk context (Rutkowska-Ziarko et al., 2022) (Żebrowska-Suchodolska, 2021, 2022). In order to compare the performance of funds, their performance indicators are determined. With their help, it is possible to compare funds within a group, between groups (Bliss and Potter, 2002) against an established benchmark (Basu and Huang-Jones, 2015), or between different markets and countries (Huij and Post, 2011). Most studies on fund performance are for the US market (Shukla and Singh, 1997). Studies for European market funds are often performed for single countries ((Leite and Cortez, 2013), (Babalos et al., 2012), (Fereira, et al., 2013), (Białkowski and Otten, 2011), (Vidal-García, 2013)) or a group of countries (Otten and Bams, 2002; Božović, 2021). Most studies indicate that funds underperform the market. European funds that have been on the market for a long time are characterized by poor results (Graham et. all, 2020), but also funds investing actively do not give better results than those that invest passively (Berk and van Binsbergen, 2012). Although there are results that exceed the market (Kosowski et al., 2006), they often lack stability (Mateus et al., 2019).

Studies of fund performance can be carried out using different types of measures. These can be both classical and non-classical indicators, which are based on the semi-standard deviation, the value at risk (Małecka, 2021) or the maximum drawdown (Żebrowska-Suchodolska, 2023). It is also important to look for factors that significantly influence fund performance (Filip and Rogala, 2021).

Due to the multitude of indicators, it is difficult for an investor to choose the right one. Research shows that many of them are correlated with each other (Żebrowska-Suchodolska, 2017), but there is still a large number of indicators to choose from. One of the methods that can be used here can be principal component analysis (PCA) (Abdi and Williams, 2010). It is used to reduce the dimension of the space under consideration, which makes it possible to obtain a description of the new variables in the new space to determine the structure of the data set under study (Jackson, 2005). The PCA method thus avoids the curse of dimensionality when dealing with linear data. Reducing redundant variables allows the elimination of those that are not very relevant. Computationally, this reduces memory consumption. The PCA method is used for many economic and social issues (Vyas and Kumaranayake, 2006). In finance, for example, it is used to reduce macroeconomic factors affecting returns (Bilson et al., 2001) and the classification of companies in terms of financial ratios (Yap et al., 2013).

The PCA method for investment funds was used by Zamojska (2013), but her research covered the period 2008-2012. These are the only studies that the author found regarding the reduction of the dimension of performance indicators. Therefore, there is a need to continue this research.

This paper fills a gap in the use of the PCA method to indicate indicators in a two-dimensional space for investment funds. In addition, the author's intention is to obtain pairs of indicators to evaluate the funds. The obtained pairs of indicators will help the investor to decide on an appropriate fund choice guided only by a minimum number of indicators.

3. Equity funds in Poland

Investment funds have been operating on the capital market in Poland for almost thirty years. They account for almost 10% of the household savings portfolio. The basic classification of funds under the Act on Investment Funds and Management of Alternative Investment Funds is the division into: open-ended funds, specialised open-ended funds and closed-ended funds. Table 1 shows the number of these funds for the last five years, i.e. the period 2017–2021.

Table 1: Number of investment funds (data as at Q4 of the year).

Funds and sub-funds	2017	2018	2019	2020	2021
Open-ended funds	334	326	327	312	304
Specialized open-ended funds	294	301	450	311	320
Closed funds	748	679	614	537	503

Source: Own compilation based on NBP.

At the end of 2021, there were 60 investment fund companies operating in Poland. They managed 1127 funds and sub-funds. At that time, there were 304 open-ended funds, which accounted for 26.97%. Over the five-year period, the percentage share of these funds in the number of total funds changed only slightly. The smallest share of open-ended funds was recorded in 2019 and they then accounted for 23.51%. The increasing number of specialised funds resulted in open-ended funds taking third place in terms of their number in 2019 and 2021.

In terms of net asset value (Table 2), closed-end funds accounted for the largest percentage of total assets. Open-ended funds came second. Considering different types of funds, equity funds ranked second in terms of net assets, after debt funds. This position did not change over the period under consideration. The net asset value of equity funds amounted to PLN 25.82 billion at the end of 2021.

Table 2: Net asset value of investment funds (in billion PLN), data as at Q4 of the year

Funds and sub-funds	2017	2018	2019	2020	2021
Open-ended funds	334	326	327	312	304
Specialized open-ended funds	294	301	450	311	320
Closed funds	748	679	614	537	503
Equity	20.04	16.85	16.85	18.41	29.53
Balanced	7.25	6.21	4.91	5.21	5.80
Debt securities	52.83	64.82	76.70	78.05	68.8
Stable growth	10.64	9.20	7.65	7.59	9.01
Other	4.93	4.98	3.43	3.7	1.24

4. Research methodology

The starting point is the daily rate of return, the risk and the performance indicators based on them. The rate of return is understood as $\frac{r_t - r_{t-1}}{r_{t-1}}$, where r_t,r_{t-1} are the values of the fund's participation units at time t and t-1.

The second important measure identifying an asset is risk. This is most commonly understood as negative and positive deviations from the mean, i.e. standard deviation. From an investor's point of view, however, what is more important is the loss that can be incurred from a given investment, or the probability of this loss. Therefore, in addition to standard deviation (S), semi standard deviation (S-), value at risk (VaR), conditional value at risk (CVaR), Ulcer index(U) and maximum drawdown (MDD) were also adopted for the study. These are described by the following formulas:

$$S = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}\left(r_t - \bar{r}\right)^2}, \tag{1}$$

where $\bar{r}$ is the average return and n is the sample size

$$S^{-} = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n} d_t \left(r_t - r_{\min}\right)^2}, \tag{2}$$

where $r_{\min}$ is the minimum required rate of return (here $r_{\min}$ =0) , and d_t is the zero if $r_t > r_{\min}$ and 1 otherwise.

$$VaR = -\left(\bar{r} + q_\alpha S\right), \tag{3}$$

where q_α is the quantile of the standardised normal distribution.

$$CVaR = \bar{r} + \frac{\varphi_{1-\alpha}}{\alpha} S, \tag{4}$$

where $\varphi_{1-\alpha}$ is density function of the standardised normal distribution.

$$U = \sqrt{\frac{1}{n}\sum_{t=1}^{n} D_t^2}, \tag{5}$$

where D_t is the relative decrease in the value of fund A shares in period *t*.

$$MDD = \min D_t \tag{6}$$

The combination of return and risk is represented by investment performance indicators, for which references include acceptable investment return, benchmark, or risk-free assets. These are taken into account by the following indicators: Sharpe, Sortino, Calmar, Martin, RVaR and CS. The selected performance indicators are described by the following formulas:

$$Sharp = \frac{\bar{r} - r_f}{S}, \tag{7}$$

where r_f is the average risk free rate.

$$Sortino = \frac{\bar{r} - r_{\min}}{S^{-}}, \tag{8}$$

$$C = \frac{\bar{r}}{MDD}, \tag{9}$$

$$M = \frac{\bar{r} - r_f}{U}, \tag{10}$$

$$RVaR = \frac{\bar{r} - r_f}{VaR}, \tag{11}$$

$$CS = \frac{\bar{r} - r_f}{CVaR}. \tag{12}$$

The large number of indicators and measures creates the need to reduce them so that the investor can make a decision on the basis of the fewest number of variables that do not duplicate information. The tau-Kendall correlation coefficient determined here makes it possible to examine the relationship between the measures adopted and the uncorrelated variables are the starting point for further considerations. Although the dimension of the uncorrelated variables is smaller than that of all variables, it is often still too large to make an investment decision. For this purpose, principal component analysis (PCA) was used. It allows to reduce the dimension of the underlying variables, leaving the most relevant ones, which makes the resulting group more homogeneous.

Principal component analysis (PCA) was first described by Pearson (1901) and developed by Hotteling (1933, 1936).

The starting point of the PCA method is the determination of the principal components, which are a linear combination of the primary variables:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p \ ,$$

where $X_1, X_2, \ldots, X_p$ are the primary variables.

The principal components are the result of determining the eigenvalues and eigenvectors from the following equation:

$$(M - \lambda I)a = 0 \,,$$

where λ are the eigenvalues of the matrix M, M - the covariance matrix of the primary variables, I - the unit matrix, and $a = (a_{i1}, a_{i2}, , , , , a_{ip})$ the eigenvector corresponding to the i-th eigenvalue. A non-zero solution exists when $|M - \lambda I| = 0$. The largest eigenvalues of the covariance matrix M are searched in order. The coefficients

of the corresponding eigenvector are the coefficients $a_{i1}, a_{i2}, \ldots, a_{ip}$ of the principal components which correspond to the i-th largest eigenvalue of the covariance matrix M.

The resulting principal components are uncorrelated with each other, and their final separation can be based on the Kaiser criterion (1960), or the percentage of explained variability by the principal components. Often, as few as two components may be sufficient here, especially if they exceed 75% of the total variability of all variables (Morison 1990).

The application of the described steps will contribute to the verification of the following research theses and hypotheses:

T1: a reduction in the indicators describing the funds in terms of their performance will help the investor in making his investment decision.

H1: the fund selection decision can be made on the basis of two indicators.

4. Results of the study

The subject of the research were fifteen equity funds that have been operating on the market for several years. They were: Allianz Polskich Akcji, Esaliens Akcji, Generali Korona Akcje, Investor Akcji Spółek Dywidendowych, Investor Akcji, Investor Top 25 Małych Spółek, Millennium Akcji, NN Akcji, Novo Akcji, Pekao Akcji Polskich, Pzu Akcji Krakowiak, Rockbridge Akcji Małych i Srednich Spółek, Rockbridge Akcji, Santander Akcji, Skarbiec Akcja. The research was based on the daily values of participation units of these funds in the period from March 12, 2020 to February 23, 2022. In the case of equity funds, the research period was characterised by an upward trend, as the reaction to the pandemic took place immediately earlier (Żebrowska-Suchodolska, Piekunko-Mantiuk 2022). Therefore, this period was selected to search for variables determining the choice of a fund and to reduce their dimensions.

For the funds, the average rate of return was calculated as well as the measures described by formulas (1) - (12) for which the tau-Kendall correlation coefficient was determined. The values of the tau-Kendall correlation coefficient are presented in Table 3.

The values of the tau-Kendall correlation coefficient indicated the existence of a relationship between many analyzed indicators. Thus, they provided a basis for removing them from further considerations. After this selection, the following groups of indicators not correlated with each other were selected:

1) $\bar{r}$, S, VaR, CVaR, MDD
2) $\bar{r}$, S, MDD, C, RVaR, CS
3) S^-
4) $\bar{r}$, VaR, MDD, RVaR, CS

5) $\bar{r}$, CVaR, MDD, RVaR, CS
6) U
7) $\bar{r}$, S, VaR, CVar, MDD, Sharp, Sortino, C, M, RVaR, CS
8) MDD, Sharpe
9) MDD, Sortino
10) S, MDD, C
11) MDD, M
12) S, VaR, CVaR, MDD, RVaR
13) S, VaR, CVaR, MDD, CS

Table 3: The values of the tau-Kendall correlation coefficient.

Specification	$\bar{r}$	S	S^-	VaR	CVaR	U	MDD	Sharp	Sortino	C	M	RVaR	CS
$\bar{r}$	1	-0.26	**-0.45**	-0.35	-0.34	**-0.50**	0.15	**0.79**	**0.79**	**0.58**	**0.79**	**1**	**1**
S		1	1	**0.81**	**0.93**	**0.49**	-0.31	**-0.47**	**-0.47**	-0.30	**-0.39**	-0.26	-0.26
S^-			1	**0.91**	**0.89**	**0.60**	**-0.39**	**-0.66**	**-0.66**	**-0.45**	**-0.54**	**-0.45**	**-0.45**
VaR				1	**0.99**	**0.59**	-0.37	**-0.57**	**-0.57**	**-0.39**	**-0.49**	-0.35	-0.35
CVaR					1	**0.58**	-0.36	**-0.56**	**-0.56**	**-0.38**	**-0.48**	-0.34	-0.34
U						1	**-0.45**	**-0.68**	**-0.68**	**-0.73**	**-0.71**	**-0.50**	**-0.50**
MDD							1	0.28	0.28	0.18	0.31	0.14	0.14
Sharp								1	**0.96**	**0.71**	**0.85**	**0.79**	**0.79**
Sortino									1	**0.68**	**0.81**	**0.79**	**0.79**
C										1	**0.79**	**0.58**	**0.58**
M											1	**0.79**	**0.79**
RVaR												1	1
CS													1

*values in bold are statistically significant at the 0.05 significance level

Source: Own calculation using Statistica.

For each group containing more than two variables, the PCA method was used to reduce the dimension and find the variables with the highest percentage of principal components explaining the variability. The first two components explained more than eighty percent of the overall variability, so on the basis of the scree plot criterion they can be considered sufficient to decide on the number of principal components.

A representation of the performance indicators in terms of the first two principal components is shown in Figure 1.

