Volume 25, Number 3, September 2024 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

ISSN 1234-7655

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 2024*
Vol. 25, No. 3, 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 2024*
Vol. 25, No. 3, pp. V–VI

Policy Statement

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

***

Abstracting and Indexing Databases

Statistics in Transition new series is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

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

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Electronic Journals Library

Elsevier – Scopus

Genamics JournalSeek

Google Scholar

Index Copernicus

J-Gate

JournalGuide

JournalTOCs

Keepers Registry

MIAR

Microsoft Academic

OpenAIRE

ProQuest – Summon

Publons

QOAM (Quality Open Access Market)

ReadCube

RePec

SCImago Journal & Country Rank

TDNet

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

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. VII–XI

From the Editor

It is with great pleasure that we present our readers with the September issue consisting of 12 articles arranged, as usual, in three sections: Research papers, Other articles, and Research communicates and letters. A wide spectrum of topics is discussed in these papers by 25 authors from a large group of countries: Czechia, Poland, India, Cameroon, Brazil, Jordan, Saudi Arabia, Nigeria, Iran, and Botswana.

Research articles

The issue starts with the paper by **Kamila Hasilová**, **Ivana Horová**, **David Vališ**, and **Stanislav Zámecník,** entitled ***A comprehensive exploration of complete cross-validation for circular data.*** The aim of the article is to propose a novel circular-specific method that is based on a crossvalidation procedure with a von Mises density used as a kernel function. Using simulated data as well as real-world circular data sets, the authors evaluate and validate the proposed method and compare it with the existing methods. This method extends the estimate of the MISE and has better theoretical properties than the LSCV. From the presented results and outcomes it was concluded that the CCV is applicable to various data types with a respective success rate. From the data-driven method of bandwidth selection, the authors focused only on the cross-validation methods which target the MISE to have the consistent group of methods to compare.

Adam Szulc's paper ***Reconstruction of the social cash transfers system in Poland and household wellbeing: 2015–2018 evidence*** examines the impact of changes in the social benefits system on the wellbeing, poverty, and economic activity in Poland. The core element of those changes was a programme of large cash transfers, referred to as Family 500+, introduced in 2016. It was intended to support families with children, especially the least affluent ones, and to foster fertility. The impact of the transfers is examined through the observation of changes in the monetary and multidimensional wellbeing of households. The study also analyzed the changes in recipients' economic activity using estimates of regression models and treatment effects. The Family 500+ programme proved to be successful as an anti-poverty tool and also resulted in the increase in the average wellbeing for the whole population.

In the paper ***On autoregressive processes with Lindley-distributed innovations: modeling and simulation*** **K. U. Nitha** and **S. D. Krishnarani** develop an autoregressive

process of order one, assuming that the innovation random variable has a Lindley distribution. The key properties of the process under investigation embrace five distinct estimation techniques employed to estimate the respective parameters. Parametric and non-parametric estimating techniques are effectively employed. The authors explored a first order autoregressive model with the Lindley error distribution and its properties. The stationarity of the process is tested using a unit root test. The application of the proposed process to the analysis of time series data is demonstrated using real data sets. Based on some important statistical measures, the analysis of the data sets reveals that the proposed model fits well, and the errors are independent and Lindley-distributed. The stationary series of additive autoregressive models could feature non-Gaussian errors and marginal.

In the next article, ***Comparing logistic regression and neural networks for predicting skewed credit score: a LIME-based explainability approach,*** **Jane Wangui Wanjohi**, **Berthine Nyunga Mpinda**, and **Olushina Olawale Awe** compare the predictive ability of Logistic Regression (LR) and a Multilayer Perceptron (MLP) using two types of data sets, with an advanced model explainability technique - Local Interpretable Model-Agnostic Explanations (LIME). The findings show that all models performed better after the data were balanced. MLP had higher scores than LR in terms of balanced accuracy, Matthews correlation coefficient, and F1 score. From the findings, this study recommends that lending companies with small amounts of data use a logistic regression model but for companies with vast amounts of data a multilayer perceptron will ease their credit offer processes. The study also highlights the importance of using explainable artificial intelligence. With the LIME explanation approach, authors were able to see how each feature influences the predicted class of a model for a given instance.

Anna Czapkiewicz's and **Katarzyna Brzozowska-Rup's** paper entitled ***The Measurement of the Gross Domestic Product affected by the shadow economy*** presents a method for balancing Gross Domestic Product (GDP) when the measurements of its components are distorted by the existence of the shadow economy using a multiple ultrastructural model (MUM), where the explanatory variables are subject to error. The expected value of GDP can be divided into two parts: the first part concerns data related to registered activities and the second part concerns unobserved data which may be partly related to the shadow economy. The empirical analysis utilizes the annual data of the Local Data Bank, for years 2000–2019. The results show that the unobservable part of the variables necessary to balance GDP on the production side does not exceed 1% of GDP, and on the expenditure side, it mostly reaches about 3% of GDP.

In the paper entitled ***Extropy and entropy estimation based on progressive Type-I interval censoring*** **Huda H. Qubbaj**, **Husam A. Bayoud**, and **Hisham M. Hilow** discuss the problem of estimation of the extropy and entropy measures based on

progressive Type-I interval censoring samples. Nonparametric-based methods involving moments and linear approximations have been proposed to this aim. The performance of the proposed estimates have been studied via simulation studies and real data sets considering various censoring schemes and three probability distributions: uniform, exponential and normal distributions. The proposed estimates of the extropy and entropy measures shown to be affected by the sample size, censoring schemes and the type of distribution. Yet, the estimates based on linear approximation outperform the other estimate in the majority of cases under study.

The paper by **R. R. Sinha, Bharti,** ***Improved estimation of the mean through regressed exponential estimators based on sub-sampling non-respondents***, discusses the issue of estimating the population mean and presents and improved regressed exponential estimators using different parameters of an auxiliary character based on sub-sampling of non-respondents. The mean square error (MSE) of the proposed estimators for the most pragmatic simple random sampling without replacement (SRSWOR) scheme have been derived up to the first order of approximation (i.e. the expression containing errors up to the power of two so that the expectation comes only in terms of the mean, variance and covariance). The optimum value of the MSE of the estimators is found, along with the necessary conditions for optimizing the MSE. The effectiveness of the suggested estimators, outperforming the existing ones in terms of their MSE, has been studied theoretically, while the empirical illustration using the simulation studies have confirmed these findings.

Idowu Oluwasayo Ayodeji's paper, ***Forecasts of the mortality risk of COVID-19 using the Markov-switching autoregressive model: a case study of Nigeria (2020–2022)***, discusses some aspects of the global pandemic due to SARS-Cov-2 and attempts to predict future occurrences of such cases in order to prevent or combat effectively consequences of the virus. This study modeled daily fatality rate in Nigeria from March 23, 2020 to March 19, 2022 and forecasts future occurrences using Markov switching model (MSM). It revealed that as of 19th March, 2022, Nigeria remained at the low-risk regime in which 1 (95%CI: 0, 1) person, on the average, died of coronavirus daily; however, the most probable scenario in the nearest future was the medium-risk state in which an average of 4 persons would die daily with 48.7% probability. The study concluded that Nigerian COVID mortality risks followed a switching pattern which fluctuated within low-, medium- and high-risks; however, the medium-risk state was most likely in the future. The results indicated that the quarantine measures adopted by the governments yielded positive results. It also underscored the need for governments and individuals to intensify efforts to ensure that the country remained at the low-risk zone until the virus would be eventually eradicated.

In the paper ***Nonparametric Bayesian optimal designs for Unit Exponential regression model with respect to prior processes (with the truncated normal as the base***

measure) **Anita Abdollahi Nanvapisheh**, **Soleiman Khazaei**, and **Habib Jafari** present a Bayesian optimal design by utilizing the Dirichlet process as *a prior*. The Dirichlet process serves as a fundamental tool in the exploration of Nonparametric Bayesian inference, offering multiple representations that are well-suited for application. Authors introduce a novel one-parameter model, referred to as the 'Unit-Exponential distribution', specifically designed for the unit interval. Additionally, a stick-breaking representation to approximate the D-optimality criterion considering the Dirichlet process as a functional tool was employed. This approach allows to explore and evaluate the performance of the nonparametric Bayesian optimal design under varying levels of concentration parameter α. The empirical results reveal interesting findings: for small parameter values, there are no two-point designs observed.

Jacek Białek's paper ***The use of the Bennet indicators and their transitive versions for scanner data analysis*** revises the price and quantity Bennet indicators and their multilateral versions for the analysis of scanner data. Specifically, instead of considering comparisons across firms, countries or regions, the transitive versions of the Bennet indicators are adapted to work on scanner data sets observed over a fixed time window. Since the scanner data sets have a high turnover of products, which can make it difficult to interpret the difference in sales values over the compared time periods, the paper also considers a matched sample approach. One of the objectives of the study is to compare bilateral and multilateral Bennet indicator results across all available products or strictly matched products over time. It also examines the impact of data filters used and the level of data aggregation on the price and quantity.

Other articles

XXXXI Multivariate Statistical Analysis 2023, Lodz, Poland. Conference Papers

The article by **Adam Idczak** and **Jerzy Korzeniewski,** ***Language independent algorithm for clustering text documents with respect to their sentiment,*** presents a novel unsupervised algorithm for documents written in any language using documents written in Polish as an example. The clustering of Polish language texts with respect to their sentiment is poorly developed in the literature on the subject. The novelty of the proposed algorithm involves the abandonment of stoplists and lemmatisation. Instead, the authors propose translating all documents into English and performing a two-stage document grouping. In the first step of the algorithm, selected documents are assigned to a class of positive or negative documents based on a set of lexical and grammatical rules as well as a set of key-terms. Key-terms do not have to be entered by the user, the algorithm finds them. In the second step, the remaining documents are attached to one of the classes according to the rules based on the vocabulary found in the documents grouped in the first step. The algorithm was tested on three corpora of documents and achieved very good results.

Research Communicates and Letters

In this section, the paper by **R. Sivasamy,** ***A finite state Markovian queue to let in impatient customers only during K-vacations,*** investigates a matrix analysis study for a single-server Markovian queue with finite capacity. During the vacation periods of the server, every customer becomes impatient and leaves the queues. If the server detects that the system is idle during service startup, the server rests. If the vacation server finds a customer after the vacation ends, the server immediately returns to serve the customer. Otherwise, the server takes consecutive vacations until the server takes a maximum number of vacation periods, e.g. K, after which the server is idle and waits to serve the next arrival. During vacation, customers often lose patience and opt for scheduled deadlines independently. If the customer's service is not terminated before the customer's timer expires, the customer is removed from the queue and will not return. Matrix analysis provides a computational form for a balanced queue length distribution and several other performance metrics.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, September 2024
Vol. 24, No. 3, pp. 1–12, https://doi.org/10.59170/stattrans-2024–024
Received – 14.08.2023; accepted – 05.06.2024

A comprehensive exploration of complete cross-validation for circular data

Kamila Hasilová[1], **Ivana Horová**[2], **David Vališ**[3], **Stanislav Zámečník**[4]

Abstract

Kernel density estimation of circular data has recently received considerable attention for its ability to model and analyse distributions on unit circles and other periodic domains. Our aim is to contribute to the literature on data-driven bandwidth selectors in circular kernel density estimation. We propose a novel circular-specific method that is based on a cross-validation procedure with a von Mises density used as a kernel function. Using simulated data as well as real-world circular datasets, we evaluate and validate the proposed method and compare it with the existing methods.

Key words: circular data, kernel density estimation, von Mises density, cross-validation method.

1. Introduction

The analysis of data in terms of directions in a plane/space, or equivalently in terms of positions of points on a circle/sphere, is required in many content areas in the Earth sciences, astrophysics, social sciences, and other fields. This type of data – known as circular data – is generally defined by the main circular measuring devices – the compass and the clock. Wind directions, animal movements, biological and social rhythms, times of occurrence of an event are illustrative examples of such observations (Mardia and Jupp, 2000; Ley and Verdebout, 2017).

The treatment of these data can also be realised by means of kernel smoothing methods. Kernel smoothing belongs to a category of nonparametric curve estimation techniques. The aim of kernel smoothing is, i.a., to estimate the entire probability density function with as few assumptions about the density as possible (Wand and Jones, 1995).

The main issue here is bandwidth selection. Among the earliest fully automatic and data-driven bandwidth selection methods are those based on cross-validation ideas (see, e.g. Rudemo, 1982; Bowman, 1984; Silverman, 1986; Scott, 1992). The main idea of these methods is a natural approximation to the mean integrated square error.

[1]Department of Quantitative Methods, Faculty of Military Leadership, University of Defence, Brno, Czech Republic. E-mail: kamila.hasilova@unob.cz. ORCID: https://orcid.org/0000-0003-1540-3489.

[2]Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno, Czech Republic. E-mail: horova@math.muni.cz.

[3]Department of Combat and Special Vehicles, Faculty of Military Technology, University of Defence, Brno, Czech Republic. E-mail: david.valis@unob.cz.

[4]Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno, Czech Republic. E-mail: zamecnik.stanislav@gmail.com.

Jones and Kappenman (1991) gave an overview of bandwidth selection methods: least square cross-validation, biased cross-validation, plug-in, and presmoothed cross-validation for linear data. Further, they proposed a new type of cross-validation. The new method is called complete cross-validation. This procedure estimates the entire mean integrated square error as opposed to least square cross-validation estimation, which omits the part depending on the unknown density.

The main point of this paper is to apply a modification of the above mentioned method to the circular data. It is not a straightforward task because circular data are fundamentally different from linear data, there is no true zero, any classification of low or high values is arbitrary. In addition, the periodic nature of the circular data complicates their analysis, as standard methods for linear data in Euclidean space are inappropriate for circular data analysis.

2. Circular kernel density estimation

Let $\Theta_1,\dots,\Theta_n \in [0,2\pi)$ be a random sample of angles drawn from a distribution with a density function f. The kernel estimate of this density at a point (angle) θ is defined as

$$\widehat{f}(\theta,\nu) = \frac{1}{n}\sum_{i=1}^{n} K_\nu(\theta-\Theta_i), \tag{1}$$

where $K_\nu(\cdot)$ is a circular kernel function and $\nu > 0$ is a concentration parameter playing the role of a smoothing parameter. For parameter ν it is required that $\nu \to \infty$, $\sqrt{\nu}n^{-1} \to 0$ for $n \to \infty$ (Oliveira et al., 2012; Tsuruta and Sagae, 2017).

As a circular kernel, the von Mises density function can be considered. The von Mises distribution, $vM(\mu,\kappa)$, is a symmetric unimodal distribution with a mean direction, $\mu \in [0,2\pi)$, and a concentration parameter, $\kappa > 0$. Its density function takes the following form

$$g(\theta;\mu,\kappa) = \left[2\pi I_0(\kappa)\right]^{-1} \exp(\kappa\cos(\theta-\mu)),$$

where $I_0(\kappa)$ is the modified Bessel function of order zero (Jammalamadaka and SenGupta, 2001).

A critical issue in the application of this estimator in practice is the selection of the smoothing parameter which controls the smoothness of the estimate. Small values of the smoothing parameter imply less concentrated data and provide oversmoothed estimates. On the other hand, large values result in undersmoothed estimates, which may reveal local structure in the data.

Data-driven procedures for selecting ν are usually based on error performance measures. One of these is the bandwidth minimising the mean integrated square error (MISE) of $\widehat{f}$, which is defined as

$$\mathrm{MISE}(\nu) = E\int_0^{2\pi} \left[\widehat{f}(\theta,\nu) - f(\theta)\right]^2 \mathrm{d}\theta.$$

Denoting $R(w,z) = E\int_0^{2\pi} w\cdot z\,\mathrm{d}\theta$ and $R(z) = R(z,z)$, we can write

$$\mathrm{MISE}(\nu) = R(\widehat{f}) - 2R(\widehat{f},f) + R(f). \tag{2}$$

3. Bandwidth selection

The above expression does not give a practical procedure for choosing the optimal smoothing parameter, because some of the MISE terms depend on the unknown density f. However, one can estimate these f-dependent quantities by those that include $\widehat{f}$ and then select the optimal bandwidth (Hall and Marron, 1987).

In the following computations, we restrict ourselves to using the von Mises density as the kernel function, i.e. $K_\nu(\theta) = \left[2\pi I_0(\nu)\right]^{-1} \cdot \exp(\nu\cos\theta)$, since it plays a similar role in the circular case as the Gaussian density does in the linear case.

Let us denote the ratio of the Bessel functions as $A_k(\nu) = I_k(\nu)/I_0(\nu)$ and the difference between the sample values as $\Theta_{ij} = \Theta_i - \Theta_j$. This notation is useful for the derivations and calculations in the following sections.

3.1. Least square cross-validation

The least square cross-validation selector aims to minimise (2) with the term $R(f)$ dropped. It is defined as the minimum of the function

$$\mathrm{LSCV}(\nu) = R(\widehat{f}) - \frac{2}{n}\sum_{i=1}^{n}\widehat{f}_{-i}(\Theta_i,\nu),$$

where $\widehat{f}_{-i}$ denotes the estimate (1) with the i-th observation omitted

$$\widehat{f}_{-i}(\Theta_i,\nu) = \frac{1}{n-1}\sum_{\substack{i=1\\ j\neq i}}^{n} K_\nu(\Theta_{ij}), \quad i = 1,\ldots,n.$$

Rewriting the formulas using a kernel function, the LSCV objective function takes the form (where $*$ denotes the convolution)

$$\mathrm{LSCV}(\nu) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_\nu * K_\nu)(\Theta_{ij}) - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} K_\nu(\Theta_{ij}). \tag{3}$$

It is possible to prove that (3) is an unbiased estimator of the quantity $\mathrm{MISE}(\nu) - R(f)$ (Hall et al., 1987).

3.2. Complete cross-validation

Jones and Kappenman (1991) investigated a class of data-driven bandwidth selection procedures for linear kernel density estimation. They proposed a new method (from the family of cross-validation methods) called complete cross-validation (CCV). Their proposal estimates the entire MISE function, as opposed to the LSCV method, which targets the difference $\mathrm{MISE}(\nu) - R(f)$.

Here, we present an adaptation of the CCV method to the circular case. Let us defined functionals T_m in a similar manner as those specified by Jones and Kappenman (1991):

$$T_m(\nu) = (-1)^m [n(n-1)]^{-1} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_\nu^{(2m)}(\Theta_{ij}),$$

where $K_\nu^{(2m)}$ is the $(2m)$-th derivative of K_ν. Then, we define the CCV function as

$$\mathrm{CCV}(\nu) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (K_\nu * K_\nu)(\Theta_{ij}) - T_0(\nu) + \frac{A_1(\nu)}{2\nu} T_1(\nu) + \frac{2A_1^2(\nu) - A_2(\nu)}{8\nu^2} T_2(\nu).$$

Lemma. $E[\mathrm{CCV}(\nu)] = \mathrm{MISE}(\nu) + o(\nu^{-2})$.

Proof. See the Appendix. □

4. Small sample behaviour

We carried out a small simulation study to compare the proposed CCV method and known LSCV method with eight simulation scenarios, see Figure 1 and Table 1.

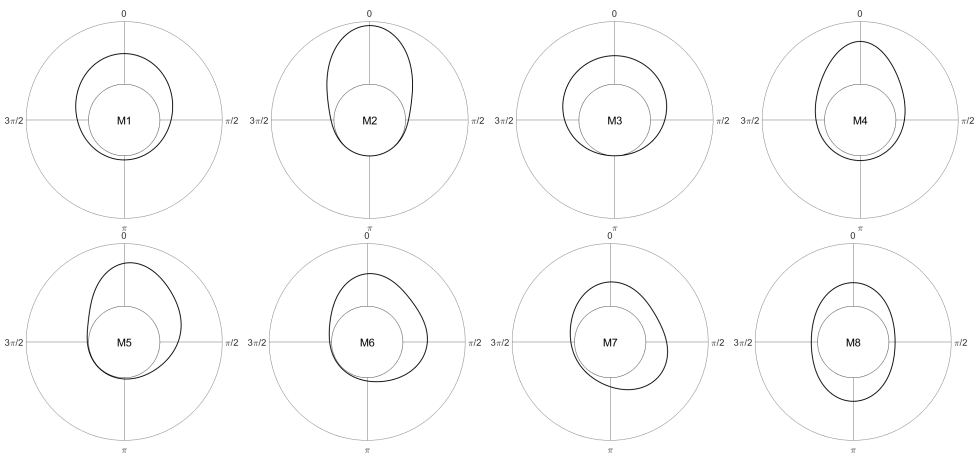


Figure 1: Graphical representation of models M1–M8

Table 1: Probability distributions and their parameters: von Mises (vM), cardioid (C), wrapped Cauchy (WC) and mixtures of two von Mises distributions

model	distribution	model	distribution
M1	$vM(0;1)$	M5	$0.5vM(0;4)+0.5vM(\pi/3;2)$
M2	$vM(0;2.5)$	M6	$0.5vM(0;3)+0.5vM(\pi/2;2)$
M3	$C(0;0.5)$	M7	$0.5vM(0;2)+0.5vM(2\pi/3;2)$
M4	$WC(0;0.5)$	M8	$0.5vM(0;2)+0.5vM(\pi;2)$

Table 2: Mean and standard deviation of ISE for all the simulation scenarios

model	method	$n = 50$	$n = 100$	$n = 200$
M1	LSCV	0.0150 (0.0226)	0.0083 (0.0096)	0.0046 (0.0046)
	CCV	0.0302 (0.0081)	0.0279 (0.0058)	0.0262 (0.0059)
M2	LSCV	0.0233 (0.0231)	0.0143 (0.0141)	0.0087 (0.0077)
	CCV	0.1197 (0.0670)	0.0904 (0.0770)	0.0525 (0.0717)
M3	LSCV	0.0135 (0.0215)	0.0076 (0.0093)	0.0044 (0.0051)
	CCV	0.0319 (0.0079)	0.0298 (0.0057)	0.0281 (0.0061)
M4	LSCV	0.0215 (0.0243)	0.0127 (0.0122)	0.0069 (0.0054)
	CCV	0.0545 (0.0165)	0.0491 (0.0199)	0.0405 (0.0251)
M5	LSCV	0.0218 (0.0272)	0.0125 (0.0132)	0.0073 (0.0071)
	CCV	0.0726 (0.0207)	0.0669 (0.0272)	0.0648 (0.0289)
M6	LSCV	0.0182 (0.0241)	0.0101 (0.0102)	0.0062 (0.0055)
	CCV	0.0415 (0.0100)	0.0388 (0.0108)	0.0363 (0.0123)
M7	LSCV	0.0173 (0.0201)	0.0093 (0.0082)	0.0057 (0.0045)
	CCV	0.0281 (0.0077)	0.0253 (0.0071)	0.0220 (0.0084)
M8	LSCV	0.0208 (0.0194)	0.0109 (0.0089)	0.0063 (0.0043)
	CCV	0.0212 (0.0097)	0.0143 (0.0105)	0.0074 (0.0075)

Taking f to be a von Mises distribution (M1, M2), cardioid distribution (M3), wrapped Cauchy distribution (M4), and mixture of two von Mises distributions (M5–M8), we generated 200 random samples of size $n = 50$, $n = 100$, and $n = 200$.

The performance and comparison of the presented circular density estimators, i.e. LSCV and CCV, is assessed by the integrated square error, $\text{ISE} = \int (\widehat{f} - f)^2 \, \mathrm{d}\theta$, which was calculated numerically using the trapezoidal rule with 500 points. The mean values of ISE (with standard deviations) are summarised in Table 2.

Even though the CCV method has better theoretical properties, the results show that when applied to simulated data, the ISE error is greater than for the LSCV method. This is due to the fact that the CCV function comes from the family of cross-validation functions based on the estimate of MISE. Cross-validation methods tend to underestimate the resulting function in the linear case, which, given the nature of the smoothing parameter, leads to an oversmoothed estimate in the circular case.