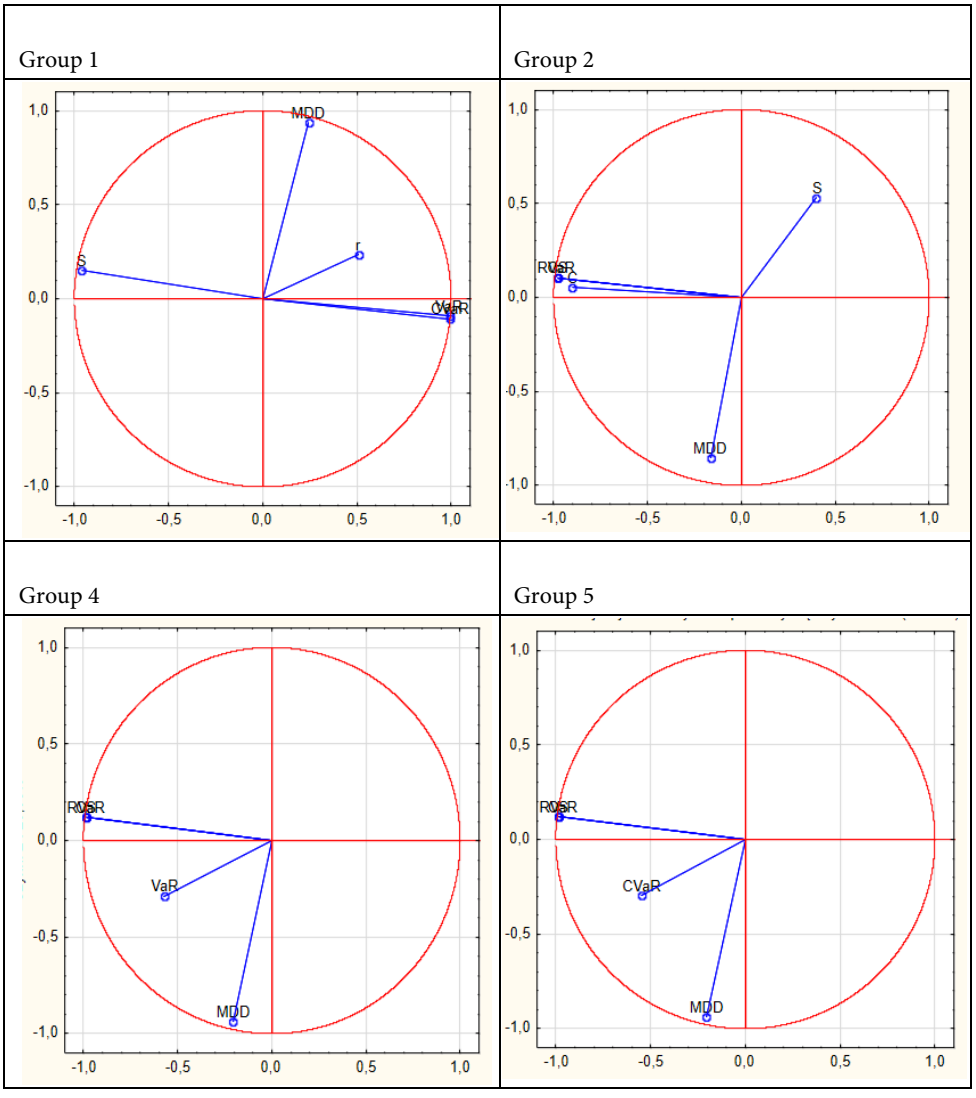


Figure 1: A representation of the performance indicators in terms of the first two principal components

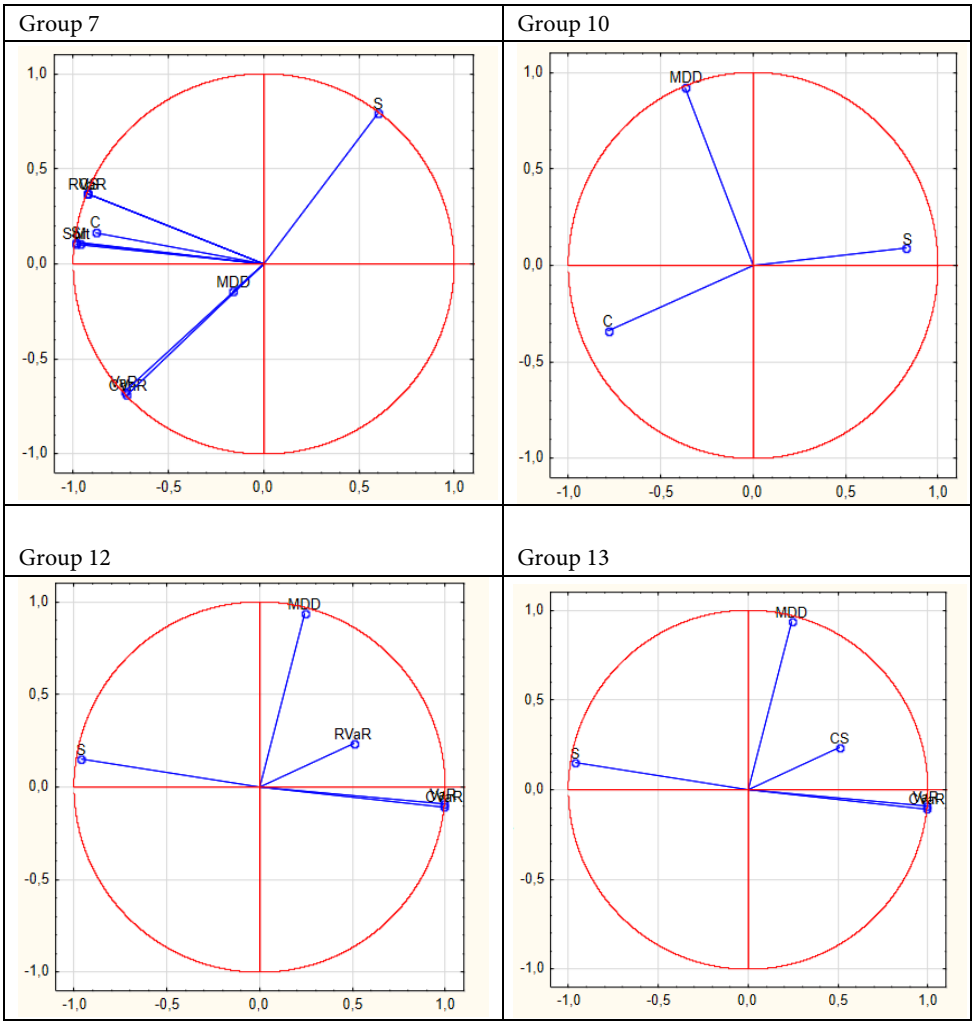


Figure 1: A representation of the performance indicators in terms of the first two principal components (cont.)

Source: Own calculation using Statistica.

In Figure 1, for each group are placed points (charges) in the unit circle. The position of the point corresponds to the information of this variable carried by the first two principal components. The closer the point is to the edge of the circle, the better it is represented by the principal components. The position of the points relative to each other, in turn, provides other information. The close position of the vectors indicates the existence of a positive correlation between the variables. Their position on the opposite side indicates negative correlation. Their perpendicular position relative to each other indicates that the variables are uncorrelated.

Projecting the indicators onto the plane of the first two components shows that in most cases the points representing the individual indicators lie at or close to the edge of the circle. This indicates that these indicators are well represented by the principal components and that they carry most of the information contained in the output indicators. In addition, the measures are located in other parts of the circle indicating that they carry quite different information. Thus, their designation here is important for the overall assessment of the fund.

The largest percentages of indicators in each principal component allow the most important indicators to be identified in terms of the importance of the information they convey. These are the following indicators in each group:

1) MDD, VaR/CVaR

2) MDD, r/RVar/CS

4) r/RVar/CS, MDD

5) r/RVaR/CS, MDD

7) Sharp/Sortino/M, S

10) S, MDD

12) VaR, MDD

13) VaR, MDD

Principal component analysis also allows investment funds to be shown in a two-dimensional factor space. The projection of the funds on the factor plane is shown in Figure 2.

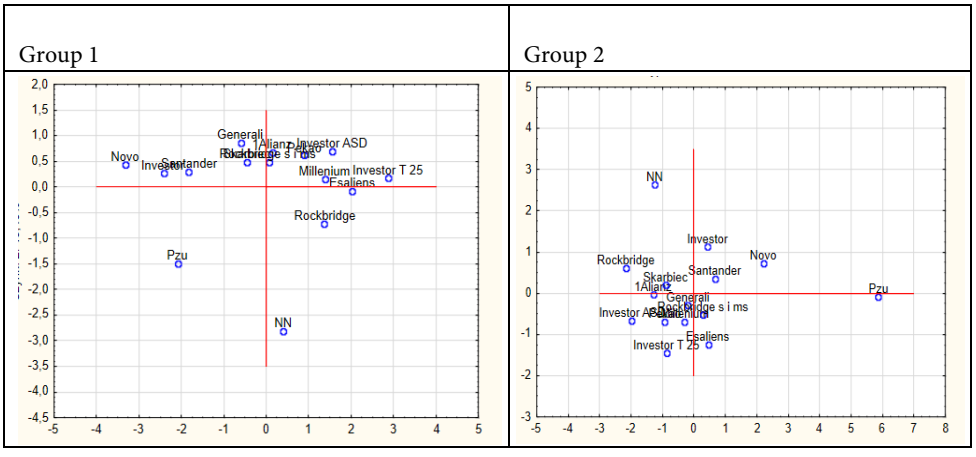


Figure 2: The projection of the funds on the factor plane

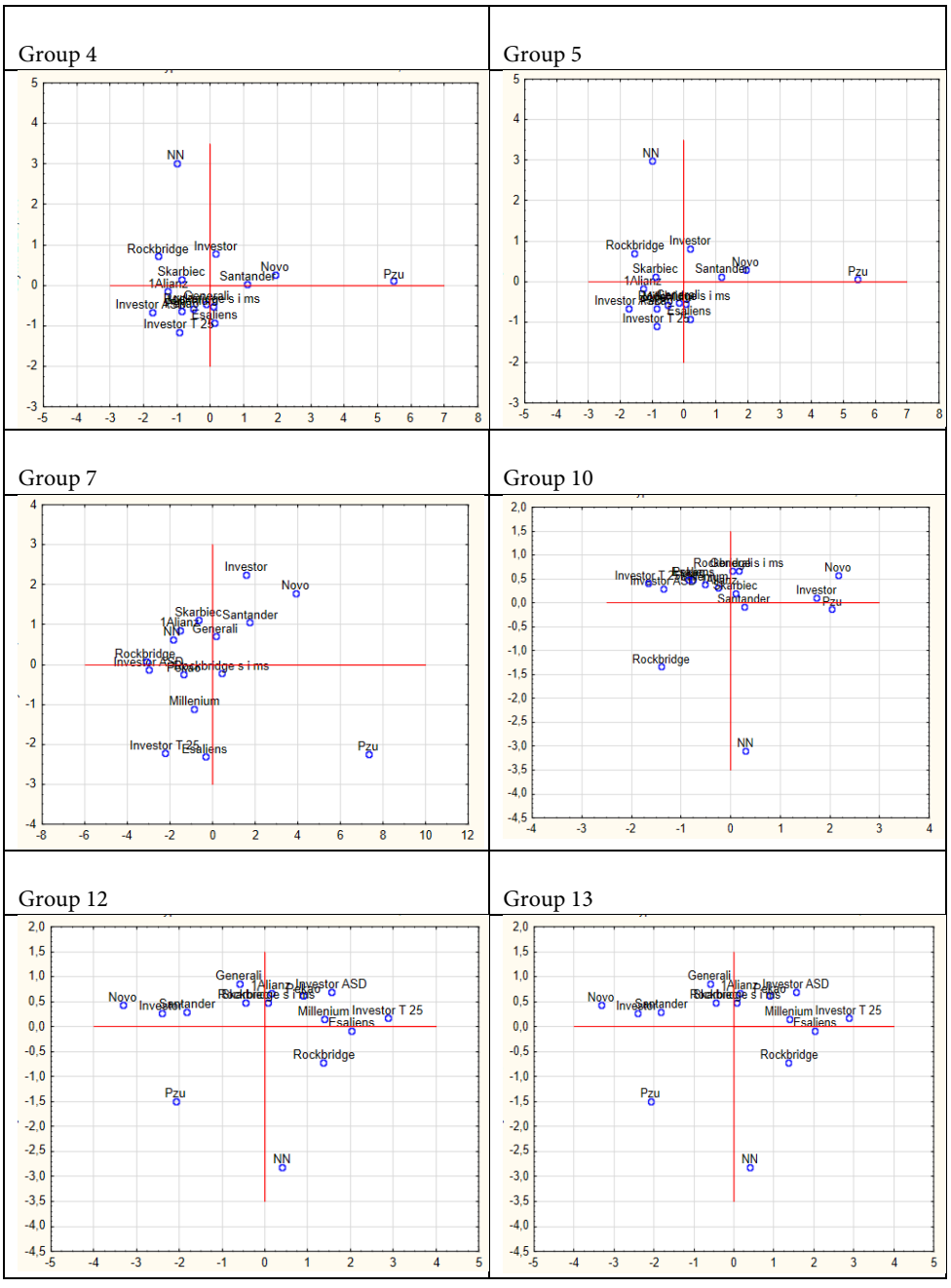


Figure 2: The projection of the funds on the factor plane (cont.)

Source: Own calculation using Statistica.

The marked points correspond to the funds. They are plotted on the plane of the first two principal components. From the graph, you can read the values of the first two components for each fund. In addition, the position of the points shows the similarity between the funds. The closer the funds are located, the more similar they are to each other.

The projection of the funds onto the plane of the first two components indicates the existence of three clusters of points in most cases. Only two clusters are discernible in the case of group 7. The first cluster contains the majority of funds, while the others contain only individual funds. Similar results in terms of the measures considered are indicative of a similar investment policy pursued by managers. This is because the funds in most cases only emulate the market, and the skills of selectivity and market timing are present in single cases (Żebrowska-Suchodolska and Karpio, 2018). Outliers of values from the largest cluster occurred for PZU, NN and Rockbridge funds in group 10. They constituted single clusters. Therefore, in the case of these funds, the results of the measures taken into account differ significantly from the others.

5. Conclusions

The aim of the research was to try to look for the main factors determining the choice of an appropriate investment fund in terms of its performance and risk. Principal component analysis was used for this. The study covered the period from March 12, 2020 to February 23, 2022 and involved fifteen equity funds that had been operating on the Polish market for several years. 13 groups were selected for the study. The groups were selected in terms of correlation of indicators. They contained from 1 to 10 indicators.

The research showed that it is possible to reduce the primary variables to two dimensions, confirming the hypothesis H1. Two indicators were also indicated by Zamojska as sufficient to assess the performance of the funds. This will help the investor to make the right decision on the fund selection (T1) by taking only two indicators. To evaluate a given investment in funds, the investor should choose the MDD measure and some measure of risk (VaR/CvaR/S).

Besides, the pairs of indicators included in the principal components have been placed in other parts of the circle, allowing the investor to assess the fund from the point of view of completely different information. The resulting indicators found in each group are based on a combination of classical and non-classical measures. It is only with this combination that the contribution of the output variable to the principal component is best. Some of the pairs contain only the risk measures themselves, which shows how important they are when evaluating a fund and from the point of view of the loss that an investor may suffer.

References

Abdi, H., Williams, L. J. (2010). Principal Component Analysis. *Wiley interdisciplinary reviews: computational statistics*, Vol. 2(4), pp. 433–459.

Babalos, V., Caporale, G. M., Philippas, N., (2012). Efficiency Evaluation of Greek Equity Funds. *Research in International Business and Finance*, Vol. 26(2), pp. 317–333.

Basu, A. K., Huang-Jones, J., (2015). The Performance of Diversified Emerging Market Equity Funds. *Journal of International Financial Markets*, Institutions and Money, Vol. 35, pp. 116–131.

Berk, J. B., Van Binsbergen, J. H., (2012). Measuring Managerial Skill in the Mutual Fund Industry. *National Bureau of Economic Research*, No. w18184.

Białkowski, J., Otten, R., (2011). Emerging Market Mutual Fund Performance: Evidence for Poland. *The North American Journal of Economics and Finance*, Vol. 22(2), pp. 118–130

Bliss, R. T., Potter, M. E., (2002). Mutual Fund Managers: Does Gender Matter? *The Journal of Business and Economic Studies*, Vol. 8(1), pp. 1–15.

Bilson, C. M., Brailsford, T. J., Hooper, V. J., (2001). Selecting Macroeconomic Variables as Explanatory Factors of Emerging Stock Market Returns. *Pacific-Basin Finance Journal*, Vol. 9(4), pp. 401–426.

Božović, M., (2021). Mutual Fund Performance: Some Recent Evidence from European Equity Funds. *Economic Annals*, Vol. 66(230), pp. 7–33.

Ferreira, M. A., Keswani, A., Miguel, A. F., Ramos, S. B., (2013). The Determinants of Mutual Fund Performance: A Cross-Country Study. *Review of Finance*, Vol. 17(2), pp. 483–525.

Filip, D., Rogala, T., (2021). Analysis of Polish Mutual Funds Performance: a Markovian Approach. *Statistics in Transition new series*, Vol. 22(1), pp. 115–130.

Graham, J. E., Lassala, C., Ribeiro Navarrete, B., (2020). Influences on Mutual Fund Performance: Comparing US and Europe Using Qualitative Comparative Analysis. *Economic research-Ekonomska Istraživanja*, Vol. 33(1), pp. 3049–3070

Hotelling, H., (1933). Analysis of a Complex of Statistical Variables Into Principal Components. *Journal of Educational Psychology*, Vol. 24, pp. 417–441.

Hotelling, H., (1936). Relations Between Two Sets of Variates. *Biometrika*, Vol. 28, pp. 321–377.

Huij, J., Post, T., (2011). On the Performance of Emerging Market Equity Mutual Funds. *Emerging Markets Review*, Vol. 12(3), pp. 238–249.

Jackson, J. E., (2005). *A User's Guide to Principal Components*, John Wiley & Sons.

Kaiser, H. F., (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, Vol. 20(1), pp. 141–151.

Kosowski, R., Timmermann, A., White, H. Wermers, R., (2006). Can Mutual Fund 'Stars' Really Pick Stocks? New Evidence from a Bootstrap Analysis. *Journal of Finance*, Vol. 61(6), pp. 2551–2595.

Kozup, J., Howlett, E., Pagano, M., (2008). The Effects of Summary Information on Consumer Perceptions of Mutual Fund Characteristics. *Journal of Consumer Affairs*, Vol. 42(1), pp. 37–59.

Leite, P. A., Cortez, M. C., (2013). Conditioning Information in Mutual Fund Performance Evaluation: Portuguese Evidence. *Asset Management and International Capital Markets*, pp. 101–122.

Małecka, M., (2021). Testing for a Serial Correlation in VaR Failures Through the Exponential Autoregressive Conditional Duration Model. *Statistics in Transition new series*, Vol. 22(1), pp. 145–162.

Mateus, I. B., Mateus, C., Todorovic, N., (2019). Review of New Trends in the Literature on Factor Models and Mutual Fund Performance. *International Review of Financial Analysis*, Vol. 63, pp. 344–354.

Morison D. F., (1990). Wielowymiarowa analiza statystyczna, PWN, Warszawa.

Otten, R. Bams, D., (2002). European Mutual Fund Performance. *European Financial Management*, Vol. 8(1), pp. 75–101.

Pearson, K., (1901). *On Lines and Planes of Closest Fit to Systems of Points in Space, Philosophical Magazine*, Vol. 2(11), pp. 559–572.

Rutkowska-Ziarko, A., Markowski, L., Pyke, C., Amin, S., (2022). Conventional and Downside CAPM: The Case of London Stock Exchange. *Global Finance Journal*, Vol. 54, 100759.

Shukla, R., Singh, S., (1997). A Performance Evaluation of Global Equity Mutual Funds: Evidence from 1988–1995. *Global Finance Journal*, Vol. 8(2), pp. 279–293.

Soongswang, A., Sanohdontree, Y., (2011). Open-Ended Equity Mutual Funds. *International Journal of Business and Social Science*, Vol. 2(17), pp. 127–136.

Sorros, J. N., (2003*). Return and Risk Analysis: A Case Study in Equity Mutual Funds Operating in the Greek Financial Market*, Managerial Finance.

Vidal-García, J., (2013). The Persistence of European Mutual Fund Performance. *Research in International Business and Finance*, Vol. 28, pp. 45–67.

Vyas, S., Kumaranayake, L., (2006). Constructing Socio-Economic Status Indices: How to Use Principal Components Analysis. *Health policy and planning*, Vol. 21(6), pp. 459–468.

Yap, B. C. F., Mohamad, Z., Chong, K. R., (2013). The Application of Principal Component Analysis in the Selection of Industry Specific Financial Ratios. *British Journal of Economics*, *Management* & *Trade*, Vol. 3(3), pp. 242–252.

Zamojska, A., (2013). Empirical Analysis of the Consistency of Mutual Fund Ranking for Different Portfolio Performance Measures. *Research Papers of Wrocław University of Economics*, Vol. 279, pp. 95–105.

Żebrowska-Suchodolska, D., (2017). The Measurement of Effectiveness of Investment Equity Funds in the Period 2004–2014 Using Drawdown Measures. *Econometrics*, Vol. 2 (56), pp. 104–115.

Żebrowska-Suchodolska, D., Karpio, A., (2018). Market Timing Models for Equity Funds Operating on the Polish Market in the Years 2003–2017. In International Conference on Computational Methods in Experimental Economics, *Springer*, *Cham*, pp. 291–309.

Żebrowska-Suchodolska, D., (2021). *Is The Size of The Fund Important in a Pandemic?* Research for Polish Equity and Bond Funds, *Innovation Management and information Technology impact on Global Economy in the Era of Pandemic*. Proceedings of the 37th International Business Information Management Association Conference (IBIMA) 30–31 May 2021, Cordoba, Spain.

Żebrowska-Suchodolska, D., (2022). Similarity of Open-Ended Mutual Funds During a Pandemic. Research for Equity and Bond Funds. In Modern Classification and Data Analysis: *Methodology and Applications to Micro-and Macroeconomic Problems Springer International Publishing*, pp. 135–146.

Zebrowska-Suchodolska, D., (2023). Risks from Investing in Open-Ended Mutual Funds – Impact of Net Asset Value. *Montenegrin Journal of Economics*, Vol. 19(4), pp. 17–27

Żebrowska-Suchodolska, D., Piekunko-Mantiuk, I., (2022). Similarity and Granger Causality in Polish and Spanish Stock Market Sectors During the COVID–19 Pandemic, Comparative Economic Research. *Central and Eastern Europe*, Vol. 25(3), pp. 90–109.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 169–178, https://doi.org/10.59170/stattrans-2023-058
Received – 15.09.2022; accepted – 23.03.2023

An empirical study of hierarchical Bayes small area estimators using different priors for model variances

Yong You[1]

Abstract

In this paper, we study hierarchical Bayes (HB) estimators based on different priors for small area estimation. In particular, we use inverse gamma and flat priors for variance components in the HB small area models of You and Chapman (2006) and You (2021). We evaluate the HB estimators through a simulation study and real data analysis. Our results indicate that using the inverse gamma prior for the variance components in the HB models can be very effective.

Key words: CPO, flat prior, inverse gamma prior, relative error, variance component.

1. Introduction

Small area estimation is very popular and important in survey data analysis due to growing demand for reliable small area estimates. Model-based estimates have been widely used to provide reliable indirect estimates. Various area level models have been proposed in the literature to improve direct survey estimates, see Rao and Molina (2015). In this paper, we use the well-known Fay-Herriot model (Fay and Herriot, 1979) as a basic model and present the Fay-Herriot model in hierarchical Bayes (HB) framework of You and Chapman (2006) and You (2016, 2021). The Fay-Herriot model has two components, namely a sampling model for the direct survey estimates and a linking model for small area parameters of interest. The sampling model assumes that a direct estimator y_i is design unbiased for a small area parameter θ_i such that

$$y_i = \theta_i + e_i,\ i = 1,\ldots,m, \tag{1}$$

where e_i is the sampling error and m is the number of small areas. It is customary to assume that e_i's are independently distributed normal random variables with mean $E(e_i|\theta_i) = 0$ and variance $Var(e_i|\theta_i) = \sigma_i^2$. The linking model assumes that the small

[1] ESMD, Statistics Canada, Ottawa, K1A 0T6, Canada. E-mail: yongyou@statcan.gc.ca.
ORCID: https://orcid.org/0009-0000-8030-1484.

area parameter θ_i is related to area level auxiliary variables $x_i = (x_{i1}, \ldots, x_{ip})'$ through a linear regression model

$$\theta_i = x_i'\beta + v_i,\ i = 1,\ldots,m, \tag{2}$$

where $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients and v_i's are random effects assumed to be independent and normally distributed with $E(v_i) = 0$ and $Var(v_i) = \sigma_v^2$. The model variance σ_v^2 is unknown and needs to be estimated. Combining models (1) and (2) leads to a linear mixed area level model given as

$$y_i = x_i'\beta + v_i + e_i,\ i = 1,\ldots,m. \tag{3}$$

Model (3) involves both design-based random errors e_i and model-based random effects v_i. For the Fay-Herriot model, the sampling variance σ_i^2 is assumed to be known in model (3). This is a very strong assumption. Generally smoothed estimators of the sampling variances are used in the Fay-Herriot model and then treated as known. Alternatively, the sampling variance σ_i^2 can be modelled together with the small area parameter θ_i. Let s_i^2 denote a direct estimator for σ_i^2. We consider a commonly used model for s_i^2 as $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$ and n_i is the sample size for the i-th area. We combine the sampling variance model $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$ with the small area model (3) to construct an integrated model in the HB framework. The integrated model borrows strength for small area estimates and sampling variance estimates simultaneously. This integrated HB modelling approach has been widely used in practice, for example, see You and Chapman (2006), Dass, Maiti, Ren and Sinha (2012), Sugasawa, Tamae and Kubokawa (2017), Ghosh, Myung and Moura (2018), Hidiroglou, Beaumont and Yung (2019) and You (2008, 2021).

In Section 2, we present two HB small area models and consider two priors for variance components, namely inverse gamma (IG) prior and flat prior. In Section 3, we conduct a simulation study to evaluate the impact of priors on small area estimation. In Section 4, we apply the models to a real data application. And in Section 5, we offer some concluding remarks.

2. Hierarchical Bayes small area models

In this section, we present two HB models with sampling variance modelling. The first model is considered in You and Chapman (2006), in which an inverse gamma model is used for the sampling variance σ_i^2 with known vague values. The second model is considered in You (2016, 2021), where a log-linear random error model is used for σ_i^2.

HB Model 1: You-Chapman Model (You and Chapman, 2006), denoted as YCM:

- $y_i|\theta_i, \sigma_i^2 \sim ind\ N(\theta_i, \sigma_i^2)$, $i = 1,\ldots,m$;
- $d_i s_i^2 \big| \sigma_i^2 \sim ind\ \sigma_i^2 \chi_{d_i}^2$, $d_i = n_i - 1$, $i = 1,\ldots,m$;

- $\theta_i|\beta,\sigma_v^2 \sim ind\ N(x_i'\beta,\sigma_v^2)$, $i = 1,\ldots,m$;
- $\sigma_i^2 \sim IG(a_i,b_i)$, where $a_i = 0.0001$, $b_i = 0.0001$, $i = 1,\ldots,m$;
- priors for unknown parameters:$\pi(\beta) \propto 1, \pi(\sigma_v^2) \sim IG(a_v,b_v)$, where a_v, b_v are chosen to be very small constants (0.0001) to reflect vague knowledge on σ_v^2.