Sometimes, cross-validation objective functions have more than one local minimum (Wand and Jones, 1995), see Figure 2, where we denote these two possible outcomes as Type I (with one minimum) and Type II (with two minima). For Type II objective functions, ν_{CCV} should be taken to be the 'largest' local minimiser of CCV.

In Table 3, we summarise the percentage of the Type I CCV objective functions in the simulation scenarios. Type II is more common for data sets larger in size and more concentrated (see models M2, M4, and M8).

Comparing the bandwidth parameters of the two methods, we can see that the ratio $\nu_{\text{CCV}}/\nu_{\text{LSCV}}$ ranges on average from 0.3 to 0.75. Although this ratio appears to be quite large and may imply that the LSCV estimate is undersmoothed (or similarly, the CCV estimate is oversmoothed), we can obtain decent results, as can be seen in Figure 3, which shows selected random samples of size $n = 200$ from the model M5 and their kernel density estimates with the bandwidth chosen by both methods. These methods ($\nu_{\text{LSCV}} = 27.81$, $\nu_{\text{CCV}} = 13.34$) provide similar results for sample (a). For sample (b), the LSCV method

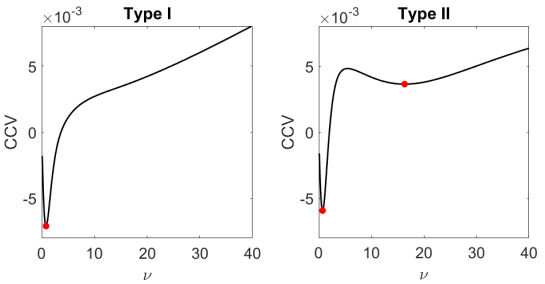


Figure 2: Two types of CCV objective functions

Table 3: Percentage of CCV function type I for the simulation models

model	$n = 50$	$n = 100$	$n = 200$
M1	93.5	93	91.5
M2	69.5	52.5	29.5
M3	92	93.5	91.5
M4	85.5	80	67
M5	82.5	80	79.5
M6	87	87.5	85
M7	89	86.5	81
M8	93.5	80.5	68

gives a very good estimate ($\nu_{\mathrm{LSCV}} = 9.52$), but the CCV method gives an oversmoothed estimate ($\nu_{\mathrm{CCV}} = 0.80$). However, the LSCV is not always an optimal method, as we can see in the graph of sample (c), which gives an undersmoothed LSCV estimate ($\nu_{\mathrm{LSCV}} = 107.47$) and a very good CCV estimate ($\nu_{\mathrm{CCV}} = 17.37$).

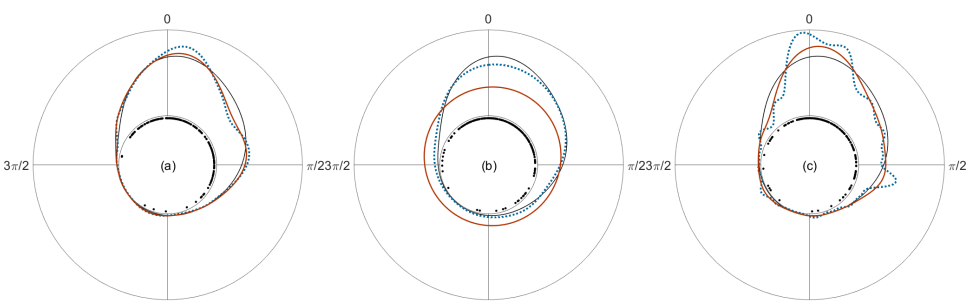


Figure 3: Selected samples from the model M5 and their estimates LSCV (blue, dotted), CCV (orange, full) with the true density (black, thin)

Admitting that there is no universally applicable method for circular kernel density estimation, we can see from the simulations that the CCV method is an acceptable alternative to the LSCV method. Looking at the graphs in Figure 3, we can consider applying the so-called averaging technique (Baszczyńska, 2017) to obtain the estimate that combines the local and global insight into the structure of the data.

5. Real data applications

The proposed method is applied to a few real data sets, each of them being similar to one of the models of the simulation study.

Example 1. Hisada (1972) studied dragonflies and their behaviour. The data set, available in the R package NPCirc (Oliveira et al., 2014), consists of 214 orientations of dragonflies with respect to the solar azimuth (measured in degrees). Figure 4(a) shows the data – zero corresponds to the azimuth of the sun – and the resulting kernel estimates obtained with optimal smoothing parameters $\nu_{\mathrm{LSCV}} = 63.87$ and $\nu_{\mathrm{CCV}} = 53.44$. As we can see, the CCV method gives almost the same density estimate as the LSCV method. As expected from the simulation study, the two methods provide almost the same results for data with two directly opposite modes (model M8).

Example 2. The Czech Hydrometeorological Institute has kindly provided the wind direction data. The data set represents the wind direction from Brno-Tuřany airport, the data were recorded every 10 minutes, resulting in 144 measurements for one day. For the analysis, 7 November 2021 was chosen. The optimal smoothing parameters obtained by the cross-validation methods are $\nu_{\mathrm{LSCV}} = 50.51$ and $\nu_{\mathrm{CCV}} = 30.77$, respectively. The data set shows bimodality (like model M7) and the resulting density estimates are similar, see Figure 4(b).

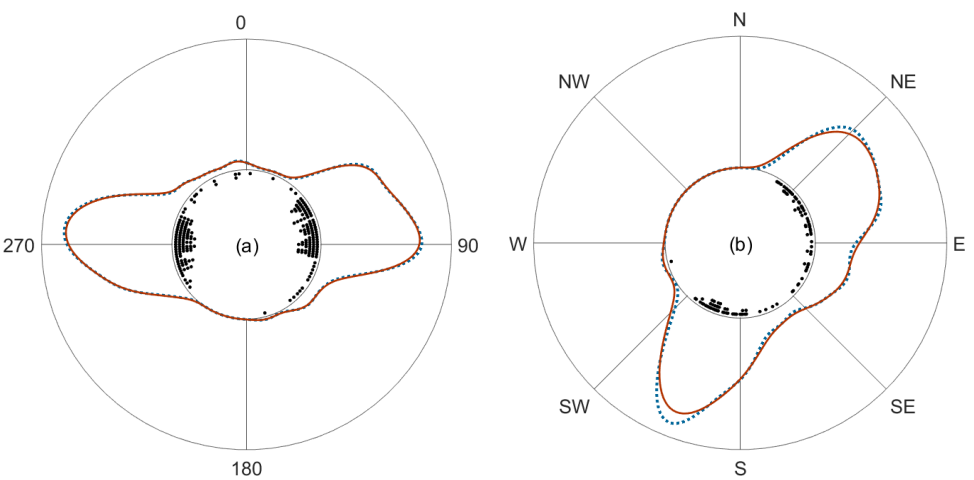


Figure 4: Kernel density estimates of the (a) dragonfly data and (b) wind data. Black points inside the circle represent the data, LSCV (dotted blue line) and CCV (full orange line) estimate

Example 3. In cooperation with the Lublin Municipal Transport Company, we collected data on bus subsystem failures over a period of more than seven years. The failures of each subsystem were recorded monthly, i.e. we only know the month of the specific failure, but not the exact date. The air conditioning (AC) subsystem ($n = 45$) was selected for investigation. The optimal smoothing parameters are $\nu_{\mathrm{LSCV}} = 2.37$ and $\nu_{\mathrm{CCV}} = 0.70$. Such small values leading to underestimated densities are due to the fact that these are aggregated data. Still, we can see that the failure occurrence is more likely to happen in early summer,

i.e. in May, June, and July. The resulting estimated densities are displayed in Figure 5.

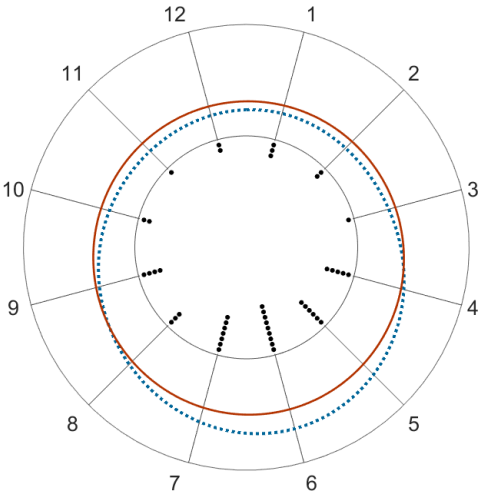


Figure 5: Kernel density estimate of the AC failure data (black points) with LSCV estimate (dotted blue) and CCV estimate (full orange)

6. Conclusion

We proposed a new complete cross-validation (CCV) method of bandwidth selection in circular density estimation. This method extends the estimate of the MISE and has better theoretical properties than the LSCV. From the presented results and outcomes we can conclude that the CCV is applicable to various data types with respective success rate.

From the data-driven method of bandwidth selection, we focus only on the cross-validation methods which target the MISE to have the consistent group of methods to compare. We have considered alternative approaches (see, e.g. Taylor, 2008; Oliveira et al., 2012; García-Portugués, 2013), but they target different type of error function – the asymptotic mean integrated square error. From the theoretical point of view also the augmented cross-validation (Tsuruta and Sagae, 2020), but this one cannot be used in real data application.

Although the CCV presents a theoretical concept, our aim is to make the method applicable to practical situations. The form and structure of the original datasets (corresponding to the true density) have to be taken into account. The size of the data set also plays an important role in producing results with some degree of validity. In datasets where data records are short, sparse or even missing, we cannot expect relevant estimates and results.

However, we are confident that the CCV method provides realistic and applicable results, not only in terms of key statistical results such as circular kernel density estimation, but also in terms of other measures that can be successfully applied in other areas such as reliability and safety engineering.

Acknowledgements

This paper has been prepared with the support of the Ministry of Defence of the Czech Republic, Partial Projects for Institutional Development LANDOPS and VAROPS, University of Defence, Brno, and Grant Agency of Masaryk University, project MUNI/A/1418/2022 "Mathematical and statistical modelling 4 (MaStaMo4)". We gratefully acknowledge the Czech Hydrometeorological Institute of Brno and the Lublin Municipal Transport Company for providing the wind direction and bus failure datasets, respectively.

References

Baszczyńska, A., (2017). One value of smoothing parameter vs interval of smoothing parameter values in kernel density estimation. *Acta Universitatis Lodziensis. Folia Oeconomica*, 6, pp. 73–86.

Bowman, A. W., (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71, pp. 353–360.

García-Portugués, E., (2013). Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electronic Journal of Statistics*, 7, pp. 1655–1685.

Hall, P., Marron, J. S., (1987). Estimation of integrated squared density derivatives. *Statistics and Probability Letters*, 6, pp. 109–115.

Hall, P., Watson, G. S., Cabrera, J., (1987). Kernel density estimation with spherical data. *Biometrika*, 74, pp. 751–762.

Hisada, M., (1972). Azimuth orientation of the dragonfly (*Sympetrum*). In Galler, S. R., et al. *Animal Orientation and Navigation*, pp. 511–522. Washington: U.S. Government Printing Office.

Jammalamadaka, S. R., SenGupta, A., (2001). *Topics in circular statistics*, Singapore: World Scientific.

Jones, M., Kappenman, R., (1991). On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, 19, pp. 337–349.

Ley, C., Verdebout, C., (2017). *Modern directional statistics*, Boca Raton: CRC Press.

Mardia, K. V., Jupp, P. E., (2000). *Directional statistics*, Chichester: Wiley.

Oliveira, M., Crujeiras, R. M., Rodríguez-Casal, A., (2012). A plug-in rule for bandwidth selection in circular density estimation. *Computational Statistics and Data Analysis*, 56, pp. 3898–3908.

Oliveira, M., Crujeiras, R. M., Rodríguez-Casal, A., (2014). NPCirc: An R Package for Nonparametric Circular Methods. *Journal of Statistical Software*, 61, pp. 1–26.

Rudemo, M., (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9, pp. 65–78.

Scott, D. W., (1992). *Multivariate density estimation: Theory, practice, and visualization*, New York: Wiley.

Silverman, B. W., (1986). *Density estimation for statistics and data analysis*, London: Chapman and Hall.

Taylor, C. C., (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, 52, pp. 3493–3500.

Tsuruta, Y., Sagae, M., (2017). Higher order kernel density estimation on the circle. *Statistics and Probability Letters*, 131, pp. 46–50.

Tsuruta, Y., Sagae, M., (2020). Theoretical properties of bandwidth selectors for kernel density estimation on the circle. *Annals of the Institute of Statistical Mathematics*, 72, pp. 511–530.

Wand, M. P., Jones, M. C., (1995). *Kernel smoothing*, London: Chapman and Hall.

Appendix

Convolution of the von Mises kernel $K_\nu(\theta)$ with itself is (Jammalamadaka and SenGupta, 2001):

$$(K_\nu * K_\nu)(\theta) = \frac{I_0\left(\sqrt{2\nu^2(1+\cos\theta)}\right)}{2\pi I_0^2(\nu)}.$$

Functionals T: Derivatives of the von Mises kernel can be expressed using the recurrence relationship

$$\begin{aligned} K_\nu(\theta) &= [2\pi I_0(\nu)]^{-1}\exp(\nu\cos\theta) & P_1(\theta) &= -\nu\cos\theta \\ K_\nu^{(n)}(\theta) &= K_\nu(\theta)\cdot P_n(\theta) & P_{n+1}(\theta) &= -\nu\sin\theta\cdot P_n(\theta) + P_n'(\theta). \end{aligned}$$

Then, the exact form of functionals T_0, T_1 and T_2 with the von Mises kernel is

$$\begin{aligned} T_0 &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} K_\nu(\Theta_{ij}) = \frac{1}{2\pi I_0(\nu)n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\exp(\nu\cos\Theta_{ij}), \\ T_1 &= -\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} K_\nu^{(2)}(\Theta_{ij}) \\ &= -\frac{1}{2\pi I_0(\nu)n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\exp(\nu\cos\Theta_{ij})\cdot\left(\nu^2\sin^2\Theta_{ij} - \nu\cos\Theta_{ij}\right), \\ T_2 &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} K_\nu^{(4)}(\Theta_{ij}) \\ &= \frac{1}{2\pi I_0(\nu)n(n-1)}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\exp(\nu\cos\Theta_{ij})\cdot\Big(\nu^4\sin^4\Theta_{ij} - 6\nu^3\sin^2\Theta_{ij}\cos\Theta_{ij} \\ &\qquad + 3\nu^2(\cos^2\Theta_{ij} - \sin^2\Theta_{ij}) - \nu^2\sin^2\Theta_{ij} + \nu\cos\Theta_{ij}\Big). \end{aligned}$$

Proof of the Lemma: First, we show expectations of functionals $T_m(\nu)$.

$$\begin{aligned} ET_0(\nu) &= E\left[K_\nu(\theta_1-\theta_2)\right] = R(f) - \frac{A_1(\nu)}{2\nu}R(f') + \frac{A_2(\nu)}{8\nu^2}R(f'') + o(\nu^{-2}), \\ ET_1(\nu) &= E\left[-K_\nu''(\theta_1-\theta_2)\right] = R(f') - \frac{A_1(\nu)}{2\nu}R(f'') + o(\nu^{-1}), \\ ET_2(\nu) &= E\left[K_\nu^{(4)}(\theta_1-\theta_2)\right] = R(f'') + o(1). \end{aligned}$$

Thus

$$E[\mathrm{CCV}(\nu)] = E[R(\widehat{f})] - ET_0(\nu) + \frac{A_1(\nu)}{2\nu} ET_1(\nu) + \frac{1}{8\nu^2}\left(2A_1^2(\nu) - A_2(\nu)\right) ET_2(\nu) + o(\nu^{-2})$$
$$= E[R(\widehat{f})] - R(f) + \frac{A_1(\nu)}{\nu} R(f') - \frac{A_2(\nu)}{4\nu^2} R(f'') + o(\nu^{-2}).$$

On the other hand, the middle term of MISE can be expressed as follows:

$$E\int \widehat{f}(\theta,\nu) f(\theta)\,\mathrm{d}\theta = \iint K_\nu(\theta-\alpha) f(\alpha) f(\theta)\,\mathrm{d}\alpha\,\mathrm{d}\theta$$
$$= R(f) - \frac{A_1(\nu)}{2\nu} R(f') + \frac{A_2(\nu)}{8\nu^2} R(f'') + o(\nu^{-2})$$

and the whole MISE reads as

$$\mathrm{MISE} = E[R(\widehat{f})] - 2E\int \widehat{f}(\theta,\nu) f(\theta)\,\mathrm{d}\theta + R(f)$$
$$= E[R(\widehat{f})] - 2R(f) + 2\frac{A_1(\nu)}{2\nu} R(f') - 2\frac{A_2(\nu)}{8\nu^2} R(f'') + R(f) + o(\nu^{-2})$$
$$= E[R(\widehat{f})] - R(f) + \frac{A_1(\nu)}{\nu} R(f') - \frac{A_2(\nu)}{4\nu^2} R(f'') + o(\nu^{-2}).$$

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 13–30, https://doi.org/10.59170/stattrans-2024-025
Received – 12.09.2022; accepted – 23.02.2024

Reconstruction of the social cash transfers system in Poland and household wellbeing: 2015–2018 evidence

Adam Szulc[1]

Abstract

This study examines the impact of changes in the social benefits system on the wellbeing, poverty, and economic activity in Poland. The core element of those changes was a programme of large cash transfers, referred to as Family 500+, introduced in 2016. It was intended to support families with children, especially the least affluent ones, and to foster fertility. The impact of the transfers has been examined through the observation of changes in the monetary and multidimensional wellbeing of households. The study also analysed the changes in recipients' economic activity using estimates of regression models and treatment effects. The Family 500+ programme proved to be successful as an anti-poverty tool and also resulted in the increase in the average wellbeing for the whole population. However, its side-effects included the fall in the economic activity of some recipients, especially in 2016 and 2017. The above trends partly reversed in 2018. As some income data in the lower parts of the distribution seemed to be flawed, income imputations, based on regression on income correlates, were applied in the study.

Key words: family benefits, monetary and multidimensional poverty, income imputation.

1. Introduction

In this study, the effects of the changes in the system of social cash transfers in Poland between 2015 and 2018 are examined. In April 2016, the state family support programme, which seriously changed the volume and structure of social benefits, was launched. It is known as "Family 500+" and ensures the monthly unconditional support of tax-free 500 PLN (złoty) per each child in families with two or more children and means-tested support of the same amount for families with one child. In 2016, 500 PLN was equal to 26% of the mean monthly equivalent income. Total spending in 2017 was equal to 6% of the state budget. Family 500+ also changed the composition of the social transfers – the contribution of all family benefits to their total amount (excluding retirement and invalid pensions) increased from 63% in 2015 to 86% in 2018. For more

[1] Institute of Statistics and Demography, Warsaw School of Economics, Poland. E-mail: aszulc@sgh.waw.pl. ORCID: https://orcid.org/0000-0003-2646-2468.

details on Family 500+, see Brzeziński and Najsztub (2017), and Ministerstwo Rodziny i Polityki Społecznej (2021). The effects of the abovementioned social policy reconstruction are examined by observing resulting changes in the standards of living, especially monetary and multidimensional poverty. The following changes in the labour activity of the recipients are also explored.

The declared goals of Family 500+ were: (i) to reduce child poverty, and (ii) to increase fertility (which was among the European lowest in 2015 in Poland). The opponents argued that there would be a negative impact on the labour supply and a low efficiency resulting from also covering non-poor families with at least two children through the programme. Among several analyses using formal quantitative methods, some of them claimed or anticipated negative effects from Family 500+. For instance, Brzeziński and Najsztub (2017), using microsimulation models, predicted a low efficiency of the program due to its structure. Magda et al. (2021) estimated, using the difference-in-difference method, its negative impact on the female labour supply, however those findings stand in opposition with the conclusions reached by Premik (2022). Wilk (2021) reported a low response to the programme in terms of the fertility rate. On the other hand, it is not surprising that Family 500+ has been effective in reducing the economic hardship of the families with children (Milovanska-Farrington, 2021). This is also claimed by simulation results presented by Brzeziński and Najsztub (2017).

In the present study, the abovementioned issues, excluding demographic ones, are analyzed further, using data for 2015–2018. The findings confirm many of the observations described above for the 2016–2017 period. However, in 2018 some trends reversed. The analysis reported here covers the effects of all social benefits not related to the social insurance system, i.e. retirement and invalid pensions. The effects do not differ much from those obtained solely for Family 500+. This suggests that this type of child allowance did not lead to unpredictable results and has generally been in line with the effects of the cash transfers observed for some post-communist countries (Szulc, 2012; Harumová, 2016).

The remaining part of the paper is organized as follows. In Section 2, the theoretical concepts employed in the study are discussed. In Section 3, basic statistics on the Polish social transfers system are provided. Section 4 is devoted to the impact of the transfers on monetary and multidimensional poverty. In Section 5, changes in economic activity of the recipients are examined. Section 6 concludes the paper.

2. Conceptual framework

2.1. Data

The individual data on households and persons employed in this research come from the household budget surveys collected annually by the central statistical office –

Statistics Poland. The yearly samples cover from 37,148 (2015) to 36,166 (2018) households. The reference period of observation is one month. Basic methodological details may be found in Główny Urząd Statystyczny – Statistics Poland (2018). The household data include a wide set of economic, demographic and sociological variables, allowing the evaluation of various aspects of households' and individuals' economic positions. Those utilised in the present study encompass, inter alia, information on household disposable income and its components, expenditures, assets, durables, dwelling conditions, demographic and socio-economic attributes, and answers to subjective income questions. There are two-year panel components covering from 15,635 (2015–2016) to 15,155 (2017–2018) households included in the samples. The survey data were weighted using the 2011 National Census results to minimise differences between the structure of the surveyed sample and the population.

There are reasons to assume that for some households their incomes are misreported, especially at the bottom of the distribution. This problem is tackled by income imputations. Some observable variables that may be assumed to be more reliable and stable in time are used to provide the estimates. They capture information on housing, consumer durables, education and some expenses related to standards of living. The income imputation is based on algorithm for missing data imputation proposed by Rubin (1987). More details may be found in Szulc (2022, pp. 4–5).

2.2. Well-being measurement

Household equivalent disposable income (employing OECD 70/50 equivalence scales) is used as a monetary well-being indicator. As it is generally accepted that monetary poverty measurement does not suffice as a tool for capturing the nature of poverty, a multidimensional approach has also been adopted in the present study. It comprises the following components: (i) income, (ii) housing and equipment, and (iii) subjective evaluations of one's own standard of living. The first one is represented by a function of equivalent income defined by eqn. (1), while the two remaining are composed of sets of single variables aggregated in one indicator. The final measure is defined as a weighted mean of indicators calculated for all three dimensions separately.

At the first stage poverty measures are calculated. If a variable describing poverty at the lowest level of aggregation is binary, it is equal to 0 when a symptom of poverty (e.g. a lack of some consumer goods) is not observed and 1 otherwise. For continuous (e.g. the equivalent income or dwelling size) and discrete ordinal (e.g. subjective evaluation of own economic conditions) variables, the concept of well-being indicator is based on the "Totally Fuzzy and Relative" (TFR) approach to multidimensional poverty measurement proposed by Cheli and Lemmi (1995). In the "fuzzy" approach, as opposed to the dichotomous approach, no single poverty line is set. Instead, for a variable y used as a single dimension well-being measure, the degree of poverty based

on preselected interval, say [y^*, y^{**}], is calculated for each unit (individual or household). A poverty measure for the i-th unit is equal to 1 when $y \leq y^*$ and equal to 0 when $y \geq y^{**}$. To measure poverty for y$\in(y^*,y^{**})$ the following TFR function is applied as the "fuzzy" poverty indicator for i-th unit:

$$p_i = p(y_i) = \begin{cases} 0 & if\ y\ \geq\ y** \\ p(y_{i-1}) + \frac{F(y_i)-F(y_{i-1})}{1-F(y*)} & if\ \ y \in (y*, y**) \\ 1 & if \leq\ y\ y* \end{cases} \quad (1)$$

where F stands for an empirical cumulative distribution function. By definition, p_i fits the interval [0; 1]. When y is a discrete ordinal variable (like subjective income questions ranging from "very bad" to "very good"), it is natural to set $y^* = y_{min}$ and $y^{**} = y_{max}$. Formally, the same choice may be applied to continuous variables, like income, but it seems to be more rational to set $y^* > y_{min}$ and $y^{**} < y_{max}$ to relax impact of outliers. Moreover, when $y^* = y_{min}$ and $y^{**} = y_{max}$, the indicator p_i is "totally relative" and its value depends on the shape of the distribution only. In this study, comparisons between years are performed, hence fixing y^* and y^{**} over time is a better choice. For income, the bottom limit is equal to the 2015 existence minimum, while the upper limit is three times the social minimum (both thresholds are calculated by the Institute of Labour and Social Affairs, 2020). For other (nearly) continuous variables, like dwelling size, y^* and y^{**} are set at the 0.05 and 0.95 percentiles obtained for 2015, respectively.

Once individual measures are defined, they should be aggregated to mid-level dimensions, i.e. equivalent income, housing and equipment, and subjective evaluations of the own economic conditions. As the first one is represented by a single variable, the problem of aggregation is relevant for the two remaining dimensions only. The weighting system within those dimensions was proposed by Cheli and Lemmi (1995). For the *j-th* item it is calculated as:

$$w_j = ln\left(\frac{1}{\overline{H_j}}\right)$$

where $\overline{H_j}$ denotes the proportion of units that are poor with respect to the *j-th* item. It is presumed that a more frequent poverty syndrome (e.g. the lack of a passenger car) is a less important symptom of poverty than a less frequent one (e.g. the lack of a refrigerator). However, that form of weighting is problematic when three mid-level components are to be aggregated in one well-being indicator, as it assumes the equal importance of all dimensions. Hence, at the highest level of aggregation arbitrary weights are applied: 0.5 for income and 0.25 for both remaining dimensions.

In the present study, the impact of social transfers on the standards of living is one of the main goals. Therefore, it is more convenient to define a multidimensional well-being measure instead of a poverty measure to make it compatible with equivalent income. This may be easily done by defining the multidimensional well-being indicator $F = 1 - P$ where P is an aggregate poverty indicator.

2.3. Measurement of the transfers' effects on the well-being

The final impact of the social transfers on the well-being depends on their volume and allocation. Comparing actual poverty indices and pre-transfer (simulated) ones allows for the evaluation of the simultaneous effect of both abovementioned attributes. Indices gauging poverty incidence and depth are used for that purpose and the poverty lines are set at the first decile and first quartile. Moreover, the elasticity of the effect with respect to the poverty threshold is evaluated using graphical methods. Poverty indices are calculated for each year separately, for all types of households together and for households with children. Simulated indices of income poverty are calculated by means of actual incomes reduced by transfer values. For multidimensional poverty, regression models are estimated to predict changes in well-being levels due to changes in incomes.

A static evaluation of the transfers is supplemented by a dynamic one aimed at answering two questions: how well the non-poor are protected from falling into poverty and to what extent transfers allow the poor to leave the poverty zone. For that purpose, joint (two year) well-being distribution is constructed using panel data. Both income and multidimensional poverty are included in the analyses. The concept applied in the present study follows the idea of protection and promotion effects proposed by Ravallion et al. (1995). If the analysis is restricted to transitions to and out of poverty, the effects may be estimated as follows. A protection effect takes the form of a relative difference between the simulated number (subscripted by *S*) of new poor and the actual one (subscripted by *A*):

$$PROT = \frac{N[y_S(0) \geq z_0 \ \& \ y_S(1) < z_1] - N[y_A(0) \geq z_0 \ \& \ y_A(1) < z_1]}{N[y_S(0) \geq z_0 \ \& \ y_S(1) < z_1]} \tag{2}$$

where $N\left[y_{A/S}(0) \geq z_t \ \& \ y_{A/S}(1) < z_t\right]$ is the number of individuals who were not poor in period 0 and became poor in period 1. Similarly, the promotion effect takes a form of a relative difference between the actual number of new non-poor and the corresponding simulated number:

$$PROM = \frac{N[y_A(0) < z_0 \ \& \ y_A(1) \geq z_1] - N[y_S(0) < z_0 \ \& \ y_S(1) \geq z_1]}{N[y_A(0) < z_0 \ \& \ y_A(1) \geq z_1]} \tag{3}$$

where $N\left[y_{A/S}(0) < z_t \,\&\, y_{A/S}(1) \geq z_t\right]$ is the number of individuals who were poor in period 0 and non-poor in period 1. The simulations are intended to answer the question: what would happen if the transfers were terminated.

Finally, the impact of the transfers on the labour activity measured by means of the changes in economically active people is gauged. A static version is based on two treatment effect estimates: a propensity score matching (see Abadie and Imbens, 2012, or Wooldridge, 2010, pp. 903–936) and inverse probability weighted regression adjustment (hereafter: IPWRA, see Wooldridge, 2007, or Cattaneo et al., 2013). In a dynamic approach, changes in outcome variables of interest are regressed on changes in transfers using two-year panels. Both matching estimation and IPWRA are intended to produce unbiased estimates of a treatment effect, which in the present case is defined as receiving a certain type of benefits.

Using panel data in both static and dynamic analysis gives an opportunity to reduce the bias caused by endogeneity (regression models) or violating the unconfoundedness assumption (the estimation of treatment effect). Both types of bias result from the occurrence of omitted variables in the model, especially unobservable ones like the capability to earn income related to psychological attributes or hidden skills. Omitting them in the model usually results in a downward bias in the estimation of the effect of benefits if they are means-tested. When panel data are available it is possible to use benefits received in the basic period as proxies for abovementioned unobservable control variables.

3. General review of the social benefits in Poland: descriptive statistics

Between 2015 and 2018, the considerable growth of mean standards of living could be observed: the mean equivalent income rose by 19.8% and the multidimensional well-being indicator by 2.8%. In 2016 and 2017, but not in 2018, the recipients of social benefits experienced higher than average well-being increases. At the same time, due to the growing number of children, for recipients of 500+ the growth rate was lower than for all the beneficiaries of social transfers. The main feature of the changes in the system of social cash transfers (see Table 1) was the huge growth of family benefits under the relative stability of the remaining ones. For the whole sample, the mean real value of the first rose by 291%, while the mean value of the latter by 10%. The mean value of all transfers increased by 188% for the whole sample and among recipients of the benefits by 99% and by 222% among households with children. Those growths were accompanied by moderate increases in the proportion of recipients of all types of transfers: from 29% to 33%. As might be expected, the proportion of family benefits recipients greatly increased, at the cost of the remaining benefits. A huge growth was also observed for households with children.

Table 1: Social benefits: basic statistics (monthly means in 2015 prices)

Type of the benefit	2015	2016	2017	2018
mean benefits				
All types	191	430	563	550
Family	121	353	488	473
incl. 500+	-	219	352	334
Other	70	78	75	77
All types – recipients only	560	996	1092	1116
All types – hh with children only	291	716	952	936
proportion of the recipients				
All types	0.287	0.328	0.369	0.331
Family	0.101	0.167	0.224	0.213
incl. 500+	-	0.112	0.199	0.190
Other	0.186	0.161	0.146	0.118
All types – hh with children only	0.367	0.529	0.682	0.680
proportion of the households with children				
-	0.338	0.333	0.326	0.316

Source: author's own calculations based on the household budget surveys.

Although expanding the value of social benefits usually results in easing the economic hardship of recipients, it may also lead to dependency on the social system (see, e.g. Kotlikoff et al., 2006 and Shepherd et al., 2011). As displayed in Table 2, the average proportion of social benefits in the household income between 2015 and 2018 rose from 6.6% to 11.8% for the whole sample and from 19.4% to 24% for the recipients. Moreover, the proportion of households for which social benefits contribute more than 50% to the whole income increased from 8.4% to 10.4%. Other side effects of the social transfers, especially the reduction of economic activity, are analyzed in Section 5.

Table 2: Contribution of the social transfers to household income

Specification	2015	2016	2017	2018
Share of the transfers: all households	0.066	0.102	0.126	0.118
Share of the transfers: all recipients	0.194	0.237	0.244	0.240
Share of the transfers: households with children	0.094	0.161	0.204	0.193
Share of the transfers < 0.2	0.730	0.656	0.641	0.646
0.2 < share of the transfers < 0.5	0.186	0.240	0.253	0.251
Share of the transfers > 0.5	0.084	0.103	0.106	0.104
% of poverty gap, poverty line at the first quartile	79	140	170	148
100% of poverty gap if poverty line at centile	21	31	35	33

Source: author's own calculations based on the household budget surveys.

4. Impact of the transfers on incidence and depth of the poverty

4.1. Static analysis

A conventional measure of an effect of social transfers takes the form of the difference between poverty indices calculated with the use of actual and pre-transfer incomes. For income poverty such a difference may be obtained just by subtracting cash transfers from actual incomes and then calculating simulated indices of poverty. For multidimensional poverty changes in indices may be predicted conditionally on changes in income. In the present study, this measure is implemented using regression models with a multidimensional well-being indicator as a dependent variable regressed on a quadratic polynomial of equivalent income. To estimate this model, the sample was restricted to the incomes between the first decile and the median. Censoring the lowest incomes is intended to reduce the impact of data errors mentioned in Section 2.1 which result, inter alia, in nonsensical relations between equivalent income and multidimensional poverty (see Szulc, pp. 19–20). Hence, indices measuring poverty are calculated for both types of incomes. The results are reported in Table 3 (income poverty) and in Table 4 (multidimensional poverty).

Table 3: Income poverty rates and depth: before and after transfers differences

Poverty measure	2015	2016	2017	2018	2015	2016	2017	2018
	declared incomes				corrected incomes			
poverty line at the first quartile, all households								
Rate, before	0.289	0.336	0.355	0.349	0.292	0.338	0.357	0.350
Difference	**0.039**	**0.086**	**0.105**	**0.099**	**0.042**	**0.088**	**0.107**	**0.100**
Depth, before	0.340	0.345	0.349	0.345	0.264	0.284	0.300	0.290
Difference	**0.071**	**0.100**	**0.123**	**0.110**	**0.078**	**0.119**	**0.140**	**0.131**
poverty line at the first quartile, households with children								
Rate before	0.372	0.429	0.447	0.442	0.376	0.433	0.454	0.445
Difference	**0.055**	**0.137**	**0.175**	**0.167**	**0.058**	**0.140**	**0.177**	**0.167**
Depth, before	0.356	0.369	0.378	0.372	0.283	0.314	0.340	0.330
Difference	**0.081**	**0.126**	**0.164**	**0.148**	**0.089**	**0.150**	**0.182**	**0.171**
poverty line at the first decile, all households								
Rate before	0.147	0.189	0.219	0.210	0.155	0.203	0.228	0.220
Difference	**0.047**	**0.089**	**0.119**	**0.110**	**0.055**	**0.103**	**0.128**	**0.120**
Depth, before	0.343	0.347	0.345	0.343	0.252	0.279	0.292	0.275
Difference	**0.096**	**0.110**	**0.120**	**0.103**	**0.122**	**0.164**	**0.183**	**0.174**
poverty line at the first decile, households with children								
Rate before	0.200	0.262	0.299	0.287	0.214	0.283	0.320	0.311
Difference	**0.068**	**0.145**	**0.198**	**0.185**	**0.079**	**0.166**	**0.211**	**0.201**
Depth, before	0.351	0.358	0.361	0.357	0.264	0.303	0.323	0.304
Difference	**0.109**	**0.129**	**0.152**	**0.128**	**0.129**	**0.190**	**0.214**	**0.205**

Source: author's own calculations based on the household budget surveys.

Table 4: Multidimensional poverty incidence and depth: before and after transfers differences

Poverty measure	2015	2016	2017	2018	2015	2016	2017	2018
	declared incomes				corrected incomes			
poverty line at the first quartile, all households								
Rate before	0.291	0.338	0.354	0.347	0.302	0.357	0.371	0.363
Difference	**0.041**	**0.088**	**0.104**	**0.097**	**0.052**	**0.107**	**0.121**	**0.113**
Depth, before	0.279	0.291	0.304	0.282	0.275	0.346	0.365	0.350
Difference	**0.077**	**0.100**	**0.121**	**0.103**	**0.079**	**0.158**	**0.180**	**0.171**
poverty line at the first quartile, households with children								
Rate before	0.364	0.420	0.437	0.436	0.374	0.440	0.455	0.451
Difference	**0.061**	**0.142**	**0.175**	**0.167**	**0.071**	**0.162**	**0.192**	**0.179**
Depth, before	0.288	0.308	0.326	0.302	0.286	0.370	0.394	0.379
Difference	**0.088**	**0.128**	**0.164**	**0.137**	**0.090**	**0.192**	**0.225**	**0.211**
poverty line at the first decile, all households								
Rate before	0.158	0.203	0.229	0.216	0.153	0.230	0.255	0.248
Difference	**0.058**	**0.103**	**0.129**	**0.116**	**0.053**	**0.130**	**0.155**	**0.148**
Depth, before	0.268	0.276	0.284	0.253	0.293	0.352	0.366	0.344
Difference	**0.094**	**0.105**	**0.119**	**0.105**	**0.122**	**0.189**	**0.201**	**0.198**
poverty line at the first decile, households with children								
Rate before	0.206	0.271	0.304	0.291	0.201	0.303	0.334	0.329
Difference	**0.084**	**0.168**	**0.213**	**0.193**	**0.079**	**0.200**	**0.240**	**0.227**
Depth, before	0.270	0.283	0.296	0.265	0.294	0.365	0.385	0.362
Difference	**0.102**	**0.125**	**0.153**	**0.135**	**0.125**	**0.216**	**0.236**	**0.229**

Source: author's own calculations based on the household budget surveys.

Poverty lines are set at the first deciles and the first quartiles of the well-being for the whole samples. Comparisons based on the raw survey data lead to two general conclusions: (i) the impact of the transfers increased sharply in 2016 and then in 2017, and (ii) the lower the poverty line, the stronger the impact. This finding is valid for both poverty incidence and depth. All estimates of the effects are significant below the 0.01 level. It is not surprising that the effects on poverty are stronger for the households with children, especially after 2015. However, even in 2015 the family allowances were sufficient to provide stronger that average reduction of the pre-transfer poverty. Comparing the abovementioned results with those obtained by means of corrected incomes yields similar general conclusions, although estimates of the effects are greater in the latter case. It may be assumed that due to removing some "fake poor" from the sample (more precisely: moving them to higher ranges of the distribution), those estimates of the effects are more reliable. To display changes in the effects of the transfers not attached to a fixed poverty line, the effects are plotted over a variable

poverty line for 2015 and 2018. When declared survey incomes are utilised, the impact on the poverty depth for poverty lines set below the first decile do not necessarily decrease with respect to the poverty threshold, which is rather a counterintuitive result. There are reasons to believe that this may be an effect of income data errors: the "fake poor" less frequently receive benefits and therefore after-transfer poverty is reduced to low extent. The plots (see Fig. 1 and Fig. 2) produced with the use of corrected incomes support this hypothesis: observed changes in the effects due to the poverty line are much more reliable.

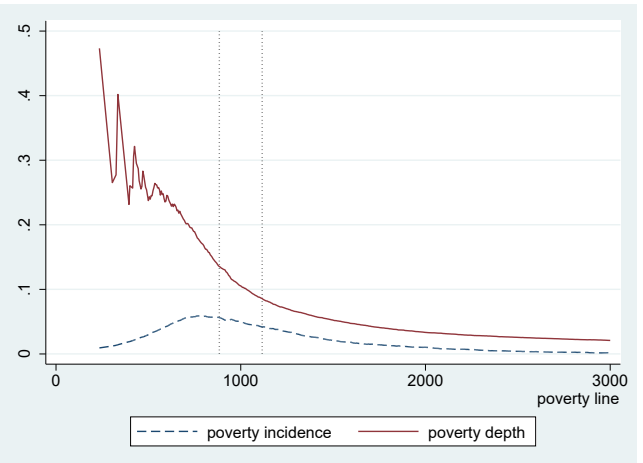


Figure 1: Difference between after and before transfer income poverty incidence and depth, corrected incomes, 2015; vertical dotted lines at the first decile and the first quartile

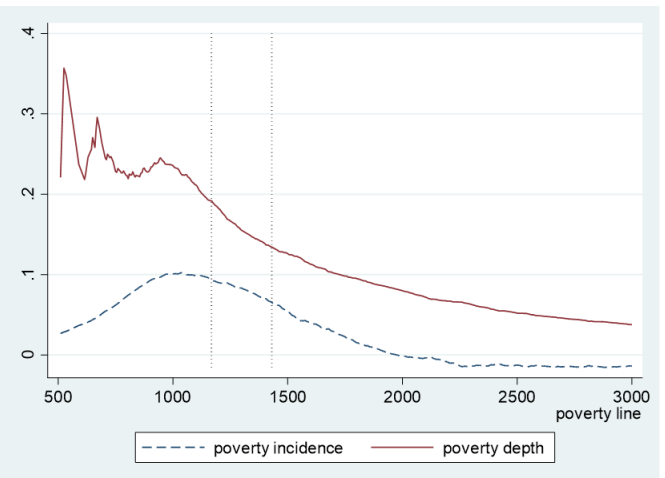


Figure 2: Difference between after and before transfer income poverty incidence and depth, corrected incomes, 2018; vertical dotted lines at the first decile and the first quartile

Conclusions on multidimensional poverty do not differ much from those derived from the estimates obtained for corrected incomes: considerable increases of the effects in 2016 and in 2017 and decreases of the effect due to the increase in the poverty line may be observed. This is true for the results attained by means of both declared and corrected incomes. The impact of the transfers on poverty depth is noticeably stronger when corrected incomes are applied.

4.2. Dynamic analysis

Due to using two-year panel data, it is possible to estimate the protection and promotion effects of the transfers described in Subsection 2.3. Both these measures display simulated transitions between the poverty and non-poverty zones due to changes in social transfers. In Table 5a and Table 5b, the simulated effects of removing all social benefits are reported. For instance, the protection effect 0.171 observed for 2015 - 2016 period suggests that the number of "new poor" in 2016 would increase by 17.1% without the transfers. The protection effect 0.278 observed for the same period should be interpreted as 27.8% increase in the number those who escaped poverty in 2016 after receiving the transfers. Unlike in the case of static effects, negative protection/promotion effects are likely and they occurred in the 2017–18 period. Naturally, this does not mean a counterproductive effect of the benefits, as simulated poverty rates are still much higher than actual ones (see Tables 3 and 4). Rather, the negative value suggests that there are sources for successfully coping with poverty other than social benefits. The increases in labour activity observed between 2017 and 2018 (see Table 6) support this hypothesis. Nevertheless, most of the protection/promotion effects appeared to be positive and significant, usually below the 0.01 level. As might also be observed in the case of static effects, the lower the poverty line, the stronger the effect. The effects for multidimensional poverty are stronger than those for income poverty.

Table 5a: Protection and promotion effect for income poverty

Specification	2015/16	2016/17	2017/18
poverty line at the first quartile			
Protection	0.171***	0.019	-0.166**
Promotion	0.278***	0.134***	-0.023
poverty line at the first decile			
Protection	0.562***	0.629***	0.642***
Promotion	0.430***	0.444**	0.304***

Source: author's own calculations based on the household budget surveys.

*Legend: ***: significant at 0.01, **: significant at 0.05, *: significant at 0.1 (bootstrap standard errors)*

Table 5b: Protection and promotion effect for multidimensional poverty

Specification	2015/16	2016/17	2017/18
poverty line at the first quartile			
Protection	0.430***	0.490***	0.469***
Promotion	0.322***	0.344***	0.236***
poverty line at the first decile			
Protection	0.669***	0.706***	0.685***
Promotion	0.396***	0.371**	0.221***

Source: author's own calculations based on the household budget surveys.

*Legend: ***: significant at 0.01, **: significant at 0.05, *: significant at 0.1 (bootstrap standard errors)*

5. Changes in economic activity following the cash transfers

Examining the household response to receiving increased benefits, especially Family 500+ is another goal of the present study. Impact on the labour activity of the recipients, measured by the changes in the numbers of economically active people, is investigated by means of univariate and multivariate analyses. The latter ones employ regression methods, with a Heckman (1976) correction when necessary, and the estimation of treatment effects (matching estimation and IPWRA). Regression models for *i-th* unit take a general form:

$$\Delta Y_i = \alpha_0 + \alpha_1 \Delta X_i + \alpha_2 \boldsymbol{Z_i} + \varepsilon_i \quad (4)$$

where ΔY_i stands for a change in indicator of well-being or economic activity, ΔX_i is a change in the benefit value, and $\boldsymbol{Z}_i$ represents a set of control variables assumed to be correlated with a response variable.

Table 6: Changes in economic activity: whole sample vs new Family 500+ recipients

Specification	2015/16	2016/17	2017/18
Change in:			
income from economic activity	0.069	0.094	0.074
income from economic activity, new 500+	0.045	0.028	-0.135
non-social income	0.071	0.092	0.097
non-social income, new 500+	0.018	0.040	-0.117
no. of active women	-0.026	-0.023	-0.023
no. of active women, new 500+	-0.024	-0.016	0.095
no. of active men	-0.034	-0.031	-0.025
no. of active men, new 500+	-0.034	-0.036	0.119

Source: author's own calculations based on the household budget surveys.

In Table 6 some univariate statistics for the whole sample and for households receiving 500+ for the first time are reported. It is not surprising that the recipients' incomes originating from economic activity (employment and self-employment, including unpaid work in a family firm/farm) were growing at a slower pace, as compared to the whole sample, and even dropped in 2018. Similar trends can be observed for all non-social incomes (including old age and invalid pensions). Due to the low number of new 500+ recipients in 2018 (1.7% of the whole sample), the results for the 2017–2018 period may be not very reliable due to sampling errors. This problem is resolved further by estimating regression models on the whole panel sample (the results are reported in Table 7) that allow controlling for changes in other variables having impact on the variable(s) of interest.

In the regression models the yearly change in economic activity (income from economic activity and the number of active people) is a dependent variable and the yearly change in 500+ is an independent one. The sets of control variables (varying between estimations) comprise changes in remaining benefits, various household attributes, and the number of economic active men and women, as well as the benefits and incomes in the basic year. All estimates are performed on a sub-sample of the households with children. There are at least two potential problems with the estimation of such a model. First, omitted unobservable variables, like people's attitudes, may result in endogeneity which, in a scarcity of potential instrumental variables, is a serious problem. As Family 500+ for households with one child is means-tested, the recipients' earning income ability is likely to be lower than that of the non-recipients. Using panel data may be an alternative to the instrumental variables estimation. It is possible to relax the impact of omitted variables by including the values of income and benefits during a basic year as proxy variables for the earning ability. Another problem with the estimation stems from the selection of the subsample of recipients. As households with children are more likely to pay more attention to family values at the cost of economic activity, they are also more likely to reach a lower income which is not necessarily a side effect of receiving benefits. In other words, people bearing children may be more likely to stay at home rather than enter the labour market if their potential earnings are lower than their reservation wage. To receive unbiased estimators of eqn. 4 regression models with a Heckman correction are applied. The results are reported in Table 7. The impact of the social benefits is also evaluated by means of estimates of the treatment effects and IPWRA using panel data sets to provide more complete information on household earning ability. Omitting it would probably result in violating the unconfoundedness assumption in the matching estimation. As in the case of regression models presented above, information on transfers of the basic year is included in the estimation. The results of the estimations are displayed in Table 8. In general, conclusions on the

negative impact of the transfers on labour/self-employment income are consistent irrespectively to the method of estimation. On the other hand, those effect in 2018 appeared to be less significant than in the previous years.