The full conditional distributions for the Gibbs sampling procedure under YCM can be found in You and Chapman (2006).

HB Model 2: You (2016, 2021) log-linear model on sampling variances, denoted as YLLM:

- $y_i|\theta_i,\sigma_i^2 \sim ind\ N\left(\theta_i,\sigma_i^2\right)$, $i = 1,\ldots,m$;
- $d_i s_i^2\left|\sigma_i^2 \right. \sim ind\ \sigma_i^2\chi_{d_i}^2,\ d_i = n_i - 1$, $i = 1,\ldots,m$;
- $\theta_i|\beta,\sigma_v^2 \sim indN(x_i'\beta,\sigma_v^2)$, $i = 1,\ldots,m$;
- $log(\sigma_i^2) \sim N\left(\delta_1 + \delta_2\, log(n_i),\tau^2\right)$, $i = 1,\ldots,m$;
- priors for unknown parameters:$\pi(\beta) \propto 1$, $\pi(\delta_1,\delta_2) \propto 1$, $\pi(\sigma_v^2) \sim IG(a_v,b_v)$, $\pi(\tau^2) \sim IG(a_\tau,b_\tau)$, where $a_v,\ b_v,\ a_\tau,\ b_\tau$ are chosen to be very small constants (say, 0.0001).

The full conditional distributions for the Gibbs sampling procedure under YLLM are given in the Appendix.

For both YCM and YLLM, we use IG priors with very small constant parameters for the variance components σ_v^2 and τ^2. Ghosh, Myung and Moura (2018) used an IG prior with some fixed values for the model variance σ_v^2. IG prior is a proper prior and conditionally conjugate for the variance components. IG prior is widely used in Bayesian literature (e.g. Gelman, Carlin, Stern and Rubin, 2004) and Bayesian software packages (e.g. WinBUGS, Lunn, Thomas, Best and Spiegelhalter, 2000). Alternatively, flat priors $\pi(\sigma_v^2) \propto 1$ and $\pi(\tau^2) \propto 1$ can be used for the model variances σ_v^2 and τ^2 in the YCM and YLLM models. Flat prior is used as a non-informative prior in the literature (e.g. Gelman, 2006). You (2021) compared the models of YCM and YLLM with the model of Sugasawa, Tamae and Kubokawa (2017) using flat priors on the variance components. In this paper we use YCM and YLLM as two studying models and compare the HB estimators using IG and flat priors through simulation study and real data analysis.

3. Simulation study

In this section, we estimate model variance σ_v^2 and small area means through a simulation study. We generate $\theta_i = \boldsymbol{x}'_i\beta + v_i = \beta_0 + x_i\beta_1 + v_i$ with $\beta_0 = 3.5$ and $\beta_1 = 1.5$ fixed through the simulation. The single covariate x_i is generated from an exponential distribution with mean equal to 1, and then fixed for the simulation study. Random effect v_i is generated from $v_i \sim N(0,\sigma_v^2)$. Following Lahiri and Rao (1995) and

Rivest and Vandal (2002), we let the number of small areas m = 30. These 30 areas are divided into five groups with different sampling variances. The true sampling variance is set at σ_i^2 = 1, 0.75, 0.5, 0.25, and 0.1 for each grouped areas, with the corresponding sample size n_i= 4, 6, 8, 10 and 12. That is, for areas from 1 to 6, σ_i^2 = 1 (i = 1,…,6) and the corresponding sample size n_i= 4 for each area (i = 1,…,6). For areas 7 to 12, σ_i^2 = 0.75 (i = 7,…,12) and the corresponding sample size n_i= 6 for each area (i = 7,…,12). And so on for other areas. We consider three choices of the model variance: the true σ_v^2 is set to be 1, 0.5 and 0.1, respectively. The direct sampling variance estimate is generated as $s_i^2 = (d_i)^{-1}\sigma_i^2\chi_{d_i}^2$, where $d_i = n_i - 1$ (e.g. Ghosh, Myung and Moura, 2018; You, 2021). For each case, we perform 5000 simulation runs. For each run, the Gibbs sampling procedure consists of 1000 burn-in period and 5000 more iterations for each simulation run.

We first compare the estimates of the model variance σ_v^2 based on YCM and YLLM using IG and flat priors on σ_v^2. Table 1 presents the estimates of σ_v^2 when the true σ_v^2 is 1, 0.5 and 0.1. It is clear from Table 1 that both YCM and YLLM lead to almost unbiased estimates of the model variance under IG prior. However, when flat prior is used, both YCM and YLLM lead to over-estimation of the model variance. The over estimation is substantially large when the true model variance is small. For example, if the true σ_v^2 is 1, under flat prior, both YCM and YLLM lead to about 22% over-estimation; if the true σ_v^2 is 0.1, the over-estimation could be more than 100%. The result in Table 1 indicates that IG prior performs much better than the flat prior for the model variance estimation.

Table 1: Estimates of model variance under YCM and YLLM using IG and flat priors

True σ_v^2	YCM		YLLM	
	IG prior	Flat prior	IG prior	Flat prior
1	1.025	1.217	1.027	1.228
0.5	0.518	0.683	0.512	0.672
0.1	0.119	0.251	0.096	0.225

To compare the small area HB estimators, we consider the average absolute relative bias (ARB) for the HB estimator $\hat{\theta}_i$ of the simulated small area mean θ_i as $\overline{ARB} = (\sum_{i=1}^{m} ARB_i)/m$, where

$$ARB_i = \left|\frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{\theta}_i^{(r)} - \theta_i^{(r)})}{\theta_i^{(r)}}\right|,$$

and $\hat{\theta}_i^{(r)}$ is the HB estimate and $\theta_i^{(r)}$ is the true mean based on the r-th simulated sample, R = 5000. The estimated average coefficient of variation (ACV) is computed as $\overline{ACV} = (\sum_{i=1}^{m} CV_i)/m$, where

$$CV_i = \frac{1}{R}\sum_{r=1}^{R} CV_i^{(r)} \text{and } CV_i^{(r)} = \frac{\sqrt{var(\hat{\theta}_i^{(r)})}}{\hat{\theta}_i^{(r)}}$$

where $var(\hat{\theta}_i^{(r)})$ is estimated posterior variance of the HB estimator $\hat{\theta}_i^{(r)}$. We also compare the average simulation relative root MSE (RRMSE), and the RRMSE is computed as $\overline{RRMSE} = (\sum_{i=1}^{m} RRMSE_i)/m$, where

$$RRMSE_i = \frac{1}{R}\sum_{r=1}^{R} RRMSE_i^{(r)}, \text{ and } RRMSE_i^{(r)} = \frac{\sqrt{(\hat{\theta}_i^{(r)} - \theta_i^{(r)})^2}}{\theta_i^{(r)}}.$$

Table 2 presents the comparison results of ARB, ACV and RRMSE under models YCM and YLLM using IG and flat priors. The HB estimator $\hat{\theta}_i$ should be unbiased for the small area parameter θ_i following the conditional posterior distribution of θ_i given in the Appendix. When the true σ_v^2= 1, the average ARB is around 1.7% to 1.8% for both models YCM and YLLM, and the average ARB becomes much smaller when the true σ_v^2= 0.1. The results of ARB also indicate that the posterior HB estimators are unbiased for the small area parameter θ_i under both the IG and flat priors. However, both YCM and YLLM have smaller average CVs and RRMSE using IG prior, and particularly, using the flat prior leads to much larger average CVs for both YCM and YLLM. For example, when the true σ_v^2= 0.1, the average CV using flat prior is 8.15% under YCM and 7.88% under YLLM, the average CV using IG prior is 5.87% under YCM and 5.56% under YLLM. The results in Table 2 indicate that both IG and flat priors lead to similar performance of the HB estimator. However, using IG prior in both YCM and YLLM leads to smaller CV and RRMSE for the HB estimator. The results in Table 2 also demonstrate that YLLM performs slightly better than YCM in terms of CV and RRMSE. This simulation result is consistent with the results shown in You (2021).

Table 2: Comparison of average ARB%, average CV (ACV%) and RRMSE%

Specification	$\hat{\theta}_i^{HB}$	YCM		YLLM	
		IG prior	Flat prior	IG prior	Flat prior
σ_v^2=1	ARB	1.83	1.78	1.73	1.77
	ACV	12.31	12.69	12.07	12.48
	RRMSE	10.49	10.46	10.24	10.25
σ_v^2=0.5	ARB	1.15	1.14	1.14	1.16
	ACV	9.99	10.93	9.87	10.71
	RRMSE	8.76	8.87	8.62	8.78
σ_v^2=0.1	ARB	0.22	0.35	0.28	0.31
	ACV	5.87	8.15	5.56	7.88
	RRMSE	5.68	5.72	5.45	5.56

4. Data analysis

In this section, we compare YCM and YLLM using IG and flat priors through a real data application. Following Hidiroglou, Beaumont and Yung (2019) and You (2021), we apply both the YCM and YLLM to a Canadian Labour Force Survey (LFS) data and compare the HB estimates of unemployment rates with the census estimates. We apply both the YCM and YLLM to the May 2016 unemployment rate estimates for the Census Metropolitan Areas (CMAs) and Census Agglomerations (CAs), and then we compare the HB estimates and the direct estimates with the census estimates. For both the YCM and YLLM, the local area employment insurance monthly beneficiary rate is used as an auxiliary variable in the model, same as in Hidiroglou, Beaumont and Yung (2019) and You (2021). We compute the absolute relative error (ARE) of the direct and HB estimates with respect to the census estimates for each CMA/CA as follows:

$$\text{ARE}_i = \left|\frac{\theta_i^{Census} - \theta_i^{Est}}{\theta_i^{Census}}\right|,$$

where θ_i^{Est} is the direct or HB estimate and θ_i^{Census} is the corresponding census value of the LFS unemployment rate. Then we take the average of AREs over CMA/CAs. For CV, we compute the average CVs of the direct and model-based estimates. We prefer a model with smaller ARE and smaller CV. We first apply both models YCM and YLLM to all the 117 CMA/CAs with sample size ≥ 2, and then apply to 92 CMA/CAs with sample size ≥ 5, and 79 CMA/CAs with sample size ≥ 7, respectively. Table 3 presents the average ARE and the corresponding average CV (in brackets) for the YCM and YLLM using IG and flat priors.

Table 3: Comparison of average ARE and average CV (in parenthesis)

CMA/CAs	Direct LFS	YCM IG prior	YCM Flat prior	YLLM IG prior	YLLM Flat prior
Average over 117 CMA/CAs (sample size ≥ 2)	0.263 (0.329)	0.149 (0.127)	0.148 (0.136)	0.135 (0.116)	0.135 (0.123)
Average over 92 CMA/CAs (sample size ≥ 5)	0.216 (0.262)	0.132 (0.115)	0.132 (0.121)	0.126 (0.112)	0.125 (0.117)
Average over 79 CMA/CAs (sample size ≥ 7)	0.181 (0.232)	0.123 (0.112)	0.122 (0.115)	0.119 (0.109)	0.118 (0.114)

It is clear from Table 3 that both the YCM and YLLM improve the direct LFS estimates substantially by reducing the average ARE and CV, and YLLM performs better than YCM. For both the YCM and YLLM, using IG and flat priors leads to about

the same ARE. However, using IG prior in the models can lead to smaller CV as shown in Table 3. For example, for YCM, the average CV over 117 CMA/CAs is 0.127 under IG prior and 0.136 under flat prior. For YLLM, the average CV is 0.116 under IG prior and 0.123 under flat prior. Thus, in our application, for the point estimation, there is no difference using IG or flat prior. However, using IG prior in both the YCM and YLLM can lead to smaller CV. This result is consistent with the simulation result reported in Table 2.

Now we present a Bayesian model comparison using conditional predictive ordinate (CPO) for both the YCM and YLLM with IG and flat priors. CPOs are the observed likelihoods based on the cross-validation predictive density $f(y_i|y_{obs(i)})$. We compute the CPO value $\text{CPO}_i = f(y_{i,obs}|y_{obs(i)})$ for each observed data point $y_{i,obs}$, and larger CPO_i indicates a better model fit. For model choice, we can compute the CPO ratio of model A against model B. If this ratio is greater than 1, then $y_{i,obs}$ supports model A. We compute the CPO ratios for YCM IG/Flat and YLLM IG/Flat, and count the number of the CPO ratios that are larger than 1. We can also plot the CPO values or summarize the CPO values by taking the average of the estimated CPOs. For more detail on applications of CPO, see for example, Gilks, Richardson and Spiegelhalter (1996), You and Rao (2000), Molina, Nandram and Rao (2014) and You (2021). Table 4 presents the average CPO values and # of CPO ratios larger than 1 over the 117, 92 and 79 CMA/CAs for the YCM and YLLM with IG vs flat priors.