Table 7: Impact of social benefits using regression: on Family 500+ and on all transfers (1 unit = 100 PLN)

Specification	2015/16	2016/17	2017/18
Income from economic activity on:			
500+ *(Heckman regression)*	-25.785***	-41.7375***	-43.2491***
all transfers *(LSQ)*	-24.4548***	-35.1543***	-28.4337***
No. of active women on:			
500+ *(Heckman regression)*	-0.0051***	-0.0033**	0.0017
all transfers *(LSQ)*	-0.0037***	-0.0009	0.0009
No. of active men on:			
500+ *(Heckman regression)*	-0.0045***	-0.0037**	-0.0008
all transfers *(LSQ)*	-0.0035**	-0.0004	0.0010

Source: author's own calculations based on the household budget surveys.

*Legend: ***: significant at 0.01, **: significant at 0.05, *: significant at 0.1 (bootstrap standard errors).*

Table 8: Impact of social benefits using treatment effect estimation: on Family 500+ and on all transfers

Specification	2015/16		2016/17		2017/18	
	matching	IPWRA	matching	IPWRA	matching	IPWRA
Income from economic activity on:						
500+	-185.2416***	x	-184.5467	x	-156.2497	x
all transfers	-327.35***	x	-89.2194	-227.03***	-212.18***	-256.84***
No. of active women on:						
500+	-0.0391**	-0.0307**	-0.1843**	-0.0366	0.0659	-0.0481
all transfers	-0.0794**	-0.0510***	-0.0200	-0.0899***	0.0097	-0.1396**
No. of active men on:						
500+	-0.0146	-0.0547***	0.1078	0.0393	0.2122**	0.1386
all transfers	-0.0217	-0.0416***	0.2924***	0.0633**	0.0862	-0.0963**

Source: author's own calculations based on the household budget surveys.

*Legend: ***: significant at 0.01, **: significant at 0.05, *: significant at 0.1 (bootstrap standard errors); x: convergence not achieved*

Table 9: Impact of social benefits using 3 level treatment effect estimation (IPWRA): on Family 500+ and on all transfers

Specification	2015/2016	2016/2017	2017/2018
Income from economic activity on:			
500+	x	-141.4552***	x
all transfers	x	-431.4383*	x
No. of active women on:			
500+	-0.0250***	-0.0126	-0.1049***
all transfers	0.0362	-0.0798	-0.2115*
No. of active men on:			
500+	-0.0253***	-0.0190**	-0.0799***
all transfers	-0.1084*	-0.0391	-0.9569***

Source: author's own calculations based on the household budget surveys.

*Legend: ***: significant at 0.01, **: significant at 0.05, *: significant at 0.1 (bootstrap standard errors); x: convergence not achieved*

The natural question is which results, based on regression or on treatment effects estimates, are more reliable. The previous method is more sensitive to the specification of the model, however it allows continuous variables representing the transfers while in a conventional matching estimation only binary variable may be used. This restriction may be partly overcome by applying multilevel treatment effect IPWRA models. In this study the social transfers are represented by a discrete variable V that equals 0 if no transfers are received by the households, 1 if transfers without 500+ are received and 2 if other transfers including 500+ are received. The respective estimates are displayed in Table 8. When the estimates are statistically significant they are close to those obtained by means of regression (changes in the number of men and women that are economically active are reported in four last rows of Table 6). They were calculated for the whole sample and for the people living in the households of recipients of Family 500+. Due to the ageing society and the increasing number of pensioners (which grew by 9.2% between 2015 and 2018), the number of economically active persons decreased over that period. Nevertheless, the number of economically active women decreased at a lower-than-average pace between 2015 and 2017 and rose substantially in 2018 (by 9.5%). The corresponding indicators for the men were slightly above the average, but the increase in 2018 was even higher than that for the women (by 11.9%). On the other hand, a small sample of the new Family 500+ recipients in 2018 probably resulted in large sampling errors. Relatively minor declines in economic activity may be easily explained by the fact that adults in household bearing children are usually below the retirement age. This suggests that the

abovementioned changes might be related to changes in social transfers. As in the previous section, this issue is analysed further by means of the estimation of regression models and treatment effects. The results are reported in last six rows of Table 7 (regression), Table 8 (matching estimation with a binary treatment) and Table 9 (IPWRA with a three-level treatment).

6. Concluding remarks

Changes in the system of social cash transfers following the introduction of Family 500+ in 2016 increased their total amount enormously. The mean value was equal to 79% of the poverty gap calculated with the use of poverty line at the first income quartile, to 140% in 2016 and to 170% in 2017. Therefore, it is not surprising that the reduction of monetary and multidimensional poverty was meaningful. Moreover, the growth of the transfer resulted in a reduction in the economic activity of the recipients, especially in 2016 and 2017. In 2018 some of the abovementioned trends reversed. This may be attributed to the relative decrease in Family 500+ values, resulting from increases in mean incomes and from the inflation (3.5% between 2016 and 2018). Distinguishing between income and multidimensional poverty generally does not lead to basically opposite results, although the impact of the transfers on multidimensional poverty is usually stronger than that on income poverty. This applies also to estimates of the protection and promotion effects in a dynamic analysis of poverty and also is confirmed by concentration curves and coefficients (see Szulc, 2022, pp. 14–16). This suggests that the administration does not employ only an income criterion when addressing the benefits (see Ravallion, 2009, for a wider general discussion of the issue).

Increases in the transfers' volume resulted in a minor reduction of the economic activity of the new recipients, especially in the first two years of the Family 500+ programme. Univariate analysis revealed drops in the numbers of economically active people, both men and women, in 2016 and 2017 followed by massive increases in 2018 (although the latter may be overestimated due to a very small sample size of new 500+ recipients). Nevertheless, the benefits appeared to be large enough to compensate for the reduction in economic activity and the incomes of the recipients were increasing at a higher pace between 2015 and 2017. In 2018 this trend partly reversed, due to a lack of indexation of the 500+. Multivariate analysis, based on regression models and on estimates of the treatment effect claimed negative impact of the benefits, of any type, on employment and self-employment incomes. Comparing the impact of Family 500+ and the social benefits altogether yields very similar conclusions.

One of the problems in analyses using household surveys is the quality of survey income data, especially at the bottom of the distribution. This problem was tackled by means of income imputations based on regressions on selected income correlates

performed for middle ranges of incomes. Applying this method suggests the underestimation of the transfer impact on poverty when declared, uncorrected incomes are utilised. Due to the non-random selection of observations replaced by the estimated incomes, formal statistical inference on imputed values could not be performed.

Acknowledgement

This study was partly sponsored by the Collegium of Economic Analyses/Warsaw School of Economics research grant. The author is grateful to the anonymous referee for the valuable comments. All remaining errors are solely the responsibility of the author.

References

Abadie, A., Imbens, G. W., (2009). Matching on the estimated propensity score. Working Paper 15301. *Harvard University and National Bureau of Economic Research,* Cambridge, Massachusetts, http://www.nber.org/papers/w15301.

Brzeziński, M., Najsztub, M., (2017). The impact of "Family 500+ programme on household incomes, poverty and inequality. *Polityka Społeczna,* Vol. 44, pp. 16–25.

Cattaneo, M. D., Drukker, D. M., Holland, A. D., (2013). Estimation of multivalued treatment effects under conditional independence. *Stata Journal*, Vol. 13, pp. 407–450.

Cheli, B., Lemmi, A., (1995). A "Totally" Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. *Economic Notes*, Vol. 24, pp. 115–134.

Główny Urząd Statystyczny - Statistics Poland, (2018). *Budżety gospodarstw domowych - Household Budget Survey*, Warsaw.

Harumová, A., (2016). Inclusive Labor Markets as a Solution of Long Term Unemployment in Slovakia. *Mediterranean Journal of Social Sciences*, Vol. 7, pp. 194–200.

Heckman, J., (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models; in: *Annals of Economic and Social Measurement*, Vol. 5 (4), pp. 475–492.

Institute of Labour and Social Affairs, (2020). *Minimum socjalne i minimum egzystencji*, Minimum socjalne i minimum egzystencji | IPiSS.

Kotlikoff, L. J., Marx, B., Rizza, P., (2006). Americans' Dependency on Social Security, Working Paper 12696, *National Bureau of Economic Research, Cambridge, Massachusetts*, http://www.nber.org/papers/w12696.

Magda, I.; Kiełczewska, A., Brandt, N., (2018). The "Family 500+" child allowance and female labour supply in Poland, OECD Economics Department Working Papers, *No. 1481*, *OECD Publishing, Paris*, https://ibs.org.pl/en/publications/the-family-500-child-allowance-and-female-labour-supply-in-poland/.

Milovanska-Farrington, S., (2021). The Effect of Child Benefits on Financial Difficulties and Spending Habits: Evidence from Poland's Family 500+ Program, IZA DP No. 14274, *IZA – Institute of Labor Economics*, https://docs.iza.org/dp14274.pdf.

Ministerstwo Rodziny I Polityki Społecznej, (2021). Rodzina 500 plus, https://www.gov.pl/web/rodzina/rodzina-500-plus.

Premik, F., (2022). Evaluating Poland's Family 500+ Child Support Programme, *Gospodarka Narodowa. The Polish Journal of Economics,* Vol. 310, pp. 1–19.

Ravallion, M., (2009). Miss-targeted or miss-measured?, *Economics Letters*, Vol. 100, pp. 9–12.

Ravallion, M., Van De Walle, D., Gautam, M., (1995). Testing a Social Safety Net. *Journal of Public Economics*, Vol. 57, pp. 175–199.

Rubin, D. B., (1987). Multiple Imputation for Nonresponse in Surveys. *New York: Wiley*.

Shepherd, A., Wadugodapitiya, D., Evans, A., (2011). Social Assistance and the 'Dependency Syndrome'. *Chronic Poverty Research Centre Policy Brief,* No. 22, https://ssrn.com/abstract=1765933.

Szulc, A., (2012). Social Policy and Poverty: Checking the Efficiency of the Social Assistance System in Poland. *Eastern European Economics*, Vol. 50, pp. 66–93.

Szulc, A., (2022). Reconstruction of the Social Cash Transfers System in Poland and Household Well-being: 2015 – 2018 Evidence, SGH KAE Working Papers Series, No. 2022/076, Warsaw, https://cor.sgh.waw.pl/bitstream/handle/20.500.12182/1134/WPKAE_2022_076.pdf?sequence =2&isAllowed=y.

Wilk, S., (2020). The Role of Family Policy in Solving Demographic Problems: Study of the Polish Program Family 500+. *European Journal of Sustainable Development*, Vol. 9, pp. 84-92.

Wooldridge, J. M., (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, Vol. 141, pp. 1281–1301.

Wooldridge, J. M., (2010). Econometric Analysis of Cross Section and Panel Data. *The MIT Press,* Cambridge*, Massachusetts,* London, England.

STATISTICS IN TRANSITION new series, September 2024
Vol. 24, No. 3, pp. 31–47, https://doi.org/10.59170/stattrans-2024–026
Received – 30.08.2021; accepted – 11.01.2024

On autoregressive processes with Lindley-distributed innovations: modeling and simulation

K. U. Nitha[1], **S. D. Krishnarani**[2]

Abstract

In this paper, we develop an autoregressive process of order one, assuming that the innovation random variable has a Lindley distribution. The key properties of the process are investigated. Five distinct estimation techniques are used to estimate the parameters and simulation studies are conducted. The stationarity of the process is tested using a unit root test. The application of the proposed process to the analysis of time series data is demonstrated using real data sets. Based on some important statistical measures, the analysis of the data sets reveals that the proposed model fits well, and the errors a re independent and Lindley-distributed.

Key words: AR(1) process, Lindley distribution, innovations.

1. Introduction

In recent years, the development of information technology has brought changes in the structure of data sets. Practitioners and data analysts have been reformulating models for analyzing such data sets. Time series analysis is concerned with data that emerge over time. The normality assumption of the innovation random variables is one of the assumptions in the analysis of time series data. This assumption is always questionable because it is far away from reality in a number of applications. As a result, researchers have been experimenting with various models based on competitive and better distributional assumptions for the innovation random variable. One of the important models for time series data analysis is the linear autoregressive process of order 1 (AR(1)) model, in the form $X_t = aX_{t-1} + \varepsilon_t$. As stated previously, the innovation random variable ε_t can be non-normally distributed, particularly in finance, hydrology, economics, and biological sciences. EAR(1) and GAR(1) models of Gaver and Lewis (1980), Andel (1988), Hutton (1990), and Jenny and Vance (1992) are some examples in this regard. See Tiku et al. (2000), Ghasami et al. (2020), Altun (2019a), Sharafi and Nematollahi (2016), and references therein for more recent attempts in this category. For the estimation methods and further studies see, Bell and Smith (1986). So, the non-normality of the observed series and of the innovations are to be taken care of while modeling time series data sets. One may recall that these kinds of situations were confronted in the case of minification models, volatility models, and other non-Gaussian time series models. This motivates us to consider an AR(1) model with non-Gaussian error. In this paper, we consider an AR(1) model with a non-normal, specifically Lindley-distributed,

[1]Farook College (Autonomous), University of Calicut, Kerala-673632, India. E-mail: nithajilesh@gmail.com
[2]Department of Statistics, University of Calicut, Kerala-673635, India. E-mail: krishnaranisd@gmail.com. ORCID: https://orcid.org/0000-0003-3584-2892.

innovation distribution. Several researchers have discussed the significance of this distribution, and the details are provided below.

A Lindley distribution with parameter θ is a mixture of exponential(θ) and gamma(θ,2) distributions. It has numerous advantages over other distributions, which several authors have thoroughly researched. The Lindley distribution, having support on the positive real line is introduced by Ghitany et al. (2008). Further, there has been extensive and detailed research conducted into its generalizations. Algarni (2021), Altun (2019b), Asgharzadeh et al. (2016), Bhati et al. (2015), Hamed and Alzaghal (2021), Ekhosuehi and Opone (2018), Elbatal et al. (2013), Oluyede and Yang (2015), Zeghdoudi and Bouchahed (2018) and Beghriche et al. (2022) are a few among them. However, time series applications of Lindley distribution are less explored except for a few listed in the next paragraph. The probability density function (p.d.f) of a random variable following the Lindley distribution $\varepsilon \overset{d}{\sim} L(\theta)$ is given by

$$f(\varepsilon) = \frac{\theta^2}{\theta+1}(1+\varepsilon)e^{-\theta\varepsilon}; \;\; \varepsilon > 0, \;\; \theta > 0. \tag{1}$$

The mean, variance, and characteristic function of the Lindley distribution are respectively given by

$$E(\varepsilon) = \frac{\theta+2}{\theta(\theta+1)}, \tag{2}$$

$$var(\varepsilon) = \frac{\theta^2+4\theta+2}{\theta^2(\theta+1)^2}, \tag{3}$$

$$and \;\; \phi_\varepsilon(s) = \frac{\theta^2(\theta-is+1)}{(\theta+1)(\theta-is)^2}. \tag{4}$$

Practitioners were using this particular distribution in reliability studies and modeling different non-negative data sets, from its beginning as seen in Sankaran (1970) and Ghitany et al. (2008). The applications of this distribution in time series are studied by introducing an autoregressive model with a Lindley distribution as marginal in Bakouch and Popovic (2016). In that model, the error distribution is a mixture distribution, which is easily derivable because of the assumptions of the marginal. The innovation series can thus be in closed form, but the marginal need not be for stationary time series data sets that may be characterized by an additive process. So, here, in this paper, an attempt is made to construct such models with innovation as Lindley, and explore further properties. The appealing algebraic formulation of the Lindley distribution is an added advantage in this study.

The paper is organized as follows. In Section 2, the first-order Lindley autoregressive error process is introduced. Estimation of the parameters and simulation studies are done in Section 3. Unit root test is carried out in Section 4 and the application of the model is illustrated with the help of two real data sets in Section 5. Concluding remarks are given in the last Section.

2. Lindley error process

Let $\{X_n, n \geq 1\}$ be a stationary process generated by an autoregressive model given by

$$X_n = aX_{n-1} + \varepsilon_n \tag{5}$$

where $0 < a < 1$, and $\{\varepsilon_n\}$ is a sequence of independent and identically distributed (i.i.d) random variables such that ε_n is independent of $X_i, (i < n)$. Here, we assume that the innovation sequence $\{\varepsilon_n\}$ follows Lindley probability distribution with parameter θ, denoted as $L(\theta)$. So, we call the process, defined by (5) as a Lindley error process of order 1, and it is abbreviated as $LER(1)$. Since an AR(1) process can be written as the sum of the innovation sequence $\{\varepsilon_n\}$, we may write X_n as,

$$X_n = \varepsilon_n + a\varepsilon_{n-1} + ... + a^{n-1}\varepsilon_1 + a^n X_0 \tag{6}$$

$$= a^n X_0 + \sum_{r=0}^{n-1} a^r \varepsilon_{n-r}, \quad \text{as } n = 1,2,3,... \tag{7}$$

This can also be written as an infinite sum,

$$X_n = \sum_{r=0}^{\infty} a^r \varepsilon_{n-r}.$$

Therefore, the characteristic function corresponding to X_n is given by,

$$\phi_{X_n}(s) = \prod_{r=0}^{\infty} \frac{\theta^2(\theta - ia^r s + 1)}{(\theta+1)(\theta - ia^r s)^2}. \tag{8}$$

Because of the complexity in the structure of (8), it is strenuous to identify the distribution of X_n analytically. We derive the basic analytical properties of $\{X_n\}$ defined by (5), using the distributional properties of $\{\varepsilon_n\}$. For this, let us assume, $\mu_0 = E(X_0)$, $\sigma_0^2 = var(X_0)$, $\mu_1 = E(\varepsilon_n)$ and $\sigma_1^2 = var(\varepsilon_n)$. Using (7), we obtain,

$$E(X_n) = a^n \mu_0 + \frac{1-a^n}{1-a}\mu_1. \tag{9}$$

Similarly, the variance of X_n is given by

$$var(X_n) = a^{2n}\sigma_0^2 + \frac{1-a^{2n}}{1-a^2}\sigma_1^2 \tag{10}$$

and covariance between X_n and X_{n+k} is,

$$cov(X_n, X_{n+k}) = a^k \left[\sigma_0^2 a^{2n} + \sigma_1^2 \frac{1-a^{2n}}{1-a^2}\right]. \tag{11}$$

Defining M_n, the average of X_n's as

$$M_n = \frac{X_0 + X_1 + ... + X_{n-1}}{n}, n = 1, 2, ...,$$

and using the recursive form of X_n,

$$\begin{aligned} M_n &= \frac{1}{n}\left(X_0 + aX_0 + \varepsilon_1 + a^2X_0 + a\varepsilon_1 + \varepsilon_2 + ... + a^{n-1}X_0 + \sum_{r=0}^{n-2} a^r\varepsilon_{n-r-1}\right) \\ &= \frac{1}{n}\left[\frac{1-a^n}{1-a}X_0 + \sum_{r=1}^{n-1}\frac{1-a^r}{1-a}\varepsilon_{n-r}\right]. \end{aligned}$$

Therefore,

$$E(M_n) = \frac{1}{n(1-a)}\left[(1-a^n)\mu_0 + \mu_1\left((n-1) - \frac{a(1-a^{n-1})}{1-a}\right)\right]. \tag{12}$$

Similarly, the variance

$$var(M_n) = \frac{1}{n^2(1-a)^2}\left[(1-a^n)^2\sigma_0^2 + \sigma_1^2\left((n-1) - \frac{2a(1-a^{n-1})}{1-a} + a^2\frac{1-a^{2(n-1)}}{1-a^2}\right)\right]. \tag{13}$$

But when n becomes very large, from (9) and (12), we have both,

$$E(X_n) \quad \text{and} \quad E(M_n) \longrightarrow \frac{\mu_1}{1-a}.$$

From (10) and (13), $var(X_n)$ and $var(\sqrt{n}M_n) \longrightarrow \frac{\sigma_1^2}{1-a^2}$.

Using (11), $cov(X_n, X_k) = 0$ as $n, k \longrightarrow \infty$.

As $n \to \infty$, $cov(X_n, X_{n+k}) = \frac{a^k}{1-a^2}\sigma_1^2$, and $corr(X_n, X_{n+k}) = a^k$.

In the next section, we discuss different estimation methods of parameters in the $LER(1)$ process.

3. Estimation methods of LER(1) process

3.1. Maximum likelihood estimation

In this section, we focus on estimating the parameters of the proposed process. The parameters involved in the process are θ and a. We take the index set in the LER(1) process defined in (5) as t with the structure, $X_t = aX_{t-1} + \varepsilon_t$. For the realizations $x_1, x_2, ..., x_n$ we can write the likelihood function of the process as,

$$L = \left(\frac{\theta^2}{\theta+1}\right)^n \left(\prod_{t=2}^{n}(1 + x_t - ax_{t-1})\right) e^{-\theta\sum_{t=2}^{n}(x_t - ax_{t-1})}.$$

The corresponding log-likelihood can be written as

$$logL = nlog(\theta^2) - nlog(\theta+1) + \sum_{t=2}^{n} log(1+x_t - ax_{t-1}) - \theta \sum_{t=2}^{n} (x_t - ax_{t-1}). \tag{14}$$

Differentiating the log-likelihood equation (14) with respect to the parameters, we get the first-order partial derivatives. It should be noted that if the second-order partial derivatives are negative, the critical point corresponds to the maximum point. So, we found the first and second-order partial derivatives as,

$$\frac{\partial logL}{\partial a} = -\sum_{t=2}^{n} \frac{x_{t-1}}{(1+x_t - ax_{t-1})} + \theta \sum_{t=2}^{n} x_{t-1}, \tag{15}$$

$$\frac{\partial logL}{\partial \theta} = \frac{2n}{\theta} - \frac{n}{\theta+1} - \sum_{t=2}^{n} (x_t - ax_{t-1}), \tag{16}$$

$$\frac{\partial^2 logL}{\partial a^2} = -\sum_{t=2}^{n} \left(\frac{x_{t-1}}{1+x_t - ax_{t-1}} \right)^2,$$

$$\frac{\partial^2 logL}{\partial \theta^2} = \frac{-2n}{\theta^2} + \frac{n}{(\theta+1)^2}.$$

Next, we equate (15) and (16) to zero to get the estimates of the parameters. However, an explicit form for the parameter estimators is not derivable from the above expressions. So, numerical maximization of the log-likelihood function is the next possible alternative. The Nelder-Mead optimization in R software is used to maximize the log-likelihood function to get the maximum likelihood estimates.

Next, we use the well-known method of moments to estimate the parameters.

3.2. Method of moments

The moment estimators are identified by equating the sample moments (m_n) with corresponding population moments of the process, which are given by,

$$m_n = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{(1-a)} \frac{\theta+2}{\theta(\theta+1)} \tag{17}$$

$$s_n = \frac{1}{n-1} \sum_{i=1}^{n-1} (x_i - m_n)^2 = \frac{1}{(1-a^2)} \frac{\theta^2+4\theta+2}{\theta^2(\theta+1)^2} \tag{18}$$

After certain algebraic calculations, we obtain,

$$\widehat{\theta} = \frac{-((1-a)m_n - 1) + \sqrt{((1-a)m_n - 1)^2 + 8(1-a)m_n}}{2(1-a)m_n}.$$

	Maximum likelihood estimate				Non-parametric approach			
True values	$a = 0.4$	$\theta = 0.1$			$a = 0.4$	$\theta = 0.1$		
Sample size	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$	$\tilde{a}$	$\tilde{\theta}$	$MSE(\tilde{a})$	$MSE(\tilde{\theta})$
30	0.4216	0.1134	0.1236	0.0161	0.4621	0.1126	0.0423	0.0162
50	0.4205	0.1149	0.0401	0.0106	0.4330	0.1065	0.0267	0.0122
100	0.4174	0.1038	0.0186	0.0070	0.4224	0.1038	0.0161	0.0076
250	0.4128	0.1023	0.0081	0.0040	0.4113	0.1021	0.0094	0.0044
500	0.4126	0.1020	0.0063	0.0033	0.4060	0.1011	0.0050	0.0033
1000	0.4117	0.1017	0.0046	0.0017	0.4002	0.1008	0.0030	0.0020
True values	$a = 0.5$	$\theta = 3$			$a = 0.5$	$\theta = 3$		
Sample size	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$
30	0.5302	3.5121	0.1132	0.0573	0.5204	3.1327	0.0170	0.4395
50	0.5261	3.4005	0.1034	0.0390	0.5121	3.1793	0.0130	0.3845
100	0.5240	3.3962	0.0859	0.0281	0.5057	2.9842	0.0054	0.2263
250	0.5217	3.2473	0.0739	0.0179	0.5020	3.0333	0.0019	0.1681
500	0.5210	3.1440	0.0630	0.0137	0.5011	3.0114	0.0011	0.1037
1000	0.5203	3.0304	0.0211	0.0086	0.5004	3.0103	0.0005	0.0831
True values are	$a = 0.6$	$\theta = 1$			$a = 0.6$	$\theta = 1$		
Sample size	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$
30	0.6392	1.3057	0.0490	0.0254	0.6187	1.079	0.0188	0.0997
50	0.6251	1.2180	0.0261	0.0194	0.6092	1.0469	0.0072	0.0907
100	0.6238	1.2137	0.0130	0.0188	0.6055	1.0311	0.0066	0.0742
250	0.6223	1.1122	0.0075	0.0113	0.6022	1.0130	0.0021	0.0492
500	0.6220	1.1073	0.0047	0.0065	0.6010	1.0053	0.0011	0.0321
1000	0.6215	1.0106	0.0039	0.0039	0.6004	1.0021	0.0004	0.0241

Table 1: Estimated values of a, θ, and corresponding MSE.

Obtaining explicit answers for the parameters is complicated once again by the non-linear character of the aforementioned problem. As a result, we fixed the values of a in the simulation study, to assess the precision of the moment estimators of θ.
Now, we propose an alternative method of estimation based on a non-parametric approach.

3.3. Estimation through non-parametric approach

Here, we use the method discussed by Bell and Smith (1986) to identify the non-parametric estimate of a. Using (5) we have the ratios,

$$\frac{X_2}{X_1} = a + \frac{\varepsilon_2}{X_1}$$

$$\frac{X_3}{X_2} = a + \frac{\varepsilon_3}{X_2}$$

$$\cdots \quad \cdots$$

$$\frac{X_n}{X_{n-1}} = a + \frac{\varepsilon_n}{X_{n-1}}$$

Since each $\frac{\varepsilon_n}{X_{n-1}}$ is positive, it is clear that $a < \frac{X_t}{X_{t-1}}$, $t = 2,3,....,n$. So, a non-parametric point estimate of a is obtained as,

$$\tilde{a} = min\left(1, \frac{X_2}{X_1}, \frac{X_3}{X_2}, ..., \frac{X_n}{X_{n-1}}\right).$$

Correspondingly, $\tilde{\theta}$ is obtained as

$$\tilde{\theta} = \frac{-((1-\tilde{a})\tilde{\mu}-1) + \sqrt{((1-\tilde{a})\tilde{\mu}-1)^2 + 8(1-\tilde{a})\tilde{\mu}}}{2(1-\tilde{a})\tilde{\mu}}.$$

3.4. Conditional least square method

In this section of the paper, we use the conditional least square (CLS) method of estimation. The sum of squared deviations from conditional expectations $(E_n(\theta))$ is minimized to get the conditional least square estimators, where

$$E_n(\theta) = \sum_{t=2}^{n} [x_t - E(x_t|x_{t-1})]^2 \tag{19}$$

$$= \sum_{t=2}^{n} \left[x_t - ax_{t-1} - \frac{\theta+2}{\theta(\theta+1)}\right]^2. \tag{20}$$

In other words, the partial derivatives of these deviations with respect to the parameters a

and θ are equated to zero and the estimates are obtained by solving these two equations.

	True value	$\theta = 0.1$	
	Sample size	$\widehat{\theta}$	$MSE(\widehat{\theta})$
	30	0.0964	0.1241
	50	0.0942	0.0187
	100	0.0993	0.0171
$a = 0.4$	250	0.0998	0.0108
	500	0.1002	0.0072
	1000	0.1000	0.0051
	True value	$\theta = 3$	
a	Sample size	$\widehat{\theta}$	$MSE(\widehat{\theta})$
	30	2.8812	0.8009
	50	2.7423	0.5012
	100	2.9215	0.4766
$a = 0.5$	250	2.9381	0.2844
	500	2.9845	0.1988
	1000	2.9947	0.1623
	True value	$\theta = 1$	
a	Sample size	$\widehat{\theta}$	$MSE(\widehat{\theta})$
	30	0.8732	0.2076
	50	0.8977	0.2042
	100	0.9420	0.1663
$a = 0.6$	250	0.9714	0.1081
	500	0.9785	0.0765
	1000	0.9983	0.0457

Table 2: Moment estimate of θ and its MSE.

3.5. Gaussian estimation method

Finally, this study also employs the Gaussian estimation methodology, a frequently utilized estimation method in time series research. Bakouch and Popovic (2016) used this method for the estimation of the parameters of the first order autoregressive model of the Lindley distribution. The conditional maximum likelihood function is given by,

$$L(a,\mu,\sigma,\lambda) = f(x_1)\prod_{t=2}^{n} f(x_t|x_{t-1}).$$

Here, $f(x_t|x_{t-1})$ and $f(x_1)$ are the conditional and marginal probability function of $X_t|X_{t-1}$ and X_1 respectively.

Now,

$$log(L(a,\mu,\sigma,\lambda)) = nlog\frac{1}{\sqrt{2\pi}} - \frac{1}{2}\sum_{t=2}^{n}\left(log\sigma^2_{x_{t-1}} + \frac{(x_t-\mu_{x_{t-1}})^2}{\sigma^2_{x_{t-1}}}\right) \tag{21}$$

where

$$\mu_{x_{t-1}} = E(X_t|X_{t-1}) = ax_{t-1} + \frac{\theta+2}{\theta(\theta+1)}$$

and

$$\sigma^2_{x_{t-1}} = Var(X_t|X_{t-1}) = \frac{1}{n-1}\sum_{t=2}^{n}(x_t - \mu_{x_{t-1}})^2 = \frac{1}{(1-a^2)}\frac{\theta^2+4\theta+2}{\theta^2(\theta+1)^2}$$

are the conditional mean and conditional variance respectively. The estimated values are the points at which the likelihood function is maximum. We use the optimization package nlminb() in R to estimate the parameters.

We conducted simulation studies to examine the efficacy of the estimating methods outlined above. To estimate the values of θ and the autoregressive parameter a, we simulated samples from the proposed process with sizes of 30, 50, 100, 250, 500, and 1000. The various values of the parameter a considered are 0.4, 0.5, and 0.6, and that of θ are 0.1, 1 and 3. In each instance, the experiment is run 100 times, and the mean of the estimates is taken as the estimated parameter values. The estimated values and corresponding mean square errors (MSE) for the maximum likelihood and non-parametric approach are displayed in Table 1. The estimates for the method of moments are shown in Table 2, and the estimates for the Gaussian approach and the CLS method are shown in Table 3. It is important to note that the MSE reduces with sample size and that there are little disparities between the estimated parameter values and true values. While looking at the MSE values, it is seen that both the non-parametric and maximum likelihood methods perform equally good for the estimation of the parameter a, although for the estimation of θ the maximum likelihood method is the best.

4. Unit root testing in LER(1)

We conduct a testing procedure for the unit root of the process defined in (5), under the null hypothesis $H_0 : a = 1$ against the alternative hypothesis $H_1 : 0 < a < 1$. The test is used for checking whether the time series is stationary. Under H_0 the likelihood function is given by

$$L_0 = \left(\frac{\theta^2}{\theta+1}\right)^n \prod_{t=2}^{n}(1+x_t - x_{t-1})e^{-\theta\sum_{t=2}^{n}(x_t - x_{t-1})}$$

and the likelihood function corresponding to the alternative hypothesis is,

$$L_1 = \left(\frac{\theta^2}{\theta+1}\right)^n \prod_{t=2}^{n}(1+x_t - ax_{t-1})e^{-\theta\sum_{t=2}^{n}(x_t - ax_{t-1})}.$$

The corresponding likelihood ratio $\frac{L_0}{L_1}$ is,

$$\frac{L_0}{L_1} = \frac{\prod_{t=2}^{n}(1+x_t - x_{t-1})}{\prod_{t=2}^{n}(1+x_t - ax_{t-1})}e^{\theta\sum_{t=2}^{n}(x_{t-1} - ax_{t-1})}.$$

	Gaussian method				Conditional least square method			
True values	$a = 0.4$	$\theta = 0.1$			$a = 0.4$	$\theta = 0.1$		
Sample size	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$	$\tilde{a}$	$\tilde{\theta}$	$MSE(\tilde{a})$	$MSE(\tilde{\theta})$
30	0.4419	0.1158	0.0900	0.0256	0.3207	0.0955	0.1683	0.0253
50	0.4209	0.1129	0.0806	0.0189	0.3531	0.0979	0.1318	0.0208
100	0.4437	0.1121	0.0593	0.0133	0.3657	0.0961	0.0809	0.0123
250	0.4367	0.1097	0.0348	0.0088	0.4050	0.1025	0.0650	0.0118
500	0.4411	0.1095	0.0273	0.0655	0.3997	0.1004	0.0411	0.0073
1000	0.4387	0.1095	0.0246	0.0060	0.3984	0.0999	0.0264	0.0048
True values	$a = 0.5$	$\theta = 3$			$a = 0.5$	$\theta = 3$		
Sample size	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$
30	0.5444	3.7667	0.0913	0.8505	0.5212	3.0188	0.1653	0.8802
50	0.5397	3.6529	0.0773	0.5456	0.5453	2.8976	0.1563	0.6631
100	0.5420	3.5322	0.0490	0.4364	0.5532	2.8961	0.1407	0.4500
250	0.5332	3.4800	0.0322	0.2922	0.4907	3.0226	0.0380	0.2226
500	0.5583	3.1440	0.0156	0.1307	0.4952	3.0179	0.0278	0.2204
1000	0.5283	3.0010	0.0153	0.0086	0.4999	3.0074	0.0001	0.0831
True values	$a = 0.6$	$\theta = 1$			$a = 0.6$	$\theta = 1$		
Sample size	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$	$\widehat{a}$	$\widehat{\theta}$	$MSE(\widehat{a})$	$MSE(\widehat{\theta})$
30	0.6718	1.3802	0.0748	0.3389	0.5319	0.9793	0.1248	0.2942
50	0.6782	1.3524	0.0596	0.2727	0.5383	0.9793	0.1246	0.2741
100	0.6742	1.3239	0.0536	0.1959	0.5644	0.9563	0.0774	0.1614
250	0.6766	1.2968	0.0252	0.1057	0.5930	0.9950	0.0526	0.1190
500	0.6763	1.2942	0.0200	0.0795	0.5959	0.9992	0.0351	0.0788
1000	0.6749	1.2846	0.0125	0.0509	0.5945	0.9897	0.0226	0.0525

Table 3: Estimated values of a, θ, and corresponding MSE.

The log-likelihood ratio test statistic used by Wilks (1938) is of the form,

$$-2log(L_0/L_1) = -2\left[\sum_{t=2}^{n} log\left(\frac{1+x_t-x_{t-1}}{1+x_t-ax_{t-1}}\right) + \theta\sum_{t=2}^{n}(x_{t-1}-ax_{t-1})\right] \tag{22}$$

that follows χ^2 distribution with n degrees of freedom. As an illustration, we simulated $\{x_t\}$ values for different sample sizes from the proposed process for different choices of values of a and θ. Then we calculated the likelihood ratio statistic for various values of significance level (α), and the corresponding probability of rejecting the null hypothesis under the alternative. The numerical computations carried out for estimation and testing purposes illustrated above are shown in Table 4. The probabilities of rejection under the alternative hypothesis increase as the sample size increases, but when the parameter a is large and the sample size is small, we get small values for the probabilities of rejection. So, the test confirms the non-stationarity behavior of the model when a is unity. In the simulation study, we can see that all the five methods that have been considered perform equally well.

5. Real life applications

In this section, we apply the LER(1) process developed in the foregoing sections to two specific real data sets to enhance the credence in this non-normal time series modeling strategy.

Data set 1: We take 100 data points which are measurements of the annual flow of the river Nile at Aswan for a period from 1871 to 1970. These data are taken from the library of the base R package. As the first step is to check the stationarity of the data, we have plotted the sample path, but it appears the data are non-stationary. The Augmented Dickey-Fuller (ADF) test for stationarity has been done and the p-value is 0.0642, indicating stationarity at a 90 percent confidence level. Log transformation is used to stabilize the variance, and the current p-value of the ADF test is 0.0472, confirming that the data are stationary. The autocorrelation function (acf) and partial autocorrelation function (pacf) are plotted in Figure $1(a)$ and Figure $2(a)$ respectively. But acf decreases to zero and pacf is significant only at lag 1, indicating that the data can be modeled by an AR(1) process. We use LER(1) process for modeling these data and the estimated values of $\hat{\theta}$ and $\hat{a}$ obtained using the maximum likelihood method employing numerical optimization are given in Table 5. Then, the p-value for the Kolmogorov-Smirnov (K-S) test was obtained as 0.4415 (see Table 5) which substantiates that the LER(1) model is fit to the data. Figure $3(a)$ is the histogram of the residual series along with the estimated Lindley density function. For the diagnostic checking, acf and pacf sketching of residuals are shown in Figure $4(a)$. Table 5 includes the results of the Box-Pierce and Ljung-Box tests, and the p-values support our conclusion that the residuals are independent.

Data set 2: The second data set contains the Canada unemployment rate from 1955 to 2021 (https://data.oecd.org/unemp/unemployment-rate.htm). As in the case of Data set 1, all the plots, tests, etc. are done in this case also. The time series plot and the p-value of 0.023 of the ADF test indicate that the Canadian unemployment data are stationary. The AR(1)

process can be used to represent these data, as shown by the acf and pacf depicted in Figure 1(b) and Figure 2(b), respectively. Table 5 provides the values of $\hat{\theta}$ and $\hat{a}$. The p-value of the K-S test, which is 0.4537 in Table 5 authenticates the validity of the LER(1) model as a good fit. Figure 3(b) shows the histogram of the residual series and fitted Lindley density is appended to it, clearly indicating that the innovation sequence is Lindley distributed. The acf and pacf of the residuals are delineated in Figure 4(b). Box-Pierce and Ljung-Box tests of the residuals are also performed and yielded the p-values in Table 5, confirming the independence of the residuals.

	True values are	$a = 0.8$	$\theta = 6$	
Sample size	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$	$\alpha = 0.1$
30	0.418	0.510	0.521	0.703
50	0.459	0.632	0.703	0.862
100	0.516	0.720	0.815	0.912
150	0.610	0.76	0.86	0.952
300	0.952	0.99	0.996	0.998
500	0.998	1	1	1
1000	1	1	1	1
	True values are	$a = 0.4$,	$\theta = 2$	
Sample size	$\alpha = 0.01$	$\alpha = 0.025$	$\alpha = 0.05$	$\alpha = 0.1$
30	0.612	0.798	0.847	0.863
50	0.820	0.938	0.968	0.974
100	1	1	1	1
250	1	1	1	1
500	1	1	1	1
1000	1	1	1	1

Table 4: Probability of rejection of the null hypothesis under different sample sizes and α values.

Data Set	Parameter estimates	K-S test	p-values of Box-Pierce test	Ljung-Box test
Nile data	$\hat{a}$=0.99 $\hat{\theta}$=7.49	0.4415	0.1071	0.1070
Canada unemployment data	$\hat{a}$=0.96 $\hat{\theta}$=1.86	0.4537	0.2820	0.2711

Table 5: Results of the analysis of real data.

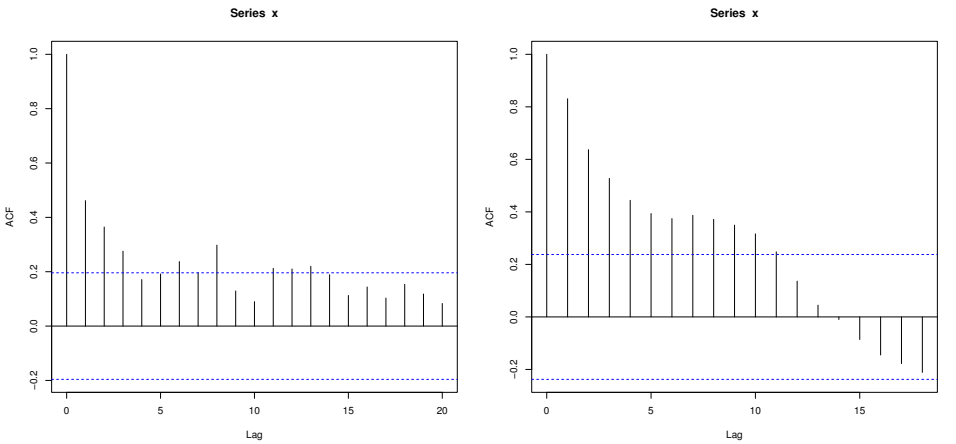


Figure 1: (a) acf of log transformed Nile data. (b) acf of Canada unemployment data.

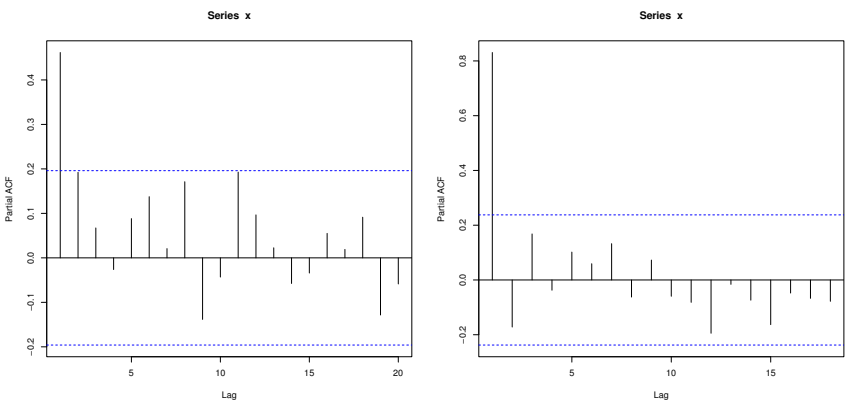


Figure 2: (a) pacf of log-transformed Nile data. (b) pacf of Canada unemployment data.

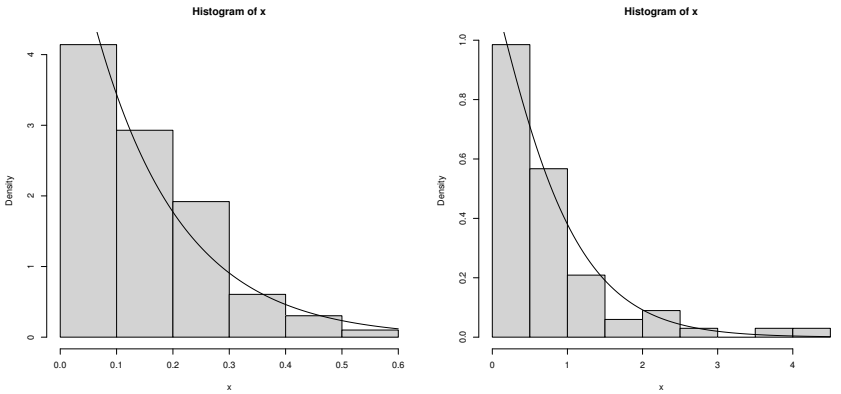


Figure 3: (a) Histogram and fitted Lindley distribution of LER(1) residuals of the Nile data. (b) Histogram and fitted Lindley distribution of LER(1) residuals of the Canada unemployment data.

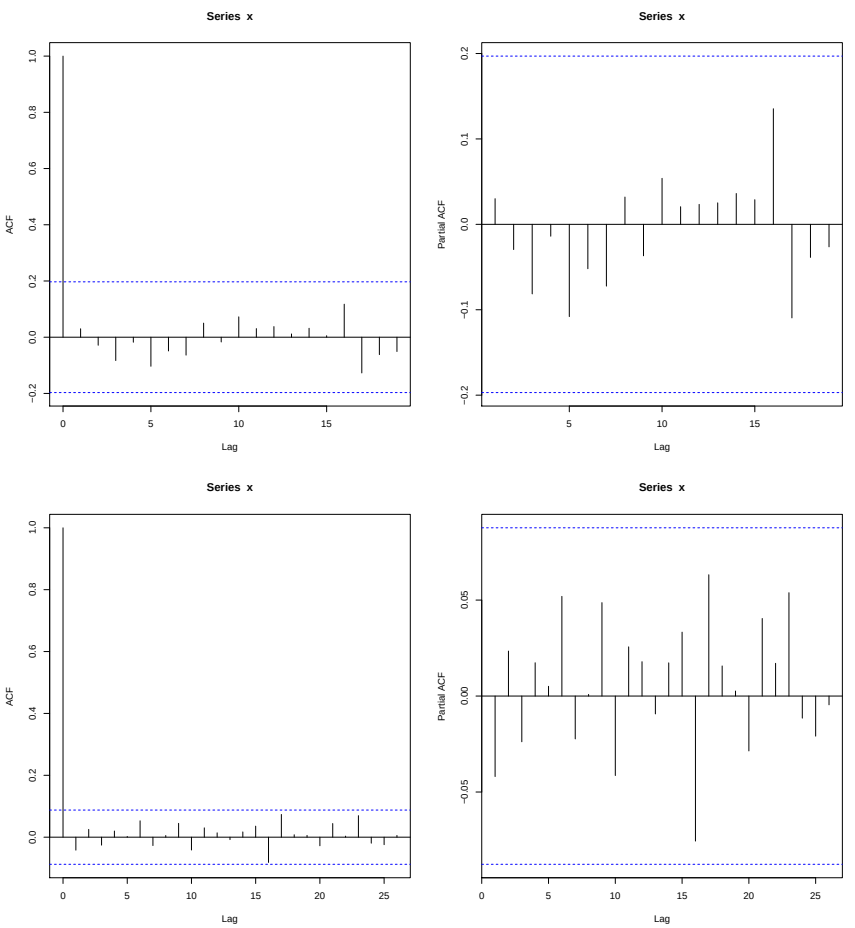


Figure 4: (a) acf and pacf plots of the residuals of fitted LER(1) to the Nile data. (b) acf and pacf plots of the residuals of fitted LER(1) to the Canada unemployment data.

6. Conclusion

The stationary series of additive autoregressive models could feature non-Gaussian errors and marginal. It would be worthwhile to look into using the Lindley distribution in time series modeling as it is a very versatile mixture distribution. In this study, we explored a first-order autoregressive model with the Lindley error distribution and its properties. Parametric and non-parametric estimating techniques are effectively employed. Additionally conducted are simulation studies and application to real-world instances. The proposed model is helpful in situations where the data are predictable in nature and the error is non-Gaussian, especially Lindley distributed. The structural or mathematical form that is produced has numerous practical uses. The model, estimating techniques, and real world examples used in this work may be extended to include non-linear modeling.