Table 4: Summary of the average CPO values and # of CPO ratios larger than 1

CMA/CAs	YCM			YLLM		
	IG prior	Flat prior	# of CPO ratio >1	IG prior	Flat prior	# of CPO ratio >1
117	0.1228	0.1222	76	0.1253	0.1242	73
92	0.1412	0.1392	61	0.1419	0.1398	59
79	0.1516	0.1491	50	0.1526	0.1517	52

It is clear from Table 4 that for both the YCM and YLLM models, IG prior has larger CPO values than flat prior, and more than half of the observations support the model with IG prior. For example, over 117 CMA/CAs, for YCM, the average CPO under IG prior is 0.1228, and 0.1222 under flat prior, and 76 observations support YCM with IG prior. For YLLM, the average CPO is 0.1253 under IG prior and 0.1242 under flat prior, and 73 observations support YLLM with IG prior. We also note that the YLLM model is better than the YCM with larger CPO values for both IG and flat priors.

5. Concluding remarks

In this paper, we have studied the performance of HB small area models using IG and flat priors on variance components through a simulation study and real data analysis. Our results indicate that both the YCM (You and Chapman 2006) and YLLM (You, 2021) models using IG and flat priors perform very well. However, using IG prior in both the YCM and YLLM leads to slightly better results (smaller CV) and better model fit. Our simulation study and real data analysis demonstrate that proper IG prior should be used in the HB small area models for variance components. Flat prior for the model variance should be avoided as using the flat prior has no advantage over the IG prior with respect to the final HB estimates. For future work, informative priors such as IG prior with parameter values based on previous survey data could also be used in the model to improve the HB small area estimators. It is also interesting to compare the HB estimators using informative priors.

Acknowledgments

I would like to thank the Editor and two referees for their careful reading of the manuscript and suggestions to improve the paper.

References

Dass, S. C., Maiti,T., Ren, H. and Sinha, S., (2012). Confidence interval estimation of small area parameters shrinking both means and variances. *Survey Methodology*, 38, pp. 173–187.

Fay, R. E., Herriot, R. A., (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, pp. 269–277.

Gelman, A., (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, pp. 515–533.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B., (2004). *Bayesian Data Analysis*. 2nd Edition, Chapman & Hall/CRC.

Ghosh, M., Myung, J. and Moura, F. A. S., (2018). Robust Bayesian small area estimation. *Survey Methodology*, 44, pp. 101–115.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Hidiroglou, M. A., Beaumont, J.-F. and Yung, W., (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, 45, pp. 101–126.

Lahiri, P., Rao, J. N. K., (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 82, pp. 758–766.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. J., (2000). WinBUGS – A Bayesian modeling framework: concepts, structure and extensibility. *Statistics and Computing*, 10, pp. 325–337.

Molina, I, Nandram, B. and Rao, J. N. K., (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach, *Annals of Applied Statistics*, 8, pp. 852–885.

Rao, J. N. K., Molina, I., (2015). *Small Area Estimation*, 2nd Edition. John Wiley & Sons, New York.

Rivest, L. P., Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, July 10–13, 2002, Ottawa, Canada.

Sugasawa, S., Tamae, H. and Kubokawa, T., (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44, pp. 150–167.

You, Y., (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, pp. 19–27.

You, Y., (2016). *Hierarchical Bayes sampling variance modeling for small area estimation based on area level models with applications*. Methodology branch working paper, ICCSMD-2016-03-E, Statistics Canada, Ottawa, Canada.

You, Y., (2021). Small area estimation using Fay-Herriot model with sampling variance smoothing and modeling. *Survey Methodology*, 47, pp. 361–370.

You, Y., Chapman, B., (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, pp. 97–103.

You, Y., Rao, J. N. K., (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, pp. 173–181.

Appendix

Full conditional distributions and sampling procedure for the YLLM model:

- $[\theta_i|y,\beta,\sigma_i^2,\sigma_v^2] \sim N\left(\gamma_i y_i + (1-\gamma_i)x_i'\beta, \gamma_i\sigma_i^2\right)$, where $\gamma_i = \frac{\sigma_v^2}{\sigma_v^2+\sigma_i^2}$, $i = 1,\ldots,m$;
- $[\beta|y,\theta,\sigma_i^2,\sigma_v^2] \sim N_p((\sum_{i=1}^m x_i\, x_i')^{-1}(\sum_{i=1}^m x_i\theta_i), \sigma_v^2(\sum_{i=1}^m x_i x_i')^{-1})$;
- $[\sigma_v^2|y,\theta,\beta,\sigma_i^2] \sim I\,G\left(a_v + \frac{m}{2}, b_v + \frac{1}{2}\sum_{i=1}^m(\theta_i - x_i'\beta)^2\right)$;
- $\left[\sigma_i^2\middle|y,\theta,\beta,\sigma_v^2,\delta,\tau^2\right] \propto f(\sigma_i^2)\cdot h(\sigma_i^2)$, where $f(\sigma_i^2)$ and $h(\sigma_i^2)$ are $f(\sigma_i^2) \sim I\,G\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2+d_i s_i^2}{2}\right)$, and $h(\sigma_i^2) = exp\left(-\frac{(log(\sigma_i^2)-z'_i\delta)^2}{2\tau^2}\right)$;
- $[\delta|y,\theta,\beta,\sigma_i^2,\sigma_v^2,\tau^2] \sim N_2\left((\sum_{i=1}^m z_i z'_i)^{-1}(\sum_{i=1}^m z_i\, log(\sigma_i^2)), \tau^2(\sum_{i=1}^m z_i z'_i)^{-1}\right)$;
- $[\tau^2|y_i,\theta\ \ ,\beta,\sigma_i^2,\sigma_v^2,\delta] \sim I\,G\left(a_\tau + \frac{m}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^m(log(\sigma_i^2) - z'_i\delta)^2\right)$.

We use Metropolis-Hastings rejection step to update σ_i^2:

(1) Draw $\sigma_i^{2^*}$ from $IG\left(\frac{d_i+1}{2}, \frac{(y_i-\theta_i)^2+d_i s_i^2}{2}\right)$;

(2) Compute the acceptance probability $\alpha\left(\sigma_i^{2^*}, \sigma_i^{2^{(k)}}\right) = min\left\{h(\sigma_i^{2^*})/h(\sigma_i^{2^{(k)}}), 1\right\}$;

(3) Generate u from Uniform(0,1), if $u < \alpha\left(\sigma_i^{2^*}, \sigma_i^{2^{(k)}}\right)$, the candidate $\sigma_i^{2^*}$ is accepted, $\sigma_i^{2^{(k+1)}} = \sigma_i^{2^*}$; otherwise $\sigma_i^{2^*}$ is rejected, and set $\sigma_i^{2^{(k+1)}} = \sigma_i^{2^{(k)}}$.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 179–190, https://doi.org/10.59170/stattrans-2023-059
Received – 09.07.2022; accepted – 17.04.2023

Application of statistical methods in socio-economic geography and spatial management based on selected scientific journals listed in the Web of Sciences database

Sławomir Dorocki[1], **Mariusz Cembruch-Nowakowski**[2]

Abstract

The aim of the paper is to present an analysis of the use of statistical methods and tools in scientific articles related to socio-economic geography and spatial management published in the years 2012–2021. In order to evaluate the intensity and diversity of this phenomenon, a query was carried out using the Web of Sciences electronic academic information database. A preliminary literature search led to the decision to focus on papers published in three selected journals relating to social geography (Geoforum), economic geography (Applied Geography) and spatial management (Landscape and Urban Planning). The paper analyses the variety of the statistical tools used in the studies presented in the aforementioned journals. The frequency and type of the applied statistical methods, computer software and computing tools is correlated with the specificity of the research area.

Key words: socio-economic geography, statistical methods, computer software, electronic academic information database.

1. Introduction

Statistical methods are frequently used by researchers as auxiliary tools for data analysis in various scientific disciplines including: exact, natural, medical, and social sciences as well as humanities. However, they are particularly useful for data collection and analysis in some sub-disciplines, such as e.g. geography including socio-economic geography and spatial management. In demography and population geography, mathematics and statistics were introduced already in the 18th century (Fleszar, 1962). However, the term "quantitative geography" appeared in the literature at the beginning of the 20th century in papers published by Wallis (1912) and Huntington (1927).

[1] Pedagogical University of Cracow, Kraków, Poland. ORCID: https://orcid.org/00000-0001-6083-0346.

[2] Pedagogical University of Cracow, Kraków &The Central Statistical Office of Poland, Culture Statistics Center, Kraków, Poland. E-mail: mariusz.cembruch-nowakowski@up.krakow.pl. ORCID: https://orcid.org/0000-0001-8443-9915.

The real revolution, which occurred at the turn of the 50s and 60s of the 20th century, resulted in a transition from descriptive geography (idiographic) to the empirical law-making (nomothetic) geography. As a consequence, more advanced statistical methods (including multidimensional ones) were used by the researchers involved in these studies (R. P Haining, R. Haining, (2003); Runge, 2006; A. Agresti, (2009); J. E. Burt et al. (2009); A. Hanushek, J. E. Jackson, (2013); R. Kitchin, (2013); O. Schabenberger, C. A. Gotway, (2017); Czyż, Chojnicki, 2019; Hauke, 2021).

Socio-economic geography and spatial management emerged as a new discipline of social sciences in 2018. The current development of social and economic geography is tightly connected with the development of statistical tools and geostatistical methods. The contemporary computing potential and availability of statistical databases (big data) supported by IT statistical tools allow the spatial interpretations of various spatial phenomena. It should be pointed out that such a research methodology usually requires the involvement of interdisciplinary research teams. It can be observed that recently such teams consisting of geographers, statisticians, and computer scientists have been created to carry out joint research in a wide area of geography.

2. Data Collection

To analyze whether and to what extent the statistical tools are currently used in socio-economic geography and spatial management the query of papers dealing with that research subject and published in the international scientific journals in the period of time 2012–2022 was performed.

Web of Science (WoS), one of the largest electronic academic information databases, was used as the source to retrieve the related publications. The publications in which the statistical methods and IT programs were used for data analysis were selected and used for further discussion.

The sample for the analysis was selected in two step procedure. In the first step, the content analysis was applied for the search in Web of Science electronic academic information database using the following key words: socio-economic geography, spatial management, statistical methods, computer software, IT programs. The search allowed to identify 30 journals (160 512 papers) matching the subject of interest. From that set of journals three of them published by Elsevier publishing house were selected: Geoforum, Applied Geography and Landscape and Urban Planning with the content of 5507 papers dealing with the problems under investigation. These journals were selected based on their scopes and reputation in scientific community (see below for details).

In the second step, the query of the papers presenting research results in the area of socio-economic geography and spatial management published in these journals in the period of time 2012-2021 was performed. That resulted in the selection of 592 papers meeting the imposed requirements. That sample was subjected to detailed analysis assuming the confidence level $\alpha = 0.95$, fraction = 0.5, maximal error 0.05%. Taking into account these parameters the sample for analysis reduced to 383 papers. However, to ensure the representativeness of the content it was decided to analyse all of the 592 papers. That resulted in increase of the maximal error to about 4%.

3. Data Selection and Analysis

The preliminary analysis indicated that the scopes of three scientific journals: Goforum, Applied Geography, and Landscape and Urban Planning are compatible with the research subjects of socio-economic geography and spatial management. In the period of time considered in the analysis (2012–2021), these journals published 5507 papers dealing with these subjects: Geoforum (1948) Landscape and Urban Planning (1832), and Applied Geography (1727). It should be pointed out that the scopes of these journals cover the representative and the most important aspects of research in geography and spatial management. Geoforum is mainly interested in papers presenting the research results of studies in social geography, including the geography of settlement and political geography. „*Geoforum is a leading international, interdisciplinary journal publishing innovative research and commentary in human geography and related fields. It is global in outlook and integrative in approach. The broad focus of Geoforum is the organisation of economic, political, social, and environmental systems through space and over time. Areas of study range from the analysis of the global political economy, through political ecology, national systems of regulation and governance, to urban and regional development, feminist, economic and urban geographies and environmental justice and resources management*"[3].

Geoforum has Cite Score 5.9, Impact Factor = 3.93, supports open access and is abstracted/indexed in 7 data bases (Current Contents, Academic Journal Guide (Chartered Association of Business Schools), Elsevier BIOBASE, Engineering Village – GEOBASE, Current Contents – Social & Behavioral Sciences, Social Sciences Citation Index and Scopus).

Applied Geography is mainly presenting papers on economic geography. "*Applied Geography is a journal devoted to the publication of research which utilizes geographic approaches (human, physical, nature-society and GIScience) to resolve human problems*

[3] Retrieved from: https://www.sciencedirect.com/journal/geoforum. 22.06.2022

that have a spatial dimension. These problems may be related to the assessment, management and allocation of the world's physical and/or human resources. The underlying rationale of the journal is that only through a clear understanding of the relevant societal, physical, and coupled natural-humans systems can we resolve such problems"[4].

Applied Geography has Cite Score equal to 8.3, Impact Factor 4.73, and supports open access. It is abstracted/indexed in 16 main data bases (Sage Public Administration Abstracts, Geography, Ecological Abstracts, Envirofiche, GeoRef, Oceanographic Literature Review, Elsevier BIOBASE, International Development Abstracts, Sage Urban Studies Abstracts, Social Sciences Citation Index, Current Contents, Current Geographical Publications, Environmental Abstracts, Environmental Periodicals Bibliography, Geographical Abstracts, Scopus).

Landscape and Urban Planning is dealing with special management. "Landscape and Urban Planning is an international journal aimed at advancing conceptual, scientific, and applied understandings of landscape in order to promote sustainable solutions for landscape change. Landscapes are visible and integrative social-ecological systems with variable spatial and temporal dimensions. Landscapes are increasingly urban in nature and ecologically and culturally sensitive to changes at local through global scales. Multiple disciplines and perspectives are required to understand landscapes and align social and ecological values to ensure the sustainability of landscapes. The journal is based on the premise that landscape science linked to planning and design can provide mutually supportive outcomes for people and nature"[5].