Acknowledgements

The authors sincerely thank the anonymous referees and editors for their valuable comments and suggestions.

References

Algarni, A., (2021). On a new generalized Lindley distribution: properties, estimation and applications. *PLOS ONE*, 16(2).

Altun, E., (2019a). A new generalization of geometric distribution with properties and applications. *Communications in Statistics Simulation and Computation*, 17(5), pp. 1481–1495.

Altun, E., (2019b). Two-sided Lindley distribution with inference and applications. *Journal of the Indian Society for Probability and Statistics*, 20, pp. 255–279.

Andel, J., (1988). On AR(1) processes with exponential white noise. *Communications in Statistics - Theory and Methods*, 17(5), pp. 1481–1495.

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

Bakouch, H. S., Popović, B. V., (2016). Lindley first-order autoregressive model with applications. *Communications in Statistics - Theory and Methods*, 45(17), pp. 4988–5006.

Beghriche, A., Zeghdoudi, H., Raman, V., Chouia, S., (2022). New polynomial exponential distribution: properties and applications. *Statistics in Transition New Series*, 23(3), pp. 95–112.

Bell, C. B., Smith, E. P., (1986). Inference for non-negative autoregressive schemes. *Communications in Statistics-Theory and Methods*, 15(8), pp. 2267–2293.

Bhati, D., Malik, M., Vaman, H., (2015). Lindley-exponential distribution: properties and applications. *Metron*, 73(3), pp. 335–357.

Ekhosuehi, N., Opone, F., (2018). A three-parameter generalized Lindley distribution: properties and application. *Statistica*, 78(3), pp. 233–249.

Elbatal, I., Merovci, F., Elgarhy, M., (2013). A new generalized Lindley distribution. *Mathematical Theory and Modeling*, 3(13), pp. 30–47.

Gaver, D. P., Lewis, P. A. W., (1980). First order autoregressive gamma sequences and point processes. *Advances and Applications in Probability*, 12, pp. 727–745.

Ghasami, S., Khodadadi, Z., Maleki, M., (2020). Autoregressive processes with generalized hyperbolic innovations. *Communications in Statistics-Simulation and Computation*, 49(12), pp. 3080–3092.

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

Hamed, D., Alzaghal, A., (2021). A new class of Lindley distribution: properties and applications. *Journal of Statistical Distributions and Applications*, 8(11).

Hutton, J. L., (1990). Non-negative time series models for dry river flow. *Journal of Applied Probability*, 27, pp. 171–182.

Jenny, N. L., Vance, L. M., (1992). Non-linear time series modelling and distributional flexibility. *Journal of Time Series Analysis*, 65, pp. 65–84.

Oluyede, B., Yang, T., (2015). A new class of generalized Lindley distribution with applications. *Journal of Statistical Computation and Simulation*, 85(10), pp. 2072–2100.

Sankaran, M., (1970). The discrete Poisson–Lindley distribution. *Biometrics*, 26(1), pp. 145–149.

Sharafi, M., Nematollahi, A. R., (2016). AR(1) model with skew-normal innovations. *Metrika*, 79(8), pp. 1011–1029.

Tiku, M. L., Wong, W. K., Bian, G., (2000). Time series models with non-normal innovations symmetric location-scale distributions. *Journal of Time Series Analysis*, 21(5), pp. 571–596.

Wilks, S. S., (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), pp. 60–62.

Zeghdoudi, H., Bouchahed, L., (2018). A new and unified approach in generalizing Lindley's distribution with applications. *Statistics in Transition New Series*, 19(1), pp. 61–74.

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 49–67, https://doi.org/10.59170/stattrans-2024–027
Received – 31.08.2023; accepted – 30.05.2024

Comparing logistic regression and neural networks for predicting skewed credit score: a LIME-based explainability approach

Jane Wangui Wanjohi[1], **Berthine Nyunga Mpinda**[2], **Olushina Olawale Awe**[3]

Abstract

Over the past years, machine learning emerged as a powerful tool for credit scoring, producing high-quality results compared to traditional statistical methods. However, literature shows that statistical methods are still being used because they still perform and can be interpretable compared to neural network models, considered to be black boxes. This study compares the predictive power of logistic regression and multilayer perceptron algorithms on two credit-risk datasets by applying the Local Interpretable Model-Agnostic Explanations (LIME) explainability technique. Our results show that multilayer perceptron outperforms logistic regression in terms of balanced accuracy, Matthews Correlation Coefficient, and F1 score. Based on our findings from LIME, building models on imbalanced datasets results in biased predictions towards the majority class. Model developers in the field of finance could consider explanation methods such as LIME to extend the use of deep learning models to help them make well-informed decisions.

Key words: credit score, logistic regression, multilayer perceptron, explainability, LIME.

1. Introduction

The credit section is a critical function for financial institutions and banks as it contributes significantly to their revenue share. One of the ways they maximize their profits is by offering more credit. With increased competition and pressure to generate more revenue, financial institutions are searching for more effective ways to attract new creditworthy customers while minimizing losses. Credit approval puts the bank at risk of losing money while disapproval may lead to loss of customers, hence competition among lenders. Most financial institutions have suffered losses due to wrong decision-making in the past led by many of their customers defaulting on payments as discussed by Taghavi et al. (2015). This has created the need for mechanisms to identify and distinguish between eligible and non-eligible loan applicants (Bolton C. 2009). In the early years, lenders used personal relationships and subjective judgments to decide whether the applicant deserved a loan offer

[1]African Institute for Mathematical Sciences (AIMS), Cameroon. E-mail: jane.wanjohi@aims-cameroon.org. ORCID: https//orcid.org/0009-0004-3295-6426.
[2]African Institute for Mathematical Sciences (AIMS), Cameroon. E-mail: bmpinda@aimsammi.org. ORCID: https//orcid.org/0000-0003-2981-6436.
[3]State University of Campinas, Brazil. E-mail: oawe@unicamp.br. ORCID: https//orcid.org/0000-0002-0442-4519.

or not. The method was reliable and less time-consuming because of the smaller number of applicants. Prior information about the applicant is important in determining credit scores (Lahsasna et al. 2010).

A comparison study between Radial Basis Function (RBF) neural networks and LR was carried out by Taghavi et al. (2015) in Tose-Taavon Bank, Guilan. The study involved 376 cases of loan instalments with 7652 total credits within approximately five years. The cases were categorized into two groups based on credit risk: good customers were those with lower credit risk and bad customers with higher credit risk. Out of the total 376 cases, 302 were classified as good customers, while the remaining 74 cases belonged to bad customers. To evaluate the efficiency of the designed LR model, the study utilized Pearson correlation analysis to identify any significant relationship between the variables. The results indicated a significant negative correlation between the loan amount and the credit decision of bank customers, and a significant positive correlation between the number of installments and the credit decisions of bank customers, both at 99% confidence level. RBFs were trained with 320 samples and 50 samples were used for testing. RBFs showed a higher prediction accuracy of 88% than 82.7% of logistic regression. The study recommended efficient databases to store customers' information for easy access.

Dumitrescu et al. (2022), proposed a new credit scoring method called penalized logistic tree regression (PLTR), which combined decision trees and logistic regression to improve the accuracy of credit risk prediction. The paper also discussed the need for interpretability in the credit scoring industry, which showed why simpler models like logistic regression were still widely used despite their limitations compared to more complex machine learning methods. Several studies have shown that neural networks are more accurate, and logistic regression has performed better than neural networks on different datasets. This means there is a need for further investigation of the performance of neural networks and logistic regression in predicting credit scores. The application of deep learning methods in credit scoring has shown promising results as shown in Imtiaz and Brimicombe (2017) and Zhao et al. (2015) compared to traditional methods. Traditional methods like logistic regression have been widely used due to their simplicity and interpretability. Despite the effectiveness of neural networks, it is often difficult for stakeholders to understand and trust their decisions due to lack of transparency. This study aims to solve this problem by comparing the effectiveness of logistic regression and neural networks in predicting credit scores, using the LIME (Local Interpretable Model-agnostic Explanations) technique on binary and multiclass datasets to enhance the explainability of the neural network models. The goal is to determine which model provides a better balance between predictive accuracy and interpretability in the context of credit scoring.

This present study comparing Logistic Regression and Neural Networks for predicting credit scores using a LIME-based explainability approach offers several key contributions to predictive analytics and model interpretability. Firstly, it enhances understanding of how different models operate, especially in the context of finance where stakeholders require trust and clarity in automated decision-making processes. It carefully analyses the trade-offs

between the simplicity and transparency of logistic regression versus the accuracy and complexity of neural networks, aiding stakeholders in selecting the appropriate model based on their specific needs for accuracy and transparency. Furthermore, the study demonstrates the practical application of LIME, showcasing how local interpretable model-agnostic explanations can be effectively generated for complex models. This not only serves as a valuable guideline for other researchers and practitioners but also advances the field of explainable AI (XAI) by providing empirical insights into the effectiveness of explainability techniques across different models. The emphasis on explainability also promotes ethical AI practices, highlighting the importance of transparency in sensitive areas like credit scoring, which could influence policy and regulatory approaches.

Additionally, by making model decisions accessible to both technical and non-technical stakeholders, this study bridges a crucial gap, fostering communication and trust among model developers, policymakers, loan officers, and customers. Overall, these contributions are instrumental in advancing machine learning applications in finance, encouraging the responsible use of complex models while upholding high standards of accountability and transparency. Adding to this introduction, the rest of the work is organized as follows. In Section 2, we present a description of the mathematics behind the methods used in this study. Section 3 presents the analyses and results of our study, Section 4 gives the explainability with LIME and Section 5 gives the conclusion.

2. Methodology

The goal of this section is to present the mathematical description of the methods and models used for the validity of our results. The flowchart in Figure 1 illustrates the various processes performed in this study.

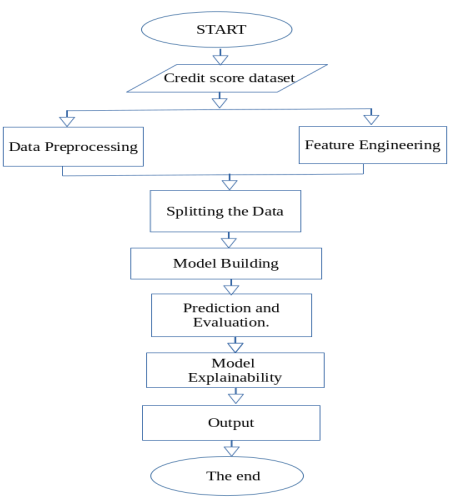


Figure 1: Credit Scoring Flowchart.

2.1. Data Preparation

2.1.1 Missing Value Imputation using MICE

Multiple Imputation by Chained Equations (MICE) is used as a data imputation technique to handle missing values in this study. The method assumes that values are missing at random implying that other columns can be used to determine a missing observation (Little and Rubin 2019). This is achieved by investigating acceptable estimates that best approximate the missing value using methods such as regression as discussed by Wulff and Jeppesen (2017). MICE creates multiple imputed datasets, where each fills in missing observations with valid data.

Given $X = (X_1, X_2, \ldots, X_k)$, a set of k features and each feature $X_j = (X_j^{observ}, X_j^{mis})$ where X_j^{observ} and X_j^{mis} are present and missing values respectively. The data imputation challenge is to get the conditional multivariate density $P(X)$ of X. Let t denote the number of iterations. With the assumption that data are missing at random, we repeat the Gibbs sampler iteration in Scheuren(2005);

$$\begin{aligned} X_1 &\sim P(X_1|X_2^t, X_3^t, \ldots, X_k^t) \\ X_2 &\sim P(X_2|X_1^t, X_3^t, \ldots, X_k^t) \\ &\vdots \\ X_k &\sim P(X_1|X_2^t, X_3^t, \ldots, X_{k-1}^t) \end{aligned}$$

A study done by Van and Oudshoorn (2000) demonstrated that when dealing with missing data in a multivariate normal distribution, iterating linear regression models like $X_1 = X_2^t\theta_{12} + \ldots + X_k^t\theta_{1k} + \varepsilon_1$ with $epsilon \sim N(0, \sigma)$ and estimates θ, can be a powerful tool for imputation as adopted in this present study.

2.1.2 Feature Selection with RFECV

The Recursive Feature Elimination with Cross-Validation (RFECV) method is applied for feature selection in this work. The method involves deleting features based on the importance it has on the model (Mustaqim et al. 2021). The final subset of selected features is chosen based on the performance of the model trained on the selected features using an independent validation set or through a nested cross-validation procedure. Suppose X is the input data matrix and y is the target vector. The algorithm is as follows:

1. Initialize a model M using all available features in X and evaluate its performance using cross-validation.

2. Rank the features based on their importance according to M.

3. Remove the least important feature from X, and evaluate the performance of M using cross-validation. If the performance metric improves, keep the feature removed. If not, add the feature back to X.

4. Repeat steps 3-4 until the desired number of features is reached or the model performance can no longer be improved. RFECV helps reduce the dimension of the dataset and that improves the performance of the model (Misra and Yadav, 2020).

2.1.3 Data Standardization

Machine learning offers distinct advantages in addressing some stringent assumptions associated with traditional statistical methods like Maximum Likelihood Estimation (MLE). Many machine learning models like the ones adopted in this study do not require assumptions about the data's distribution and can handle complex, nonlinear relationships automatically. This flexibility allows them to adapt more naturally to the actual structure of the data. Machine learning techniques such as logistic regression (borrowed from statistical methods) and neural networks are particularly adept at managing non-linearities and interactions without explicit modeling. Moreover, these methods often demonstrate robustness against outliers and noisy data, a common challenge in real-world datasets. In this study, we standardized the data before analysis to mitigate the possible effect of outliers. The mean is subtracted from each feature to standardize the data and then divided by the standard deviation. Standardization was used in our study to ensure all the inputs were on the same scale to avoid creating biased models. Given a dataset with n observations and p features, X_j ,$j = 1,2,\ldots p$, the mean and standard deviation of feature j is shown in Equations 1 and 2 respectively.

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\sigma_j = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_j)^2 \tag{2}$$

where μ_j and σ_j are the mean and standard deviation of feature j respectively, x_i is the i-th observation in feature j, and n is the total observations. Standardization of an observation i of feature j using z-score normalization is expressed as:

$$\mathbf{X}_{\text{standardized}} = \frac{x_i - \mu_j}{\sigma_j} \tag{3}$$

Standardization is an important technique in machine learning that helps improve the robustness and generalizability of models.

2.1.4 Data Balancing with SMOTE

Synthetic Minority Oversampling (SMOTE) was applied to address class imbalance. The method was first introduced by Chawla et al. (2002). It creates synthetic samples in the minority class for fair distribution between the classes. This method works closely as K-Nearest Neighbours (KNN). Let X be the input matrix with minority class instances, k be the number of nearest neighbours in consideration and m be the synthetic samples to generate. Let x_i be a minority data point x_{ij} an i^{th} observation of j^{th} feature and k random

selected neighbours. The new feature vector $X_{i,new}$ is given as:

$$X_{j,new} = X_j + (X_j - X_r) \times \alpha, \tag{4}$$

where X_j is the original feature vector, X_r is the feature vector selected at random and α is a random number between 0 and 1. SMOTE oversampling helps overcome overfitting observed in random sampling. Applying this technique enhances the accuracy of classifiers.

2.2. Classification Algorithms

Credit score prediction is a classification task in classification algorithms that classifies a customer as a defaulter or a non-defaulter. In supervised learning, the machine learns from labelled data automatically and improves its prediction capability with experience. The target label depends on the input features in the data and these inputs are meant to give accurate predictions of unseen data. Given a set of features $x_i \in X$ the model seeks to find a function f that maps the features to the output $y_i \in Y$, $f : X \to Y$. This function is used to make predictions of new data after it has learned from a set of labelled data. The model seeks also to get a function that gives a minimal difference between Y and the models' prediction $f(X)$, for $x \in X$. This study focuses on two supervised learning classification algorithms; logistic regression and multilayer perceptron.

2.2.1 Logistic Regression

Logistic Regression is a supervised classification algorithm applied when predicting a dependent categorical variable. The algorithm captures the probability of an outcome (e.g., 0 or 1) as a function of one or more input variables. The output of the logistic regression model is a probability value between 0 and 1. The algorithm is an extension of linear regression where the sigmoid function transforms the linearity (Imtiaz and Brimicombe 2017). There are various types of logistic regression. This study utilizes binary and multinomial logistic regression. The approximated value of y as a linear function of x is given as:

$$g_\beta(x) = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \beta_0 + \sum_{j=0}^{p} \beta_j x = \beta^{'} x, \tag{5}$$

where $g_\beta(x)$ is a linear combination of the input variables $x_1, x_2, \ldots, x_p$ and their associated weights $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ and $x_0 = 1$. As we saw earlier in supervised learning the main goal is to find a function that captures the relation between the input and output data. The sigmoid function transforms $g_\beta(x)$ in such a way that it takes values between 0 and 1 in the form:

$$h(\beta^{'} x) = \frac{1}{1 + e^{-\beta' x}}, \tag{6}$$

$$h(z) = \frac{1}{1 + e^{-z}}, \tag{7}$$

where $z = \beta^{'} x$ and $h(z)$ is the sigmoid or logistic function. From the notion of statistics, the regression coefficients (i.e. the parameters) provide information about each indepen-

dent variable's influence on the variations in the response variable (Hossain 2022). The conditional probabilities of the labels (0 & 1) for a given observation i are defined by:

$$p(y_i|x;\beta) = (g_\beta(x))^{y_i}(1-g_\beta(x))^{1-y_i}. \tag{8}$$

The likelihood function is used to find the values of parameters that maximize the likelihood of the observed data to train logistic regression (Hand and Henley 1997, Sewpaul et al. 2023), Mahmood 2024). The observations are taken to be independent. This function is given by:

$$L(\beta_0,\beta_1,\beta_2,\ldots,\beta_p) = \prod_{i=1}^{p}((g_\beta(x))^{y_i}(1-g_\beta(x))^{1-y_i}). \tag{9}$$

where p is the number of observations in the training data, y_i is the binary outcome for the i-th observation (0 or 1), and $g_\beta(x)$ is the linear combination of input variables with their associated weights for the i-th observation (Taghavi et al. 2015). An optimization algorithm such as gradient descent is then used to estimate the values of $\beta^{'}$. Gradient descent iteratively updates the parameters and finds the values that minimize the loss function (Zhao et al. 2015). The gradient of the log-likelihood function with respect to the parameters is given by:

$$\nabla L(\beta_0,\beta_1,\beta_2,\ldots,\beta_p) = \sum_{i=1}^{n}(g_\beta(x)-y)x_i. \tag{10}$$

The loss function is given by:

$$\mathscr{L}(\beta_0,\beta_1,\beta_2,\ldots,\beta_p) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i\log(g_\beta(x)) + (1-y_i)\log(1-g_\beta(x))\right]. \tag{11}$$

For prediction, a trained logistic regression model takes in a new instance and gives a probability output. A threshold such as 0.5 is set to make a binary prediction. The probability of the classes present adds up to one.

2.2.2 Multilayer Perceptron

Multilayer Perceptron (MLP) is commonly used in credit scoring. Unlike a percepton, MLP contains more than one hidden node or layer. MLP comprises three types of layers; the input, hidden and the output layer. Each layer contains interconnected nodes that transmit signals. The behaviour of hidden neurons is influenced by the input units and the weights connecting them, while the output neurons' behaviour is determined by the activities of the hidden neurons and the weights connecting them to the output neurons (Rodrigues et al. 2020). The main intention in neural networks is to get the ultimate parameters that best predict an output. The first step is forward propagation where each attribute is fit in the input layer. These inputs are assigned to random parameters called weights and passed to the first hidden layer. The hidden layer does some processes, like applying an activation function to each neuron to determine the output, passing this output to the next hidden layer for the same process, and then passing it to the output layer. The output layer gives the

predicted outcome, as shown in Equation 12:

$$\widehat{y_i} = h(\sum_{i=1}^{n} w_i x_i + b), \tag{12}$$

where $\widehat{y_i}$, h, w_i, x_i and b are the output, activation function, weights, inputs and the bias term respectively. The loss function is then used to compare the predicted output with the actual output. The cross-entropy loss function is used for classification purposes as shown in Zhang et al. (2023):

$$crossentropy = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} y_{i,j}\log(\widehat{y}_{i,j}), \tag{13}$$

where n is the number of inputs, p represents the number of different classes present in Y, y_{ij} is the actual value and $\widehat{y}_{ij}$ is the predicted value. An optimizer is then introduced to the network. Optimizers are computational techniques used to adjust certain parameters of a neural network, such as the weights and bias term, to minimise the loss function and achieve precise outcomes. The learning rate, a hyperparameter, determines the magnitude of the step taken at each iteration as the algorithm moves towards a minimum of the loss function. Gradient descent is the most commonly used optimizer in neural networks (Agarwal et al., 2021). Equation 14 shows how the value of a parameter w_{ij} is updated, by subtracting the product of the learning rate $\eta, 10^{-}6 < \eta < 1$ and the partial derivative of the loss function $\mathscr{L}$ with respect to w_{ij} from its current value. This process is repeated for each parameter during training to minimize the loss function and improve the neural network's performance. The learning rate controls how much the parameters of a multilayer perceptron are updated during training.

$$w_{ij+1} = w_{ij} - \eta\frac{\partial\mathscr{L}}{\partial w_{ij}}, \tag{14}$$

$$b_{ij+1} = b_{ij} - \eta\frac{\partial\mathscr{L}}{\partial b_{ij}}, \tag{15}$$

where, w_{ij+1}, w_{ij}, η, and $\mathscr{L}$ are the new weights of observation i feature j, initial weight of observation i feature j, the learning rate and the loss function respectively. b_{ij+1} in Equation 15 is the updated bias term. The collective process of adjusting the parameters using an optimizer, computing the gradient of the loss function, and propagating the error backwards through the network is known as backpropagation as discussed in Zhang et al. (2023). After a multilayer perceptron is trained, it can be used for the prediction of unseen data and the output is determined by the learned parameters.

2.3. Performance Evaluation Metrics

Evaluation metrics are intended to estimate the model's ability to generalize unseen data in this study. Below are the metrics used to evaluate binary classification tasks in this work. A confusion matrix calculates evaluation metrics such as accuracy, balanced

accuracy, precision, and F1 score.

Table 1: Representation of a Confusion Matrix.

Class	**Positive Prediction**	**Negative Prediction**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The accuracy(AC) performance metric gives the proportion of the correct predictions out of all the predictions made. The accuracy of a classification model is given as:

$$AC = \frac{TP+TN}{TP+TN+FP+FN}. \tag{16}$$

The accuracy metric lies between, 0 and 1, where values close to 1 show a good model performance. The accuracy metric is a commonly used evaluation metric. However, the metric is sensitive to imbalanced data. In cases of imbalanced data, balanced accuracy is a more efficient evaluation metric.

Balanced Accuracy (Bal AC) metric gives the average of the true positive rate (sensitivity) and the true negative rate (specificity). It ranges from 0 to 1. The metric is expressed as:

$$BalAC = \frac{Sensitivity + Specificity}{2}. \tag{17}$$

With

$$Sensitivity = \frac{TP}{TP+FN}$$

and

$$Specificity = \frac{TN}{TN+FP}$$

.

Additionally, this metric guarantees that all classes present are considered which gives a better view of the model's performance.

Matthews Correlation Coefficient (MCC) is the association between the actual and the predicted classes. This metric is commonly used when dealing with imbalanced datasets because it considers true positives, false positives, and false negatives of the model's predictions. MCC is expressed as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \tag{18}$$

The value of MCC ranges from -1 to 1, with values close to 1 showing a good model performance. Additionally, an MCC score of 1 indicates a perfect prediction, 0 shows a random prediction, and -1 indicates a wrong prediction.

2.4. Machine Learning Explainability

The success stories in solving diverse problems have led to a rapid expansion of the field, often referred to as a "golden age" because its end cannot be seen (Tran et al. 2022). However, there are several challenges that need to be addressed for further improvement in the performance and applicability of machine learning. Interpretability is one of the major challenges that researchers are working hard to solve so that the models are easy to understand and explain how they function. Some of the main reasons for machine learning explainability are: to validate the previous studies made on these algorithms, to increase the trust of end users of these models, and also the explanation methods may bring in the validation techniques (Molnar 2020). In this study, a Local Interpretable Model-agnostic Explanations (LIME) model explainability technique is applied. The aim of this method is to reduce the difference between artificial intelligence and humans by providing an explanation of how each feature is influencing a classifier. The method's main focus is to create a local approximation of our complex model (the trained model) for a particular discussed by Ribeiro et al. (2016).

Given a set of predictors $x \in X$, our trained model is denoted as f, and $g \in G$ where g is a simple model that comes from G, a family of interpretable models such as linear regression. An explainer $\xi(x)$ is defined as shown below:

$$\xi(x) = argmin_{g \in G} \mathscr{L}(f, g, \pi_x) + \Omega(g), \tag{19}$$

where π_x denotes the local neighbours of an instance x and $\Omega(g)$ represents a measure of the complexity that $g \in G$ does not explain. Equation 19 is an optimization task whose aim is to get a good approximation $\mathscr{L}$ and a minimum complexity $\Omega(g)$. Basically, the main idea is to look for a simple model g that approximates the trained model f in the local function $\mathscr{L}$ and keep its simple nature. To accomplish this, a new dataset is generated randomly with similar features as the original set. Predictions are made using the new data and the complex model f. The first loss term $\mathscr{L}$ is minimized by getting the highest accuracy on the new dataset using a simple linear model. Ribeiro et al. (2016), get the loss term by getting the difference in the sum of squared distances between the labels predicted by model f and the predictions of model g. π_x is also included to weigh the loss according to how close the data point is.

$$\mathscr{L}(f, g, \pi_x) = \sum_{y,y'} \pi_x(y)(f(y) - g(y'))^2. \tag{20}$$

The loss term $\Omega(g)$, ensures that the model keeps its simplicity nature. Ribeiro et al. (2016) state that in LIME, a sparse linear model is used to maintain simplicity. This model takes care of the second loss function since the model aims at producing as many zero weights as it can. This can also be achieved by using regularization techniques that help get a simple model with relevant features (Molnar 2020).

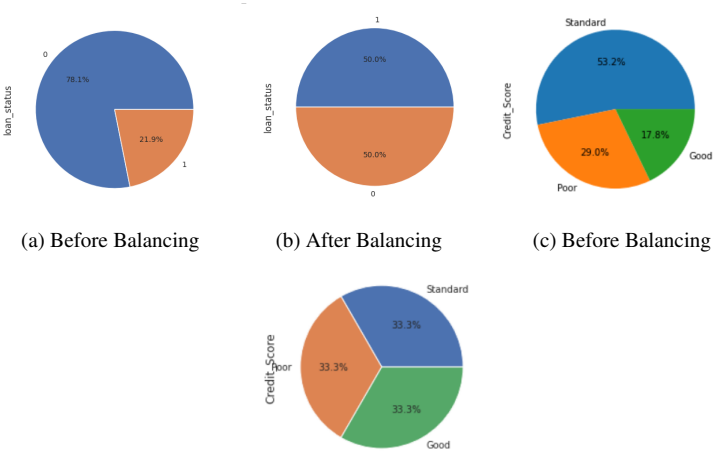
3. Data Analyses and Results

This section aims to provide a description of our datasets, an evaluation of the model's performance, and an explanation of the model's prediction. We used Python 3 software to

implement the various methods applied. Python libraries used include Pandas, Matplotlib, Seaborn, Scikit-learn as indicated by Pedregosa et al. (2011) and TensorFlow.

3.1. Data Description and Preprocessing

We used two types of datasets, a binary and multiclass case dependent variable. Both datasets contained categorical and numerical variables. The binary case dataset contained 32580 observations with a total of 12 columns while the multiclass case had 100,000 observations and 28 columns. In the binary case data, the dependent variable, 'loan status', takes a value of 0 or 1, where 0 indicates that the customer is a non-defaulter, while 1 is a defaulter. Conversely, in the multiclass case, the dependent variable, 'credit score', consists of three categories: good, standard, and poor credit scores. The simulated dataset was obtained using the make classification inbuilt function in scikit-learn library in Python. The binary case dataset contained 4011 missing values, but no missing values in the multiclass case. Various methods were employed, such as imputation of missing values, feature selection, standardization, data splitting, and balancing of classes in the training set to ensure that the data were consistent, complete, and appropriate. Good quality data and relevant features improve the accuracy and generalization of the model (Laborda and Ryoo 2021).



(a) Before Balancing (b) After Balancing (c) Before Balancing

(d) After Balancing

Figure 2: Distribution of Classes.

Figure 2a shows the distribution of customers' loan status, with approximately 78% non-defaulter customers and 22% defaulters. After preprocessing, the dataset was reduced to 32414 rows and 8 features for the binary case and 100000 rows and 16 variables for the multiclass case, using recursive feature elimination with cross-validation method. Figure 2c shows the proportion of customers' credit scores, with approximately 53% of the customers having a standard credit score, 29% having a poor one and 18% having a good score.

3.2. Model Assessment and Discussion

The performance evaluation metrics such as accuracy, balanced accuracy, Matthews Correlation Coefficient (MCC), precision and AUC-ROC score were used to examine the performance of the models used in the study. To assess the model's performance, we divided our data where 80% of the data trained the models and 20% for evaluation. The parameters used to train the MLP are shown in Table 2.

Table 2: Multilayer Perceptron Hyperparameters.

Parameters	**Levels in Binary Classification**	**Levels in Multiclass Classification**
Hidden Layers	2(8, 6 nodes in each layer)	2(10, 8 nodes in each layer)
Output Layer	2 nodes	3 nodes
Activation Functions	relu, relu, sigmoid	relu, relu, softmax
Dropout	0.1	0.1
Loss Function	binary cross entropy	categorical cross entropy
Optimizer	Adam	Adam
Iterations	50	50

3.2.1 Binary Case

Table 3 gives a view of how the models performed based on different evaluation metrics.

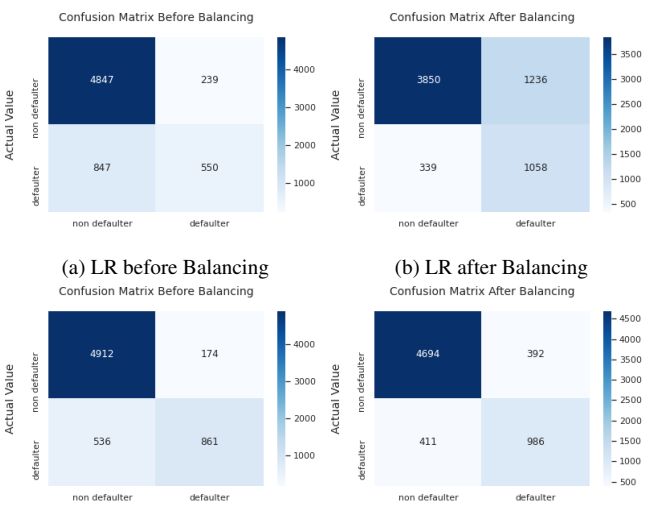
Table 3: Performance of LR and MLP for Binary Classification.

Evaluation Metrics	**Empirical Data**				**Simulated Data**			
	Imbalanced Data		Balanced Data		Imbalanced Data		Balanced Data	
	LR	MLP	LR	MLP	LR	MLP	LR	MLP
Accuracy	0.832	0.888	0.755	0.878	0.910	0.939	0.870	0.928
Balanced Accuracy	0.673	0.759	0.754	0.812	0.820	0.871	0.860	0.902
Precision	0.820	0.861	0.820	0.880	0.890	0.900	0.890	0.940
F1 Score	0.810	0.870	0.770	0.880	0.900	0.940	0.880	0.940
MCC	0.436	0.624	0.437	0.648	0.701	0.813	0.657	0.856
AUC-ROC score	0.827	0.759	0.828	0.812	0.924	0.871	0.925	0.902

From Table 3, we observe that the accuracy of all created models decreases after balancing the data. Given the sensitivity of accuracy on imbalanced data, we consider balanced accuracy metric. MLP exhibits a higher performance for both imbalanced and balanced data than logistic regression for both accuracy and balanced accuracy metrics. This highlights MLP's ability to predict the credit payment ability of customers accurately. The models demonstrate higher MCC scores after data balancing. MLP has higher MCC scores compared to LR. This shows MLP's ability to predict both non-default and default customers more accurately than LR.

The precision scores of all the models are found to be high on both balanced and imbalanced data. Upon comparing the performance of LR and MLP, we observe that MLP exhibited higher precision scores than LR. This indicates that MLP is more capable of accurately classifying true positive instances out of the total positives compared to LR, which means it is better at capturing fewer false non-default customers. The F1 scores of LR and MLP are reasonably high, indicating that both models are able to capture the underlying patterns in the dataset. Based on F1 score MLP performs best. Both models demonstrated a good performance based on the AUC-ROC score, as all models achieved scores higher than 0.5. This suggests that the models were able to effectively distinguish between non-default and default loan applicants. Notably, LR outperforms MLP by achieving a higher AUC-ROC score. This means that LR has a higher ability to distinguish loan applicants than MLP. We observe that the results of all the metrics appear similar in both the empirical and simulated data.

The confusion matrices in Figure 3 provide a summary of the number of correct and incorrect predictions made by the two models we have created on imbalanced and balanced data in our binary classification. The non-default class, which is the majority class, shows a higher count of instances that are correctly predicted on imbalanced data. However, after balancing the data, a more balanced distribution is achieved across all classes. This indicates that for imbalanced data, the models are more likely to predict the non-default class, but after balancing the data, the models become better at predicting the minority class as well. In addition, MLP demonstrates a greater ability to accurately predict the true classes and a lower number of misclassifications shown in the off-diagonal elements of the confusion matrices in comparison to LR.



(a) LR before Balancing (b) LR after Balancing

(c) MLP before Balancing (d) MLP after Balancing

Figure 3: Binary Confusion Matrix.

3.2.2 Multiclass Case

This section focuses on the model performance of our second dataset which has a dependent variable with three classes. The models created are meant to predict whether a customer has a good, standard or poor credit score. A good credit score means that a customer is likely to pay a given loan on time; a standard credit score, the customer pays the loan on time on average; and a poor credit score means that the customer is likely to default on a loan. A visual of the model's performance is shown in Table 4 for both imbalanced and balanced data. The simulated data showed similar performance scores for both imbalanced and balanced data.

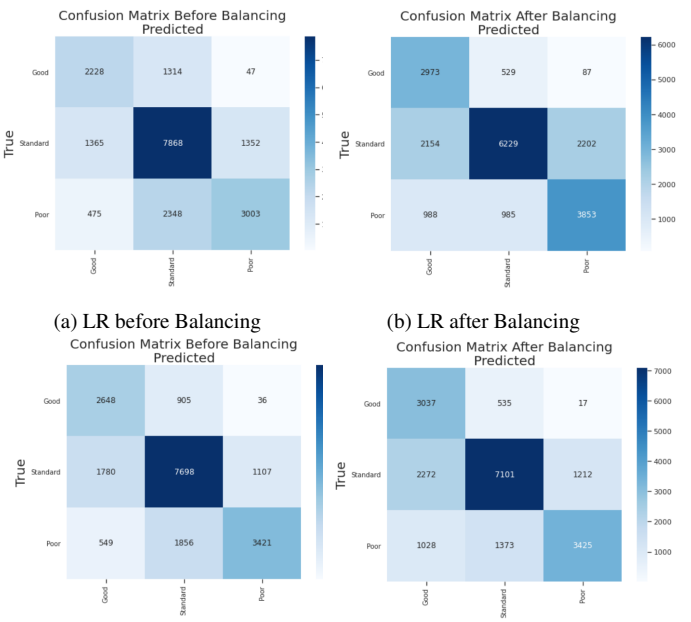
Table 4: Performance Metrics of LR and MLP for Multiclass Classification.

Evaluation Metrics	**Empirical Data**				**Simulated Data**	
	Imbalanced Data		Balanced Data			
	LR	MLP	LR	MLP	LR	MLP
Accuracy	0.645	0.680	0.65	0.678	0.629	0.687
Balanced Accuracy	0.600	0.709	0.685	0.718	0.525	0.621
Precision	0.660	0.710	0.690	0.720	0.620	0.700
F1 Score	0.650	0.670	0.650	0.680	0.610	0.680
MCC	0.422	0.485	0.469	0.515	0.345	0.473
AUC-ROC score	0.806	0.855	0.803	0.849	0.770	0.840

An analysis of the performance of our models with respect to the scores presented in Table 4 revealed that MLP exhibited a higher accuracy and balanced accuracy on both balanced and imbalanced data compared to LR. This is also observed in the simulated data. This means that MLP model was better at predicting the probability of a good, standard or poor credit score than LR. MLP has a higher precision score, 71% for imbalanced and 72% for balanced data compared to LR. MLP model also showed higher MCC scores of 0.485 and 0.515 for imbalanced and balanced data, respectively, in comparison to logistic regression, which showed MCC scores of 0.422 and 0.469 for imbalanced and balanced data, respectively. This indicates that MLP has a higher capacity to accurately predict the three classes of customers compared to LR. This means that MLP can capture fewer false positives. The MCC values can vary depending on factors such as class distribution, the separation of classes, interclass correlations, and the complexity of the model. It is important to consider these factors and evaluate MCC values on a case-by-case basis, as the interpretation of MCC values depends on the specific problem, dataset, and model performance in both binary and multiclass classification scenarios (Chicco et al. 2020). All the models perform well on AUC-ROC scores because they have scores greater than 0.5. This indicates that both models can distinguish between loan applicants with good, standard and poor credit scores. Considerably, MLP has higher AUC-ROC scores than LR. This shows that MLP has a higher ability to distinguish loan applicants than LR. Moreover, on the simulated data, MLP performs better than LR for all the performance metrics.

Figure 4 presents the confusion matrices of the multiclass classification models. Before balancing, as shown in Figures 4a and 4c, the standard credit score class had the highest

count owing to its majority class status, whereas the good credit score class had the lowest count as a result of being the minority class. Following the balancing of our data, in Figures 4b and 4d, the count of majority classes reduced therefore achieving a fair distribution across all classes. Our analysis of the confusion matrices presented in Figures 4b and 4d shows that the MLP model outperformed logistic regression in terms of correctly predicting class elements for the balanced data. This observation is consistent with the higher balanced accuracy of the MLP model, as reported in Table 4.



(a) LR before Balancing (b) LR after Balancing

(c) MLP before Balancing (d) MLP after Balancing

Figure 4: Multiclass Confusion Matrix.

4. Explainability with LIME

In order to obtain explanations of a model's prediction we use the LIME package in Python. We compile a list of the attributes used to train the model. We then define class labels (i.e. non-defaulter and defaulter for binary classification, and good, standard and poor for the multiclass classification), and then we create a function that provides the probabilities of each feature, fed it as an array. Passing all these components to the LIME explainer object, we input an observation into the explainer, which yields a prediction and offers insights into how each feature contributes to the classes present. Figures 5, 6, 7, and 8 display a visual representation of the prediction explanations for two instances provided by LIME. The results of a binary class prediction instance for imbalanced and balanced datasets are illustrated in Figures 5 and 6, respectively.

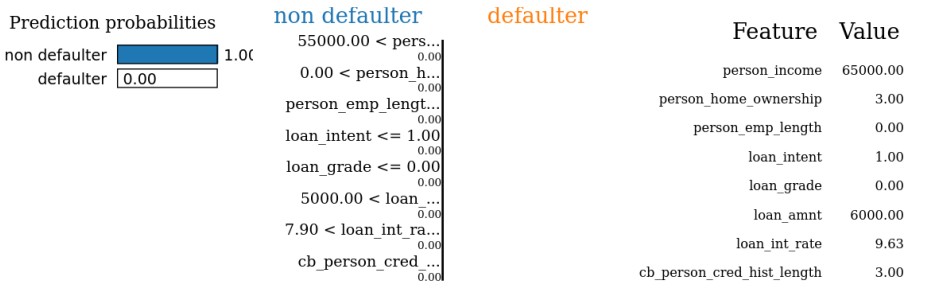


Figure 5: LIME explanations for a given observation on Binary Imbalanced data.

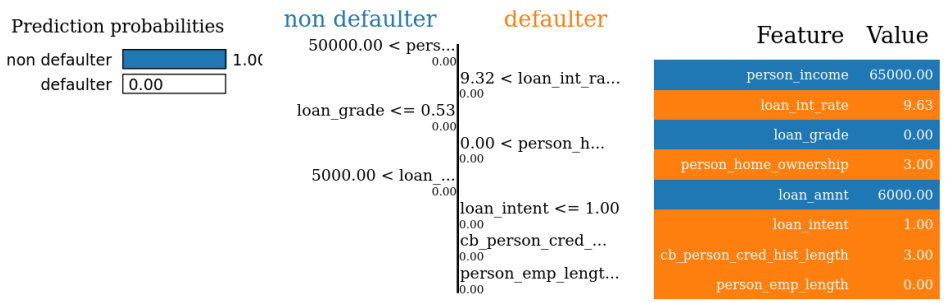


Figure 6: LIME explanations for a given observation on Binary Balanced data.

Figure 5 displays the prediction of a non-default customer, where all features were relevant in contributing to the non-default class prediction. In the scenario of balanced data in Figure 6, we observe a similar prediction of a non-default class on the same instance. However, in this case only three features significantly contribute to the non-default class: the applicants' income, loan grade, and loan amount. In both cases, the models give 100% assurance that the given loan applicant is a non-defaulter meaning that the given customer was likely to pay the loan on time. Figures 7 and 8 present the LIME explanation for a given instance on both imbalanced and balanced multiclass data, respectively.

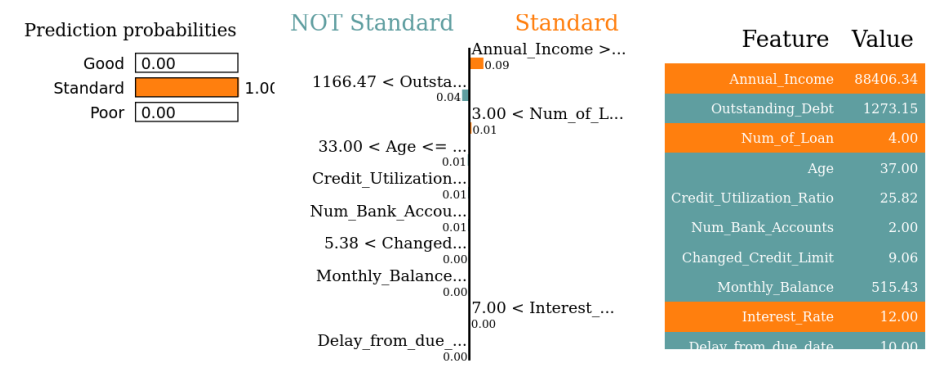


Figure 7: LIME generated explanations for a given observation on Multiclass Imbalanced data.

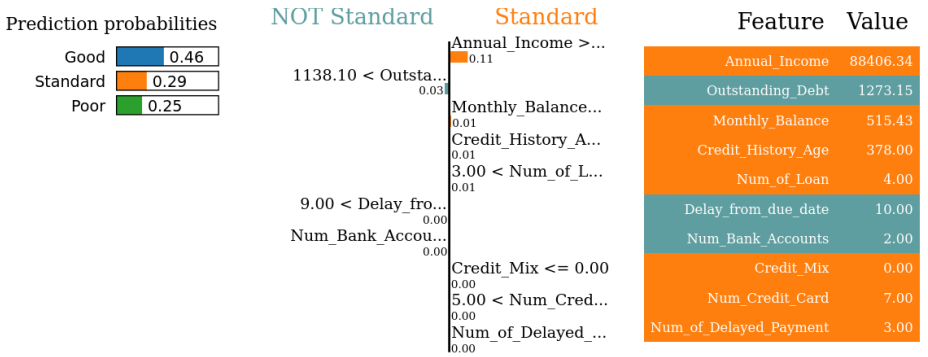


Figure 8: LIME generated explanations for a given observation on Multiclass Balanced data.

In Figure 7, we observe that the model predicts a standard credit score for the given loan applicant, where most features contribute to the decision. Additionally, the model shows 100% confidence in its prediction, suggesting a high likelihood that the customer will not pay the loan on time entirely. However, the prediction for the same loan applicant differs when it comes to balanced data, as shown in Figure 8. In this scenario, the model predicted with 46% assurance that the given applicant had a good credit score, indicating a likelihood of timely loan repayment. This information proves to be valuable to financial regulators in their decision-making process regarding loan offers.

5. Conclusion

In this study, we have compared the predictive ability of Logistic Regression (LR) and a Multilayer Perceptron (MLP) using two types of datasets, with an advanced model explainability technique - Local Interpretable Model-Agnostic Explanations (LIME). The findings show that all models performed better after the data were balanced. MLP had higher scores than LR in terms of balanced accuracy, Matthews correlation coefficient, and F1 score. From our findings, this study recommends that lending companies with small amounts of data use a logistic regression model but for companies with vast amounts of data a multilayer perceptron will ease their credit offer processes. The study also highlights the importance of using explainable artificial intelligence. With the LIME explanation approach, we were able to see how each feature influences the predicted class of a model for a given instance. We also found out that, models developed on imbalanced data are likely to show biased results, which may cost the lenders in the future. From what we were able to interpret, the LIME framework will enable developers to clarify to end-users the rationale behind a specific decision.

Disclosure of Competing Interests

The authors declare that they have no competing interests.

References

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R. and Hinton, G. E., (2021). Statistical regression modeling with R. *Advances in Neural Information Processing Systems*, 34, No. 1, pp. 4699–4711.

Bolton, C., (2009). Logistic regression and its application in credit scoring.*University of Pretoria (South Africa)*.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P., (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, No. 1, pp. 321–357.

Chicco, D., Jurman, G., (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21, pp. 1–13.

Dumitrescu, E., Hué, S., Hurlin, C. and Tokpavi, S., (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297, pp. 1178–1192.

Hand, D. J., Henley, W. E., (1997). Statistical classification methods in consumer credit scoring: a review.*Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 160(3), pp. 523–541.

Hossain, M. M., (2022). Statistical regression modeling with R. *Oxford University Press*, Vol. 5, No. 1, pp. 63–72.

Imtiaz, S., Brimicombe, A. J., (2017). A Better Comparison Summary of Credit Scoring Classification. *International Journal of Advanced Computer Science and Applications*, 8(7).

Laborda, J., Ryoo, S., (2021). Feature selection in a credit scoring model. Mathematics. *Mathematics*, 16, 9(7), p. 746.

Lahsasna, A., Ainon, R. N. and Teh, Y. W., (2010). Credit Scoring Models Using Soft Computing Methods: A Survey. *Int. Arab J. Inf. Technol.*, 7(2), pp. 115–123.

Little, R. J. Rubin, D. B., (2019). Statistical analysis with missing data, Vol. 793. John Wiley & Sons.

Mahmood, E. A., (2024). Robust Estimation of Multiple Logistic Regression Model.

Misra, P., Yadav, A. S., (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. *Int. J. Emerg. Technol*, 11(3), pp. 659–665.