Landscape and Urban Planning, has Cite Score equal to 12.7, IF = 8.12, supports open access. It is abstracted/indexed in 13 main data bases: Science Citation Index, Elsevier BIOBASE, LandSearch, Engineering Village – GEOBASE, Applied Ecology Abstracts, BIOSIS Citation Index, Current Contents, Environmental Periodicals Bibliography, Geographical Abstracts, Scopus, Cambridge Scientific Abstracts, Environmental Abstracts, Urban Studies Abstracts. It can be summarized that these journals are representative of the subject under investigation. The statistical sample of 592 papers published in these journals in the period of time 2012–2021 was carefully reviewed to search for the application of statistical methods in data analysis (see Table 1). The number of papers selected from various journals is proportional to the total number of papers published in each of them. It should be noticed that Applied Geography is a quarterly journal while Geoforum is a monthly journal, which translates to the total number of papers published in each of them.

[4] Retrieved from: https://www.sciencedirect.com/science/article/abs/pii/S014362282100028X. 22.06.2022

[5] Retrieved from: https://www.sciencedirect.com/journal/landscape-and-urban-planning. 22.06.2022

Table 1: Number of analyzed papers published in the selected journals in the period of time 2012–2021

Journal/year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Sum	IF
Applied Geography	10	12	7	8	9	9	7	11	23	17	113	4.60
Geoforum	42	25	27	11	24	10	33	20	27	32	251	3.77
Landscape and Urban Planning	28	28	18	26	17	24	20	17	24	26	228	6.85
Sum	80	65	52	45	50	43	60	48	74	75	592	

Source: own studies.

The authors of 63% papers analyzed used statistical methods as research tools (Table 2). Most often, these methods were present in the works on the special management, papers published in Landscape and Urban Planning and in economic geography, papers published in Applied Geography. The share of papers in which statistical methods were used was almost ten times smaller in the case of publication in Geoforum than for the first two magazines. This is mainly due to the fact that papers published in Geoforum were of more theoretical nature or they presented research based on the survey results, which were often purely descriptive. It is noticeable that in the articles under investigation several (more than one) statistical methods were usually used; on average, about 3.5 methods were used per article. In this analytical approach, the publications in Geoforum did not stand out considerably from the average methods used in the other two inspected journals.

In the conducted study, not only the methods used were analyzed but also the way of presenting the results in the form of charts and diagrams was considered. However, not all charts, but only those used as the basis for the statistical inference were taken into consideration. These graphical methods prevailed in Applied Geography (in 69% of papers) while in Geoforum they were rarely present (in 1.6 % of papers). It should be observed that in Landscape and Urban Planning the choropleth maps and diagrams accompanying the geostatistical methods were identified. However, they were not considered in the current analysis because there are basic substantive and methodological difficulties in the expectance of their applicability in the social and political geography where they are rarely, if ever, used.

Table 2: Fraction of papers in which the statistical methods were used in the selected journals in the period of time 2012–2021

Journal	Statistical methods used (%)	The average number of methods used per one paper	Charts used (%)
Landscape and Urban Planning	88.60	3.77	59.65
Applied Geography	88.50	3.16	69.03
Geoforum	9.96	2.28	1.59
Sum	63.34	3.47	36.82

Source: own studies.

As a result of the query, the following statistical methods were distinguished: average measures (mean), measures of variation - including standard deviations (SD), positional measures, e.g. median, dominant, etc. (position), correlation-regression (R2), statistical model (model), standardization or normalization (normal). In addition, the most commonly used advanced statistical methods and programs used to carry out calculations were reviewed. It should be noticed that in some cases there was no information on the type of statistical tools used although the results presented clearly indicated that they were applied.

Table 3: Fraction of papers in which the selected statistical methods were used in selected journals in the period of time 2012–2021

Journal	r2	mean	model	SD	position	normal
Applied Geography	58.41	46.02	46.90	39.82	43.36	45.13
Geoforum	3.19	7.57	3.19	4.78	1.59	2.39
Landscape and Urban Planning	70.61	65.35	58.33	59.65	53.51	26.32
Sum	39.70	37.16	32.77	32.60	29.56	19.76

Source: own studies.

Among these statistical methods, the correlation-regression (R2) was the most frequently used (in 39.7% of analyzed papers). That method was mostly applied to analyze the economic and social phenomena in studies carried out in the area of spatial management (published in Landscape and Urban Planning) and economic geography. Interestingly, although the analyzed phenomena had spatial dimension (geographic coordinate system) the autocorrelation of variables was neglected in their analyzes (Dorocki, Jenner 2016). Thus, these methods were applied for selection as well as for observation of the relationship between variables.

The arithmetic average was in the second place amongst the statistical methods used by the authors of the papers published in analyzed journals. That simple method

was used most often in the studies in the area of social geography (papers published in Geoforum). Two others identified by our methods such as measures of differentiation our statistical models were used to a similar extent.

Surprisingly, it was observed that the frequency of using statistical methods in the papers published in the time frame 2012-2021 was not dependent on the year of publication. It was expected that in the most recent papers there will be a higher probability of using statistical tools but that assumption was not confirmed. The use of statistical methods was shown to be dependent mainly on the research subject presented but also on the quality and quantity of data collected. In the case of publications from the area of social geography (Geoforum) usually a survey and scientific observations served as a base for analysis. The analysis of this kind of data was often limited to the very description of the relative values and the survey responses obtained. The most common methods, in this case, were average and positional measures and the Cronbach Alpha test. In papers from the area of economic geography the most frequently analyzed are secondary statistical data derived from databases shared by public institutions. In such cases, the study of dependencies between statistical variables, confidence testing as well as creation and verification of statistical models were most often undertaken. In the case of papers from the area of spatial management, the studies involved mainly spatial data. Thus, the researchers applied mostly the geostatistical methods (GIS) and analyses related to environmental protection such as risks analysis and matrix methods. Also, the spatial or mixed models were considered.

Out of all advanced statistical methods used in the literature analyzed the following, most often repeated methods are listed in alphabetical order: (ANOVA), Amsterdam mode, Cronbach Alpha, Chi-2, exploratory factor analysis (EFA), Generalized Linear Mixed Models (GLMM), generalized linear models (GLM), Kolmogorov–Smirnov test, Gaussian distribution, Kruskal–Wallis test, mixed methods, Monte Carlo method, Mann Whitney U test, generalized cross-validation (GCV), Manova (Wilk's lambda), multiple linear regression (MLR), Moran I, OLS, PCA, Pearson, Spearman, multidimensional regression tree (MRT), Shannon diversity index, index Kappa. It is worth mentioning that the use of many statistical methods is also related to the accessibility of computer statistical programs. Out of the statistical software used the most popular was R Project for Statistical Computing (R), which served as an analytical tool in 53 papers, of which 43 were published in Landscape and Urban Planning. The popularity of that program comes not only from its computing capabilities but also from its high availability; it is free statistical software working in the UNIX, Windows, and in the MacOS environments. Additionally, various packages extending the possibilities of the program are delivered.

In the second place among the statistical software used was Statistical Package for the Social Sciences (SPSS). It was originally addressed mainly to the social sciences but currently it is widely used also in other scientific disciplines. One of the most popular geostatistical tool is ArcGIS software. It allows the creation and processing/modification of the existing maps, analysis of spatial data, their visualization, and data management in geodatabases. That program was used mainly by the authors publishing papers in Landscape and Urban Planning. Most often the NVivo package was used for the analysis of qualitative data obtained from surveys, e.g. text - interview content, etc. Also, the very popular MS Excel package and built-in statistical functions are utilized. As with the selection of analytical methods for a given type of data, also in the case of statistical software, there is a diversification of the programs used (Table 4).

Table 4: Share of statistical software used by the authors of papers published in selected journals in the time frame 2012–2021

Program	Applied Geography	Geoforum	Landscape and Urban Planning	sum
R	9	1	43	53
SPSS	14	5	14	33
ArcGIS	1		17	18
NVivo		5	2	7
MS excel	3	2	2	7
GeoDa	2	1	3	6
PC-ORD			6	6
ArcMap			5	5
Matlab	1		1	2
MAXQDA			2	2
MINITAB			2	2
Qgis			2	2
SigmaPlot			2	2

Source: own studies.

4. Discussion

In the current paper the application of statistical methods in studies in socio-economic geography and spatial management was presented based on the analysis of the papers published in three deliberately selected prestigious scientific journals. These journals are published by the Elsevier publication house and have high citation sores and impact factors. These can serve as evidence that they are of interest to a wide population of the researchers. As indicated above, the results of research on the subjects related to the socio-economic geography and spatial management are also presented in several other scientific journals, including some of Polish journals. Some of them are

not recognized in WoS while others, e.g. Bulletin of Geography-Socio-Economic Series, European Spatial Research and Policy, Geographia Polonica, Miscellanea Geographica, Studies of the Industrial Geography Commission of the Polish Geographical Society, Quaestiones Geographicae are indexed only recently. It could be recommended to consider, in the extension of the analysis presented within this paper, to include the contents of some of these journals in the near future. This would allow to consider the interest of many other researchers in application of statistical methods in analysis of data relevant to the socio-economic geography. It should be observed that Polish researchers, e.g. T. Czyż, W. Ratajczak, P. Czapliński, Z. Chojnicki, D. Jędrzejczyk, T. Stryjakiewicz, I. Jażdżewska, J. Hauke, (2021) belong to that group.

The sample of 592 papers published in the journals indicated below in the period of time 2012–2021 was subjected to the detailed analysis. Interestingly, the fraction of papers in which the statistical methods reached almost 89% for publications in Landscape and Urban Planning and Applied Geography while only about 10% for publications in Geoforum. The correlation-regression (R2) was the most frequently used method. It was applied to analyze data in almost 40% of papers reviewed. The simple, arithmetic average was the second most popular statistical method used by the authors of the papers published in analyzed journals. This was followed by differentiation our statistical models, which were used to a similar extent. Unexpectedly, that order of popularity of the statistical methods did not changed during the decade under consideration (2012–2021). Various types of statistical software were used as the analytical tools. R Project for Statistical Computing (R) was the most popular one. It was used as an analytical tool in the research resented in 53 papers. Interestingly, 43 of them were published in Landscape and Urban Planning. Such high frequency of using that software can be explained considering its high computing capabilities combining with availability, as it is free of charge statistical software working in UNIX, Windows, and in the MacOS environments. The second place most popular statistical software was Statistical Package for the Social Sciences (SPSS), currently widely used also in various scientific disciplines. For the spatial analysis the ArcGIS software is applied. The NVivo package is used for the analysis of qualitative data obtained from surveys while MS Excel package is helpful in the quantitative analysis.

5. Conclusions

The literature query allowed drawing a picture of differentiation and intensity of the statistical methods and statistical software application in studies carried out in the area of socio-economic geography and spatial management. Differences observed between the sub-disciplines reflect the specificity of the research practices characteristic

for them. For example, the papers dealing with the subject related to spatial management usually present the conclusions reached within the applied research projects which are carried out by large multidisciplinary, often international research teams. In that case, advanced analyzes based on statistical models and methods (often multidimensional methods and triangulation) are used. On the other hand, studies in the area of social geography are less costly and usually carried out by individual researchers. Thus, the application of less advanced and simpler analytical methods prevail, producing more general sets of results (often only as percentage values). The papers in the area of economic geography present mainly the results of secondary data analysis based on the tests and dependency research carried out with the use of statistical methods and models.

There is no doubt that conducting research with the support of statistical methods increases the credibility and reliability of their results as well as ensures the correctness of inference. This is particularly important for the analysis of spatial phenomena which is becoming more and more complex.

It can be concluded that geographers should constantly develop and improve their competencies in the area of applied statistics. Also, the involvement of professional statisticians in the development of statistical tools better suited for studies of various geographic phenomena would be highly beneficial for the discipline.

Limitations and future studies.

The conclusions presented above are based on the analysis of the representative but relatively small sample of the literature resources available. This can be considered as the limitation of the certainty of the results presented. The analysis of a larger sample and observation of possible changes which have occurred in recent years will be performed to draw a wider and more precise picture of the phenomena under observation.

References

Agresti, A., (2009). *Statistical methods for the social sciences*, Vol. 207. Upper Saddle River, NJ: Pearson Prentice Hall.

Burt, J. E., Barber, G. M. and Rigby, D. L., (2009). Elementary statistics for geographers. *Guilford Press*.

Czapliński, P., (2008). Problematyka badawcza przemysłu w geografii na tle nauk ekonomicznych, *Prace Komisji Geografii Przemysłu Polskiego Towarzystwa Geograficznego*, 11, pp. 46–52.

Czyż T., Ratajczak W., (1986). Mathematical methods in economic geography. *Concepts and Methods in Geography*, 1, pp. 99–126.

Czyż, T., (2016). Metoda wskaźnikowa w geografii społeczno-ekonomicznej. *Rozwój Regionalny i Polityka Regionalna*, (34), pp. 9–19.

Czyż, T., Chojnicki, Z., (2019). Rola poznańskiego ośrodka geograficznego w implementacji metod i modeli matematycznych w geografii społeczno-ekonomicznej. *Rozwój Regionalny i Polityka Regionalna*, (45), pp. 9–21.

Dorocki, S., Jenner, B., (2016). Recepta na nienormalność rozkładu i współzależność obserwacji z wykorzystaniem testów randomizacyjnych i testu Mantela – na przykładzie rozmieszczenia zasobów ludzkich w regionach Francji. *Prace Komisji Geografii Przemysłu Polskiego Towarzystwa Geograficznego*, 30(2), pp. 186–197. https://doi.org/10.24917/20801653.302.12.

Fleszar, M., (1962). Zarys historii geografii ekonomicznej w Polsce do 1939 r. Warszawa. *Wydawnictwa Geologiczne*.

Haining, R. P., Haining, R., (2003). Spatial data analysis: theory and practice. Cambridge university press.

Hanushek, E. A., Jackson, J. E., (2013). Statistical methods for social scientists. *Academic Press*.

Hauke J., (2021). Metody statystyczne w geografii społeczno-ekonomicznej: szkic historyczny oraz ograniczenia i korzyści stosowania w dobie cyfryzacji. *Czasopismo Geograficzne*, 92(1), 73–93. https://doi.org/10.12657/czageo-92-04.

Huntington, E., (1927). The quantitative phases of human geography. *The Scientific Monthly*, 25(4), pp. 289–305.

Jażdżewska, I., (2021). Od nauk geograficznych w kierunku nauki o geoinformacji. *Wydawnictwo Uniwersytetu Łódzkiego*.

Jędrzejczyk, D., (2000). Mathematical Method of Analysing Forms of the Rural Settlements of Benon Janowski. Miscellanea Geographica, *Regional Studies on Development*, 9(1), pp. 159–164.

Kitchin, R., (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography*, 3(3), pp. 262–267.

Runge, J., (2006). Metody badań w geografii społeczno-ekonomicznej–element metodologii, wybrane narzędzia badawcze, *Wyd. Uniwersytetu Śląskiego, Katowice*.

Schabenberger, O., Gotway, C. A., (2017). Statistical methods for spatial data analysis. *CRC press*.

Stryjakiewicz T., (2003). Nowe interpretacje starych wskaźników w geografii społeczno-ekonomicznej. [W:] H. Rogacki (red.), Problemy interpretacji wyników metod badawczych stosowanych w geografii społeczno-ekonomicznej i gospodarce przestrzennej. *Bogucki Wydawnictwo Naukowe*, Poznań, pp. 135–142.

Wallis, B. C., (1912). The Use of Formulæ in Quantitative Geography. *Geographical Journal*, pp. 175–178.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 191–196, https://doi.org/10.59170/stattrans-2023-060
Received – 20.04.2023; accepted – 10.07.2023

40th International Conference MSA'2022 joined with MASEP

Marta Małecka[1], **Artur Mikulec**[2], **Elżbieta Zalewska**[3]

The jubilee 40th International Conference MSA'2022 joined with MASEP took place on November 7–9, 2022, at the Training and Conference Center of the University of Lodz, Kopcińskiego 16/18 St. The conference was organized by the Department of Statistical Methods of the University of Lodz. This edition of the conference was organized in cooperation with the MASEP (Measurement and Assessment of Social and Economic Phenomena) conference. The conference was co-organized by: the Institute of Statistics and Demography (University of Lodz), the Department of Economic and Social Statistics (University of Lodz), the Statistics and Econometrics Committee of the Polish Academy of Sciences, the Polish Statistical Association, branch in Lodz and the Statistical Office in Lodz. The honorary patronage over the conference was taken by: Rector of the University of Lodz – prof. Elżbieta Żądzińska and the President of Statistics Poland – Dominik Rozkrut. The substantive partner of the conference was also StatSoft Polska Sp. z o. o.

The head of the Scientific Committee was prof. Czesław Domański and the head of the Organizing Committee was prof. Alina Jędrzejczak. The scientific secretaries of the conference were Elżbieta Zalewska, Marta Małecka, Artur Mikulec and Łukasz Ziarko.

The main objectives of the MSA'2022 conference were twofold: first, the presentation of the latest achievements in the field of multidimensional statistical analysis, and second, exchange of experience resulting from its application.

The 40th International Conference MSA'2022 joined with MASEP was organized in a stationary mode. The conference was attended by 179 people (including 74 employees of the Statistical Office in Lodz, who participated only on the first day of

[1] Department of Statistical Methods, University of Lodz, Poland. E-mail: marta.malecka@uni.lodz.pl. ORCID: https://orcid.org/0000-0003-4465-9811
[2] Department of Statistical Methods, University of Lodz, Poland. E-mail: artur.mikulec@uni.lodz.pl.
[3] Department of Statistical Methods, University of Lodz, Poland. E-mail: elzbieta.zalewska@uni.lodz.pl. ORCID: https://orcid.org/0000-0003-1544-300X.

the conference) from various academic centers in Poland, including Gdansk, Katowice, Cracow, Lodz, Poznan, Radom, Szczecin, Warsaw and Wroclaw, representatives of Statistics Poland, the Statistical Office in Lodz as well as guests from the Czech Republic, Italy and Ukraine. During the conference, three plenary and fourteen panel sessions were held, during which a total of 59 papers were presented.

The conference began on November 7 with an opening made by the head of the Organizing Committee, prof. Alina Jędrzejczak, Vice-Rector for External Relations of the University of Lodz, prof. Agnieszka Kurczewska, Vice-Dean for Scientific Research at the Faculty of Economics and Sociology of the University of Lodz prof. Ewa Kusideł, and the head of the Scientific Committee prof. Czeslaw Domański. Due to the jubilee edition of the MSA conference, during the opening of the conference, prof. Czesław Domański mentioned the first editions and reminded the history of the Lodz statistical conference.

After the official opening, the first plenary session took place. According to the conference tradition, the session was dedicated to the history of the Polish statistics and to memories of recently deceased statisticians. It was chaired by prof. Bronisław Ceranka, Poznan University of Life Sciences. The following papers were presented in this session:

- Integration of the statistician environment in the light of the ruby jubilee of the MSA – prof. Czesław Domański, University of Lodz, prof. Agata Szczukocka, University of Lodz,
- Tadeusz Gerstenkorn (1927–2021) – probabilist and statistician – Andrzej Łuczak, University of Lodz,
- Jan Kordos (1930–2021) – statistician, classical scientist, animator of scientific and organizational life on many levels – Czesław Domański, University of Lodz, Dominik Rozkrut, Statistics Poland, Włodzimierz Okrasa, Cardinal Stefan Wyszyński University in Warsaw,
- Tadeusz Bednarski (1949–2021) – an outstanding statistician from Wrocław, doctor of statistics at the University of California in Berkeley (USA) – Grzegorz Wyłupek, University of Wrocław.

The second plenary session was held after the historical session and was chaired by prof. Mirosław Szreder, University of Gdansk. The participants of the conference were given two invited lectures:

- Testing dependencies in time series – Marie Huskova, Charles University, Prague,
- Statistics in transition: five trends that accelerate the evolution of official statistics – Dominik Rozkrut, Statistics Poland.

The third plenary session was held on the third day of the conference, before its official closing. It was chaired by prof. Elżbieta Gołata, Poznan University of Economics and Business. The last session included two invited lectures entitled:

- Longevity dividend – multidimensional risk analysis – Grażyna Trzpiot, University of Economics in Katowice,
- Relative taxonomy method in data analysis – approaches, simulation studies, applications – Marek Walesiak, Wroclaw University of Economics and Business, Grażyna Dehnel, Poznan University of Economics and Business, Andrzej Dudek, Wroclaw University of Economics and Business.

The conference hosted a banquet on the occasion of the 80th birthday of prof. Miroslaw Krzyśko, which was held on the first day of the conference. On the same day, there was also a jubilee of the 60th anniversary of the Statistical Office in Lodz, during which Anna Luchowska (Statistical Office in Lodz) gave a paper entitled "60 years of operations of the Statistical Office in Lodz", presenting the history and key information on the activities of the Office. The co-author of the speech was Piotr Ryszard Cmela (Director of the Statistical Office in Lodz).

During the ceremony, the President of Statistics Poland, Dominik Rozkrut, presented seventy-four employees of the Statistical Office in Lodz with the decorations "For merits for the statistics of the Republic of Poland", which are the honorary distinctions awarded for special achievements in the field of statistics. The jubilee session ended with thanks from the President of Statistics Poland and the Director of the Statistical Office in Lodz for the employees' work and commitment, which greatly contributed to the achievements and prestige of the Office.

During the fourteen parallel sessions of the conference a broad area of topics was covered. The scope of the topics included, in particular, the following groups of issues:

1. Theory of statistical methods. A very wide range of topics were discussed in this group, including: multivariate measures, in particular asymmetry measures, statistical tests, the problem of optimal sample size, concept of noise according to Daniel Kahneman, the question of measuring dissimilarities, estimating multivariate stochastic volatility models, breakpoint detection, multimodel prediction, Bayesian procedures. Within the presented studied dedicated to statistical theory, there was also a line of discussions on current problems, challenges and solutions related to national censuses.
2. Macroeconomic applications. The main macroeconomic topic discussed in accordance with a number of papers presented during the conference was inflation, which is currently an important social issue. Within this topic, modern

methods of measuring inflation based on scrapped data were also discussed. Moreover, the conference debates included topics such as consumer confidence index and development measures.

3. Demographic and social issues. The range of socio-demographic issues discussed in the papers presented at the conference was very wide. In the field of demography, there were papers on current demographic problems and methods of mortality forecasting. Many papers concerned the labor market: the status of young people on the labor market, sectoral and spatial differentiation of the employment rate in the COVID-19 era, labor demand, duration of the young workers' first job. Other topics concerned income and poverty, including: modeling the distribution of income, material situation and life satisfaction, income expectations in families, poverty lines. The problem of seniors' social activity and spatial accessibility were also discussed. Moreover, the conference participants presented papers on current problems such as estimating the length of stay of foreigners in Poland or integration of immigrants in the European Union.
4. Sustainable development. Among the economic and social topics, a group of topics closely related to the problems of sustainable development and ecology was distinguished. In this regard the conference papers included the implementation of the goals of sustainable development in EU countries, sustainable energy in European countries, forecasting demand for water for households and the relationship between non-ferrous metals and the green bond market.
5. Business applications. Many papers covered the application of statistical methods to business. Presented topics included, among others, the strategic dimensions of corporate entrepreneurship, the development of organizations based on the logic of effectuation, document grouping with respect to their sentiment or the issue of using data coming from Internet sources.
6. Financial market. Topics in the field of statistical methods used in financial markets included, among others: the use of a kernel estimator to measure the extreme risk of returns on the Warsaw Stock Exchange, estimation of value at risk, liquidity measures in cryptocurrency markets, measures of insurance market integration, the use of models of artificial intelligence in commercial banks and predicting consumer insolvency.

In addition to topics closely related to statistical methods, the conference also featured a discussion on teaching statistics at universities. This discussion was based on a paper joining issues of new academic program creation, educational assessment, labor market, national committee of accreditation and econometric model.

The parallel sessions were chaired by:

November 7th

SESSION IIIA	prof. Andrzej Sokołowski (Cracow University of Economics, Collegium Humanum – Warsaw Management University)
SESSION IIIB	prof. Krzysztof Jajuga (Wroclaw University of Economics and Business)
SESSION IVA	prof. Danuta Strahl (Wroclaw University of Economics and Business, WSB University)
SESSION IVB	prof. Grażyna Dehnel, prof. UEP (Poznan University of Economics and Business)

November 8th

SESSION IA	prof. Grzegorz Kończak (University of Economics in Katowice)
SESSION IB	prof. Iwona Bąk, prof. ZUT (West Pomeranian University of Technology in Szczecin)
SESSION IIA	prof. Iwona Markowicz, prof. US (University of Szczecin)
SESSION IIB	prof. Maria Grzelak (University of Lodz)
SESSION IIIA	prof. Eugeniusz Gatnar (University of Economics in Katowice)
SESSION IIIB	prof. Elżbieta Roszko-Wójtowicz (University of Lodz)
SESSION IIIC	prof. Józef Dziechciarz (Wroclaw University of Economics and Business)

November 9th

SESSION IA	prof. Wojciech Zieliński (Warsaw University of Life Sciences)
SESSION IB	prof. Dominik Krężołek (University of Economics in Katowice)
SESSION IC	prof. Alina Jędrzejczak (University of Lodz).

A detailed list of presenting authors and topics is available at https://www.uni.Lodz.pl/msa.

The social program of the conference included a trip to the Museum of Cinematography, which was organized after the sessions on the second day of the conference. In addition to visiting the museum, the participants listened to a lecture on "Special effects in cinematography", thanks to which they learned about the most popular techniques of creating trick shots, their use in cinema and television, and followed their development over the years. The lecture was illustrated with photos and fragments of selected films using special effects. The story began with the first productions at the turn of the 19th and 20th centuries, and ended with today's films.

The conference was summed up and closed officially by the head of the Scientific Committee – prof. Czesław Domański and the head of the Organizing Committee - prof. Alina Jędrzejczak. They thanked all the participants for their active participation in the conference. They also thanked the co-organizers, partners and all institutions cooperating in the organization of the conference. They emphasized the fact that, thanks the joint efforts of participants and organizers, the jubilee edition of the conference turned out to be a great success.

Finally, all participants were invited to the 41st Conference on Multivariate Statistical Analysis (MSA'2023), which is planned to take place on November 6–8, 2023, at the Training and Conference Center of the University of Lodz.

STATISTICS IN TRANSITION new series, September 2023
Vol. 24, No. 4, pp. 197–202

About the Authors

Aidi Khaoula is a lecturer of probability and statistics, Laboratory of probability and statistics LaPS, University Badji Mokhtar, Annaba, Algeria. Her research interests include probability theory and validation testing.

Ali M. Masoom is George and Frances Ball Distinguished Professor Emeritus of Statistics and Professor Emeritus of Mathematical Sciences at Ball State University, USA. Dr. Ali has published widely and extensively in leading statistical journals in areas such as finite sampling, order statistics, inference based on optimal spacing, multivariate statistics, characterization problems, mixtures of distributions, ranking and selection, survival analysis, estimation of tail probabilities, parametric estimation, Bayesian inference, skew-symmetric distributions, and generalized distributions. He is a Fellow of the American Statistical Association USA, Royal Statistical Society, UK, Institute of Statisticians UK, and Bangladesh Academy of Sciences, and Elected Member of the International Statistical Institute, Netherlands. Dr. Ali has served as editor, associate editor, and editorial board member of a number of international statistical journals.

Alizadeh Morad is an Assistant Professor of Statistics at the Department of Statistics, Faculty of Science of the Persian Gulf University. His research interests are distribution theory and statistical inference and data analysis. Dr. Alizadeh has published more than 160 research papers in international/national journals and conferences.

Aslam Muhammad received the MSc degree in statistics from the University of the Punjab, Pakistan, and the PhD degree in statistics from the University of Wales, UK in 1996. In 1980, he started his career as a Lecturer of statistics with the University of Baluchistan, Pakistan. He has served Quaid-i-Azam University, Pakistan, about 25 years, where he was the Head of the Department of Statistics, for six years. He has 42 years of teaching and research experience at graduate level. He has supervised 18 PhD degree scholars and 145 MPhil degree students. He has more than 190 research publications in international repute journals. He is currently working as a Professor of statistics at Riphah International University, Pakistan. His research interests include Bayesian inference and mathematical statistics.

Bhatti M. Ishaq is a Professor of Finance at the School of Business and Economics, Universiti Brunei Darussalam, and holds adjunct Professorship at LaTrobe University

(Melbourne), SP Jain School of Global Management (Sydney), and Australian National, King Abdul Aziz and Minhaj Universities. With an impressive academic record, he has published above 200 articles, 11 books, and amassed 4420 citations (h-index of 33) in top-tier journals. Dr. Bhatti is a renowned teacher, recognized for his excellence in the field. His expertise lies in financial data analytics and statistics. As a book series editor (Islamic Business and Finance Series - Book Series - Routledge & CRC Press) and the Editor-in-Chief of the Journal of Statistical Theory and Application, he plays a significant role in shaping academic discourse. His work can be accessed through ORCID, ResearchGate, http://orcid.org/0000-0002-5027-7871: and google scholar https://scholar.google.com/citations?pli=1&authuser=1&user=gJhCA9MAAAAJ.

Cembruch-Nowakowski Mariusz is an Associate Professor at the Department of Geographic Education and Logistics, Institute of Law, Economics and Administration, Pedagogical University of Krakow. Simultaneously he holds the position of an expert in the Centre for Cultural Statistics, Statistical Office in Kraków, His main areas of interest include: management, entrepreneurship and innovation in tourism, particularly in the hospitality, cultural and creative sectors. He researches the role of small and medium-sized enterprises and public institutions in the functioning and development of these areas of the economy. Currently he is a member of editorial board (deputy editor) of Entrepreneurship and Education journal.

Chaudhuri Arijit is currently a honorary Visiting Professor at the Indian Statistical Institute (ISI), after serving there as a Professor and later as CSIR Emeritus Scientist. His main area of research has been sample surveys and for a brief period in reliability and life testing. He guided successfully 10 PhD students. He has published about 150 papers in peer-reviewed journals. He also published 14 books on Survey Sampling with Marcel Dekker, Taylor & Francis, CRC Press, North Holland, LAP (Germany), and Prentice Hall of India as the publishers.

Chłoń-Domińczak Agnieszka – Vice Rector for Science and the head of the Institute of Statistics and Demography at SGH. She is a country team leader for the SHARE project in Poland. Between 2009 and 2017 she led the project developing the Polish Qualifications Framework and the team working on Education and Labour Market and the Educational Research Institute in Warsaw. Member of networks: Network of independent experts on education, European Social Policy Network and between 2010 and 2017 a member of the European Qualifications Framework Advisory Group. Twice a Deputy Minister of Labour and Social Policy as well as former Director of the Department of Economic Analyses and Forecasting in the same Ministry. She was the vice president of Social Protection Committee of the European Council and chair of the Working Party on Social Policy of Employment, Labour and Social Affairs Committee

of the OECD. An author and co-author of many publications in the field of pensions and labour markets.

Dorocki Sławomir is an Associate Professor at the Department of Social and Economic Geography, Institute of Law, Economics and Administration, Pedagogical University of Krakow. His main areas of interest focuses on regionalization issues and particularly on the diversity of European expanse resulting from historical and cultural conditions. Currently he is a member of editorial board (Editor of on-line publication) of Entrepreneurship and Education journal.

Dudek Hanna is an Associate Professor at the Department of Econometrics and Statistics, Institute of Economics and Finance, Warsaw University of Life Sciences. Her main research fields include applied econometrics, measurement of material deprivation and food insecurity, multidimensional poverty analysis, and demand systems. She has published over 100 papers published in scientific journals and monographs.

Eftekharian Abbas is an Assistant Professor at the Department of Statistics, Faculty of Science, University of Hormozgan. His main areas of interest include: ordered data, ranked set sampling, statistical inference, non-parametric estimation.

Ibrahim Mohamed is an Associate Professor of Applied Statistics and Mathematics, Department of Applied, Mathematical and Actuarial Statistics, Faculty of Commerce, Damietta University, Damietta, Egypt. He has an MSc in Applied Statistics, 2009, and a PhD in Applied Statistics, 2015 from Mansoura University, Egypt. He has a total of 65 publications and total citations: 1060 (Google Scholar), h-index: 21 (Google Scholar), h-index: 14 (Scopus) and i10-index: 32 (Google Scholar). His research interests include probability theory, continuous distributions, discrete distributions, continuous G families, discrete G families, Bayesian analysis, semi-parametric, parametric, nonparametric regression and new goodness-of-fit tests. Dr. Ibrahim has served as Reviewer member of several international statistical journals.

Kharazmi Omid is an Associate Professor in the Department of Statistics at Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran. He received his BS degree in Statistics from Shahid Bahonar University, Kerman, Iran, in 2007, and completed his MS and PhD degrees in Statistics at the University of Isfahan, Iran, in 2009 and 2019, respectively. He has actively collaborated with researchers both locally and internationally, fostering a global exchange of knowledge and ideas. His research findings have been disseminated in reputable journals, contributing to the advancement of statistical science. His research interests include applied probability modelling, reliability theory, information theory, data science, and Bayesian analysis.

Kiani Sania Khawar is an Assistant Professor at the Department of Rehabilitation and Allied health sciences, Riphah International University. She received her MPHIL

degree in Statistics from Quaid I Azam University, Islamabad Pakistan in 2012. She started her career as Lecturer in 2013. She has more than 10 years teaching and research experience at graduate and post graduate level. She gives her expertise in statistical analysis of rehabilitation research projects. Simultaneously, she is providing statistical consultancy in Journal Riphah College of Rehabilitation Sciences. She has published about 15 articles. Her research interests include Bayesian inference and biostatistics.

Landmesser-Rusek Joanna is an Associate Professor at the Department of Econometrics and Statistics, Institute of Economics and Finance, Warsaw University of Life Sciences. Her main research interests focus on microeconometric modelling: counterfactual scenarios analysis, decomposition of income inequalities, hazard models, and multidimensional poverty. She has published more than 80 research papers in scientific journals and three monographs.

Małecka Marta has graduated from University of Lodz, Poland, where she obtained degrees in three study programmes run at two faculties: Faculty of Economics and Sociology and Faculty of Mathematics and Informatics. In 2014 she obtained PhD degree in Economics at the University of Lodz. She was granted the Award of the Polish Financial Supervision Authority for the doctoral thesis in finance in 2015. In years 2014-2018 she was a coordinator of the National Science Centre project „Hypothesis Testing in Market Risk Evaluation". She is a member of the Polish Statistical Association. For over 10 years her academic writing has explored various aspects of statistical testing in finance. Current areas of focus include market risk management, extremal risk measures and testing risk models.

Mikulec Artur is an Assistant Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. Simultaneously he holds the position of a deputy director at the Statistical Office in Lodz. His main areas of interest include: linear ordering methods, cluster analysis, duration analysis of enterprises and application of statistical methods in biological sciences.

Olanrewaju Rasaki Olawale is a Research Associate at the Department of Business Analytics and Value Networks (BAVNs), Africa Business School (ABS), Mohammed VI Polytechnic University (UM6P). He holds a Doctor of Philosophy in Mathematics (Statistics option) from the Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI). He has received his Master of Science (Proceed to PhD grade), Bachelor of Science (First Class Honor), and Professional Diploma in Statistics (Distinction), all in Statistics from the prestigious University of Ibadan. His research interests are time series analysis, statistical modelling, Bayesian methods, statistical inference, machine learning, mathematical statistics, stochastic processes, econometric, high-dimensional analysis, and data analysis in particular. Rasaki has published more than thirty-eight (38) research publications in peer-reviewed

international/national journals and conferences. He is an active member of many scientific and statistical professional bodies.

Olanrewaju Sodiq Adejare is currently rounding-up his Bachelor of Science in Statistics at the Department of Statistics, Faculty of Science, University of Ibadan. His research interests entail time series analysis, statistical modelling, Bayesian methods, statistical inference, machine learning. Sodiq has published seven (7) research publications in peer-reviewed international/national journals.

Omodolapo Waliyat Isamot is currently a research assistant at the Department of Epidemiology and Medical Statistics, University College Hospital. She received her Master of Science, Bachelor of Science, and National Diploma in Statistics from University College Hospital, University of Ibadan and Federal School of Statistics. Her main areas of interest include biostatistics, medical statistics, time series analysis, and sample surveys.

Pal Sanghamitra is in the Department of Statistics, West Bengal State University, India. Her main areas of interest include Survey Sampling: Development of theories in unequal probability sampling, randomized response techniques, adaptive cluster sampling, and small area estimation. She is an active reviewer of many journals.

Panichkitkosolkul Wararit is an Associate Professor at the Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University. Simultaneously, he holds the position of an editor-in-chief of Thailand Statistician: Journal of the Thai Statistical Association. His main areas of interest include: statistical inference, probability theory and time series analysis. Currently, he is a member of two editorial boards: Thailand Statistician and Science and Technology Asia.

Patra Dipika is in the Department of Statistics, Seth Anandram Jaipuria College, India. Her main areas of interest include survey sampling, randomized response techniques, adaptive cluster sampling, Kalman filtering.

Priyanka Kumari is a Full Professor of Mathematics. Her research interests are sampling theory, statistical inference, sensitive estimation theory and missing data analysis. Professor Kumari Priyanka has published more than 44 research papers in peer-reviewed international journals of repute, delivered many invited talks, and participated in many conferences/seminars. She has authored a book with reputed publisher and is also serving as referee/reviewer of several SCI indexed International Journals. She is the life-time member of many academic societies/associations and is also associated with many research groups in India and abroad.

Ptak-Chmielewska Aneta is an Associate Professor at the Institute of Statistical and Demography, Collegium of Economic Analysis, Warsaw School of Economics. Simultaneously, she holds the position of Credit Risk Model Validation Department

Lead in ING Tech Poland in Warsaw. Her main areas of interest include: business demography, enterprises survival and bankruptcy measurement, data mining and machine learning techniques. Currently, she is a member of Scientific Council of Economy and Finance Discipline in Warsaw School of Economics.

Ranjbar Vahid is an Assistant Professor of Statistics at the Department of Statistics, Faculty of Science of the University of Golestan. His research interests are distribution theory, censored data analysis, statistical inference and data analysis. Dr. Ranjbar has published more than 60 research papers in international/national journals and conferences.

Trisandhya Pidugu is an Assistant Professor of Mathematics. Her research interests are Sampling theory, Statistical Inference and Sensitive estimation theory. Dr. Pidugu Trisandhya has published 13 research papers in peer-reviewed international journals of repute and participated in many conferences/seminars.

You Yong is a Senior Research Methodologist at Statistics Canada. His research interest is small area estimation and Bayesian inference for survey data analysis. Dr. You published dozens of research papers in international/national journals and conferences. Dr. You won the 2021 Tom Symons Research Award of Statistics Canada for the recognition of the publication of outstanding research work.

Yousof. Haitham M. is an Assistant Professor of Statistics, Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Benha University, Egypt. He has an MSc in Applied Statistics, 2011, and a PhD in Applied Statistics, 2015 from Benha University, Egypt. He has a total of 230 publications and total citations: 6002 (Google Scholar), h-index: 46 (Google Scholar), h-index: 30 (Scopus) and i10-index: 134 (Google Scholar). His research interests include Probability theory, continuous distributions, discrete distributions, continuous G families, discrete G families, Bayesian analysis, semi-parametric, parametric, nonparametric regression and new goodness-of-fit tests. Dr. Yousof has served as an editor, associate editor, and editorial board member of several international statistical journals. Dr. Yousof has a total of 18 years of experience as a practitioner and teacher of Statistics.

Żebrowska-Suchodolska Dorota is a doctor of economic sciences in the discipline of finance. She works at the Warsaw University of Life Sciences. She is the author and co-author of several dozen scientific articles, chapters in monographs and one monograph. Her scientific interests focus on the information efficiency of the capital market, the efficiency of investment fund management, multidimensional data analysis, and the use of quantitative methods in the capital market.

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 CAPITALS**. 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.**, **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).