Molnar, C., (2020). Interpretable machine learning. *Lulu. com*.

Mustaqim, A. Z., Adi, S., Pristyanto, Y. and Astuti, Y., (2021). The effect of recursive feature elimination with cross-validation (RFECV) feature selection algorithm toward classifier performance on credit card fraud detection. *In 2021 International conference on artificial intelligence and computer science technology (ICAICST)* , pp. 270–275, IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., (2011). Scikit-learn: Machine Learning in Python. *the Journal of machine Learning research*, 12, pp. 2825–2830.

Ribeiro, M. T., Singh, S. and Guestrin, C., (2016). Why should i trust you?" Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledgediscovery and data mining*, pp. 1135–1144.

Rodrigues, P. C., Awe, O. O., Pimentel, J. S. and Mahmoudvand, R., (2020). Modelling the behaviour of currency exchange rates with singular spectrum analysis and artificial neural networks. *Stats*, 3(2), pp. 137–157.

Scheuren, F., (2005). Multiple imputation: How it began and continues. *The American Statistician*, Vol. 59(4), pp. 315–319.

Sewpaul, R., Awe, O. O., Dogbey, D. M., Sekgala, M. D. and Dukhi, N., (2023). Classification of Obesity among South African Female Adolescents: Comparative Analysis of Logistic Regression and Random Forest Algorithms. *International Journal of Environmental Research and Public Health*, 21(1), No. 1, p. 2.

Taghavi Takyar, S. M., Aghajan Nashtaei, R. and Chirani, E., (2015). The Comparison of Credit Risk between Artificial Neural Network and Logistic Regression Models in Tose-Taavon Bank in Guilan. *International Journal of Applied Operational Research-An Open Access Journal*, 5(1), pp. 63–72.

Tran, K. L., Le, H. A., Nguyen, T. H. and Nguyen, D. T., (2022). Explainable machine learning for financial distress prediction: Evidence from Vietnam. *Data*, 7(11), p. 160.

Van Buuren, S., Oudshoorn, C. G. M., (2000). Multivariate imputation by chained equations: Mice v1. 0 user's manual.

Wulff, J. N., Jeppesen, L. E., (2017). Multiple imputation by chained equations in praxis: guidelines and review *Electronic Journal of Business Research Methods*, 15, pp. 41–56.

Zhang, A., Lipton, Z.C., Li, M. and Smola, A. J., (2023). Dive into deep learning *Cambridge University Press*.

Zhao, Z., Xu, S., Kang, B. H., Kabir, M. M. J., Liu, Y. and Wasinger, R., (2015). Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications*, 42(7), pp. 3508–3516.

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 69–81, https://doi.org/10.59170/stattrans-2024–028
Received – 12.01.2023; accepted – 18.03.2024

The Measurement of the Gross Domestic Product affected by the shadow economy

Anna Czapkiewicz[1], **Katarzyna Brzozowska-Rup**[2]

Abstract

The article presents a method for balancing Gross Domestic Product (GDP) when the measurements of its components are distorted by the existence of the shadow economy. Our proposal to measure GDP is based on a multiple ultrastructural model (MUM), where the explanatory variables are subject to error. We show that the expected value of GDP can be divided into two parts: the first part concerns data related to registered activities and the second part concerns unobserved data which may be partly related to the shadow economy. The empirical analysis is based on the annual data for individual voivodeships in Poland for the years 2000–2019. The data are obtained from the Local Data Bank of Statistics Poland. Two approaches to measuring GDP are considered: from the expenditure side and from the production side. The results show that the unobservable part of the variables necessary to balance GDP on the production side does not exceed 1% of GDP, and on the expenditure side, it mostly reaches about 3% of GDP.

Key words: measurement error, ultrastructural model, Gross Domestic Product, shadow economy.

1. Introduction

According to Regulation (EU) No 549/2013 of the European Parliament and of the Council of 21 May 2013 on the European system of national and regional accounts in the European Union, one of the main aggregates in the ESA is gross domestic product (GDP). It is a measure of the total economic activity taking place on a given economic territory. In 1995, ESA documents introduced the concept of including shadow economy data in GDP. According to the definition formulated by the European Commission, the shadow economy is an economic sector comprising a group of economic activities that are productive, in line with the SNA/ESA definition of production, legal in terms of compliance with legal norms and regulations, but hidden from public authorities. Methods for measuring the shadow economy can be divided into two groups: direct methods and indirect methods. The first group includes all survey research, while the second is limited to the analysis of "traces" in macroeconomic data.

In Poland, the size of the shadow economy is determined by national statistical offices on the basis of survey data. However, due to the problems of survey research, other methods

[1]Department of Applications of Mathematics in Economics, Faculty of Management, AGH University of Science and Technology, Cracow, Poland & Centre for Non-observed Economy Studies, Statistical Office in Kielce, Kielce, Poland. E-mail: czapkiewicz@agh.edu.pl. ORCID: https://orcid.org/0000-0002-6144-8381.

[2]Faculty of Management and Computer Modelling, Kielce University of Technology, Kielce, Poland. & Centre for Non-observed Economy Studies, Statistical Office in Kielce, Kielce, Poland. E-mail: krup@tu.kielce.pl. ORCID: https://orcid.org/0000-0003-1231-8027.

of estimating the shadow economy are discussed in the literature. Methods for estimating the shadow economy are described by Adair (2017, 2021), Błasiak (2018), Medina and Schneider (2019), Malczewska (2019), among others.

In the literature, special attention is paid to econometric models used to describe the relationships between some macroeconomic data. Such methods can be useful for estimating the shadow economy. One of the most used econometric models for this purpose is the Multiple Indicators Multiple Causes (MIMIC). The properties of the MIMIC model in the context of shadow economy estimation are discussed, for example, in Schneider et al. (2000, 2005a, 2005b, 2018), Trebicka (2014), Breusch (2005), Buszko (2017), Dybka et al. (2017, 2019). However, MIMIC is a confirmatory rather than an exploratory statistical technique. As Kirchgassner (2016) points out, the conclusion that the variable is considered a statistically significant determinant of the shadow economy may not be fully justified. The model, like many other hidden variable models used to measure an unobservable phenomenon, is based on the assumption that some relationship exists (Dybka et al., 2019).

This article presents a model that includes unobservable variables that is similar to the MIMIC model but has a slightly different structure. The proposed model determines a linear relationship between random variables where the dependent variable and some of the explanatory variables are subject to measurement error. The problems of linear models with errors in the variables have been discussed by Kendall and Stuart (1973), Dolby (1976), Chan and Mak (1984), Gillard (2006), and others. Dolby (1976) considers an ultrastructural relationship with only a single explanatory variable. The model discussed in this paper extends the reasoning of Dolby (1976) for a multivariate case, hereafter called the Multiple Ultrastructural Model (MUM). A maximum likelihood estimation algorithm is also presented, which allows us to effectively estimate the unknown model parameters and data relationships. The structure of the MUM allows to determine the relationships between GDP and its components, among other things. Gross domestic product is measured using three different approaches - output (production), income and expenditure - which are then compiled (properly balanced) to give a final estimate of GDP. Possible differences between these methods may result from the existence of shadow economy. Therefore, the possibility of defining SE as the difference between GDP in terms of production and expenditure has been considered, for example, by Madzarevic-Sujster (2001), Mikulic (2002), Czapkiewicz and Brzozowska-Rup (2021).

The empirical study of this paper examines the relationships between published data on GDP and its components. The data are taken from the Local Data Bank of Statistics Poland and cover the years 2000-2019. These data already include the informal economy. Therefore, the MUM model used to describe these relationships checks the correctness of the balance between different approaches to calculating GDP, taking into account SE.

The purpose of this article is twofold. First, to test whether the MUM model provides a good forecast of GDP, and second, to examine the balance of GDP resulting from the production and expenditure approach. In addition, this methodology allows us to estimate the percentage of GDP accounted for by unobservable data that may not have been included in the GDP calculations.

The paper is organized as follows. Section two defines the model and discusses its methodological features. Section three presents the data and results of the empirical study. A brief summary and conclusions are presented in the last section.

2. Model

Let $(X_t^1,\ldots,X_t^W,Y_t)$ $(t=1,\ldots,T)$ refer to observations that are subject to measurement error. True but unknown values of the explanatory variables are denoted as s_t^w and

$$X_t^w = s_t^w + \varepsilon_t^w \quad w=1,\ldots,W.$$

Let assume that

$$Y_t = \gamma_1 s_t^1 + \ldots + \gamma_W s_t^W + \eta_t$$

where ε_t^w and η_t are normally distributed measurement errors. However, under these assumptions the model is unidentifiable. Therefore, the dependent and explanatory variables are replicated N times to avoid the unidentifiability problem. They are further represented by the vector $(X_{i,t}^1,\ldots,X_{i,t}^W,Y_{i,t})$, where $i=1,\ldots,N$. The true values of the explanatory variables $(s_{i,t}^1,\ldots,s_{i,t}^W)$ satisfy a linear relation

$$Y_{i,t} = \gamma_1 s_{i,t}^1 + \ldots + \gamma_W s_{i,t}^W + \eta_{i,t}. \tag{1}$$

Furthermore, we assume that

$$s_{i,t}^w \sim N(s_i^w,\sigma_{s,w}), \quad X_{i,t}^w \sim N(s_i^w,\sigma_{\varepsilon,w}),$$

$$\varepsilon_{i,t}^w \sim N(0,\sigma_{\varepsilon,w}), \quad \eta_{i,t} \sim N(0,\sigma_\eta).$$

Assuming that $s_{i,t}^w,\varepsilon_{i,t},\eta_{i,t}$ are independent, the vector $(X_{i,t}^1,\ldots,X_{i,t}^W,Y_{i,t})$ is normally distributed with a mean of $(s_t^1,\ldots,s_t^W,\sum_{w=1}^W \gamma_w s_t^w)$ and a covariance matrix

$$\Sigma = \begin{bmatrix} V_{11} & V_{12} \\ V_{21}^T & V_{22} \end{bmatrix}$$

where $V_{11} = [V_1,\ldots,V_W] \times I_W$ and $V_w = \sigma_{s,w}^2 + \sigma_{\varepsilon,w}^2$,
$V_{22} = \sum_{w=1}^W \gamma_w^2 \sigma_{s,w}^2 + \sigma_\eta^2$, $V_{12} = [U_1,\ldots,U_W]^T$ and $U_w = \gamma_w \sigma_{s,w}^2$.

The unknown parameters of the model are estimated by the maximum likelihood method. To avoid technical problems in determining the estimators of the unknown parameters, $\sigma_{s,w}^2 = 0$ is assumed, then the vector of unknown parameters is

$$\Phi = (\gamma_1,\ldots,\gamma_W,\sigma_\eta,\sigma_{\varepsilon,1},\ldots,\sigma_{\varepsilon,W},s_1,\ldots,s_W).$$

Consider the vector

$$Z_i = (X_{i,1}^1,\ldots,X_{i,T}^1,\ldots,X_{i,T}^W,\ldots,X_{i,T}^W,Y_{i,1},\ldots,Y_{i,T}).$$

Its expectations (μ) and covariance matrix (V) take the form of

$$\mu = (s_1^1, \ldots, s_T^1, \ldots, s_1^W, \ldots, s_T^W, \gamma_1 s_1^1 + \ldots + \gamma_W s_1^W, \ldots, \gamma_1 s_T^1 + \ldots + \gamma_W s_T^W)$$

and

$$V = \Sigma \otimes I,$$

where $\otimes$ is the Kronecker matrix multiplication. The log-likelihood function corresponding to N replications has a form

$$L = C - \frac{N}{2} \ln |V| - \frac{1}{2} \sum_{i=1}^{N} d_i V^{-1} d_i \text{ where } d_i = Z_i - \mu. \tag{2}$$

Let V_ϕ be the matrix of derivatives computed with respect to ϕ ($\phi \in \Phi$), we get

$$\frac{\partial \ln L(\Phi)}{\partial \phi} = N \left(\frac{1}{2} tr(P V_\phi) - d_\phi^i V^{-1} d \right) \tag{3}$$

where $d = \sum_{i=1}^{N} d_i$, $P = V^{-1}(D - V)V^{-1}$ and $D = \frac{1}{N} \sum_{i=1}^{N} d_i d_i^T$.

By performing the appropriate calculations, we obtain estimators for the unknown parameters of the model. Let $\left(X_{.,t}^w\right)$ and $(Y_{.,t})$ denote averages of N replications at each time t and let

$$R_t^w = \left(X_{.,t}^w - s_t^w\right) = \frac{-z_t \gamma_w \sigma_{\varepsilon,w}^2}{\sigma_\eta^2 + \gamma_1^2 \sigma_{\varepsilon,1}^2 + \ldots + \gamma_W^2 \sigma_{\varepsilon,W}^2}, \quad w = 1, \ldots, W, \tag{4}$$

where

$$z_t = Y_{.,t} - \gamma_1 X_{.,t}^1 - \ldots - \gamma_W X_{.,t}^W. \tag{5}$$

The unknown parameters $\gamma_1, \ldots, \gamma_W$, σ_η^2 and $\sigma_{\varepsilon,w}^2$ satisfy the following non-linear equations

$$\sum_{t=1}^{T} (X_{.,t}^w - R_t^w)(\gamma_1 R_t^1 + \ldots + \gamma_W R_t^W + z_t) = 0$$

$$\sigma_\eta^2 = \frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{i,t} - Y_{.,t})^2 + \frac{1}{T} \sum_{t=1}^{T} (\gamma_1 R_t^1 + \ldots + \gamma_W R_t^W + z_t)^2$$

$$\sigma_{\varepsilon,w}^2 = \frac{1}{TN} \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{i,t}^w - X_{.,t}^w)^2 + \frac{1}{T} \sum_{i=1}^{N} (R_t^w)^2.$$

The next step is to estimate the unobservable parts of the explanatory variables. We found that the expected value $E(Y_{i,t})$ can be represented by two components $E(X_{i,t})$ and SE

$$E(Y_{i,t}) = \gamma_1 s_t^1 + \ldots + \gamma_W s_t^W = \gamma_1 (X_{.,t}^1 - R_t^1) + \ldots + \gamma_W (X_{.,t}^W - R_t^W) =$$

$$\gamma_1 X_{.,t}^1 + \ldots + \gamma_W X_{.,t}^W - (\gamma_1 R_t^1 + \ldots + \gamma_W R_t^W) = E(X_{i,t}) - SE.$$

The first component refers to the observed data. The second component refers to the unobservable part of the explanatory variables, which can be estimated using the formula

$$| E(Y_{i,t}) - E(X_{i,t}) | = | \hat{\gamma}_1 \hat{R}_t^1 + \ldots + \hat{\gamma}_W \hat{R}_t^W | = | SE | . \quad (6)$$

3. Empirical study

Gross Domestic Product is the primary indicator of economic activity, which can be estimated in three independent and theoretically equivalent ways. The results of these methods may differ, so a balancing process is carried out to obtain the final GDP (SNA93, ESA95). The article discusses two approaches: from the production side and from the expenditure side.

In the production approach, GDP is the gross value added of institutional sectors or industries (GP) minus the value of intermediate consumption (IC). In the expenditure approach, GDP is equal to the sum of the final uses of goods and services (all uses except intermediate consumption) plus exports and minus imports of goods and services. It corresponds to the expenditure of all purchasers of final goods produced during the year, including consumption (Cn), investment (In), government expenditure (Gov) and net exports (Exports (Ex)-Imports (Im)). In empirical study the MUM model is used to check the correctness of the balance between two approaches to calculating GDP when its components include unknown values related to the shadow economy.

3.1. Data

The analysis considers voivodeship data (there are 16 voivodeships in Poland) from 2000-2019. The voivodeship observations for each year are assumed to be replications in the MUM model. These data are transformed to ensure the desired statistical properties. Taking into account the differences in the logarithms of the observations, it is possible to compare values while avoiding large differences between voivodships.

The dependent variable is defined as $Y_{i,t} = \log\left(\frac{GDP_{i,t}}{GDP_{i,t-1}}\right)$. In the first model (the production approach) the explanatory variables are $P_{i,t} = \log\left(\frac{GP_{i,t}}{GP_{i,t-1}}\right)$ and $Z_{i,t} = \log\left(\frac{IC_{i,t}}{IC_{i,t-1}}\right)$, whereas in the second model (the expenditure approach) the explanatory variables are: $C_{i,t} = \log\left(\frac{Cn_t}{Cn_{t-1}}\right)$, $I_{i,t} = \log\left(\frac{In_t}{In_{t-1}}\right)$, $G_{i,t} = \log\left(\frac{Gov_{i,t}}{Gov_{i,t-1}}\right)$, $E_{i,t} = \log\left(\frac{Ex_{i,t}}{Ex_{i,t-1}}\right) - \log\left(\frac{Im_{i,t}}{Im_{i,t-1}}\right)$ and $i = 1\ldots,16$. The consumption and investment are noted by voivodeship, whereas government expenditures and net export are aggregated data for Poland. Hence, the model includes also data related to voivodeship budget expenditures, $B_{i,t}$.

Figure 1 shows the data considered in the empirical study.

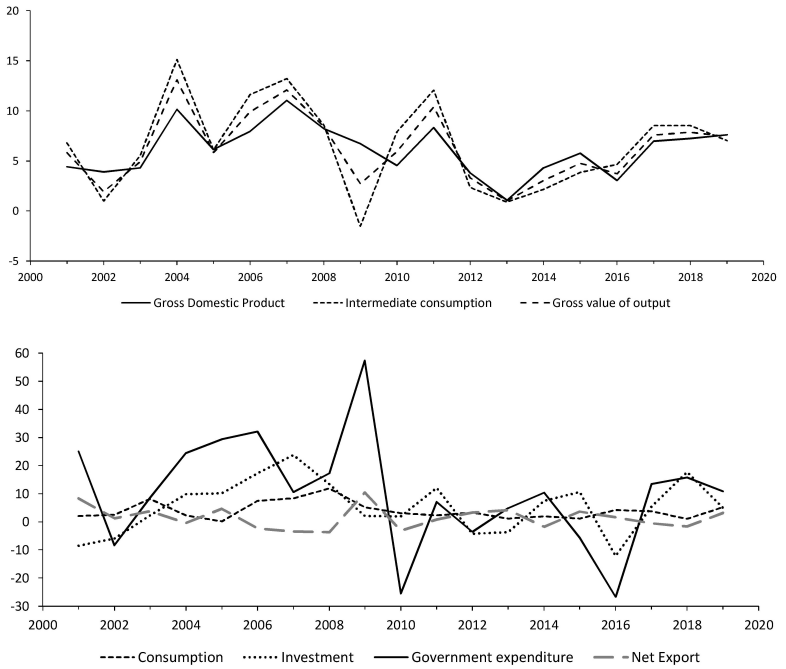


Figure 1: The average values (calculated by voivodeship) of transformed data related to GDP, intermediate consumption and gross value of output ($Y_{i,t}$, $P_{i,t}$, $Z_{i,t}$) (first panel); and the average values of transformed data related to consumption, investment, governed expenditure and net export ($C_{i,t}$, $I_{i,t}$, $B_{i,t}$, $E_{i,t}$) (second panel).

In order to illustrate the lack of meaningful differences in the replication by voivodeships taken into account in the MUM model, Figure 2 is presented. It shows four voivodeships with the lowest GDP growth (first panel) and four voivodeships with the highest GDP growth (second panel). The highest growth in the period under review was observed in the Małopolskie voivodeship and the lowest in the Zachodniopomorskie voivodeship. It should be noted, however, that the increases in GDP within each voivodeship are relatively small, so the data from the voivodeships can be used as replications in the MUM model.

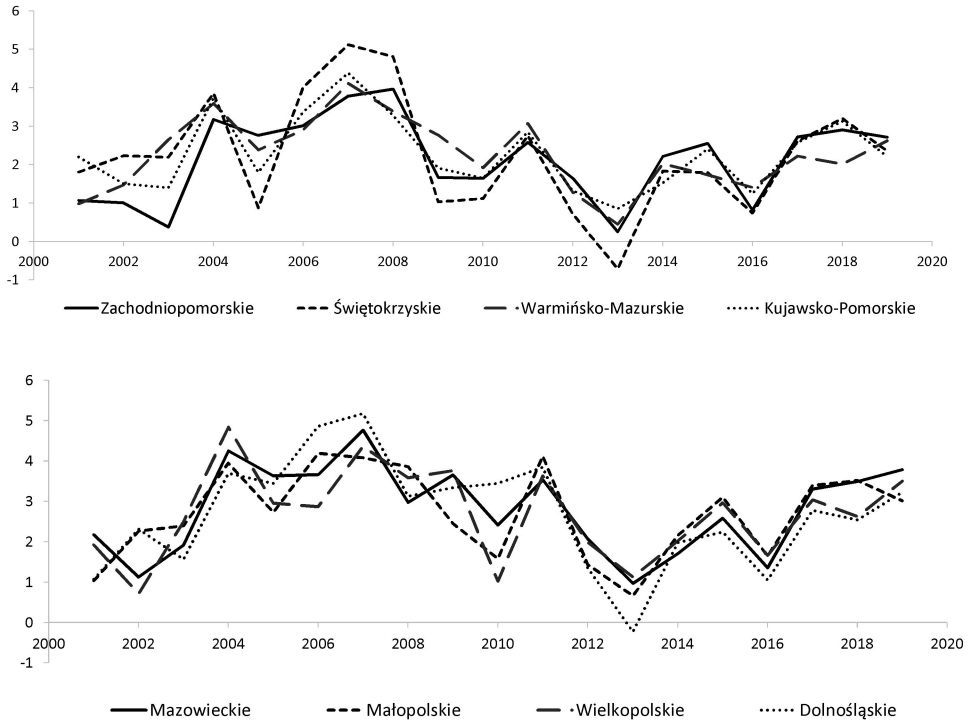


Figure 2: GDP growth in percent for voivodeships: voivodeships with the lowest GDP (first panel), voivodeships with the highest GDP (second panel).

3.2. Results

3.2.1 GDP - production approach

In this subsection GDP is calculated using global production and intermediate consumption. In this case, the MUM model has a form

$$P_{i,t} = s_t^P + \varepsilon_{i,t}^P$$

$$Z_{i,t} = s_t^Z + \varepsilon_{i,t}^Z$$

$$Y_{i,t} = \gamma_1 s_t^P + \gamma_2 s_t^Z + \eta_{i,t} \tag{7}$$

where

$$\varepsilon_{i,t}^P \sim N(0, \sigma_{\varepsilon,p}),\ \varepsilon_{i,t}^Z \sim N(0, \sigma_{\varepsilon,z}),\ \ \eta_{i,t} \sim N(0, \sigma_\eta),\quad i = 1, \ldots, N,\ \ t = 1, \ldots, T.$$

and s_t^P, s_t^Z denote true but unknown (unobserved) values of explanatory variables.

Table 1 shows the ML estimates of the parameters γ_1 and $\gamma 2$. For comparison, these parameters are also estimated from the Ordinary Least Square (OLS) regression model. To evaluate the goodness of fit of two models, the mean squared error (MSE) is calculated:

$$MSE = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(Y_t - \hat{Y}_t)^2},$$

where Y_t denotes annual GDP published by Statistics Poland and $\hat{Y}_t$ are estimates derived from the given model. The last column of Table 1 shows the *MSE* results for both models.

Table 1: Estimates of the model parameters of the relationship between GDP and its components: global production and intermediate consumption, obtained from the OLS and MUM model.

Method	γ_1	γ_2	*MSE*
OLS	1.948^{***}	-0.962^{***}	0.449
MUM	2.036^{***}	-1.041^{***}	0.053

Note that the MUM model's mean square error is much smaller than the OLS model's (0.053 and 0.449, respectively).

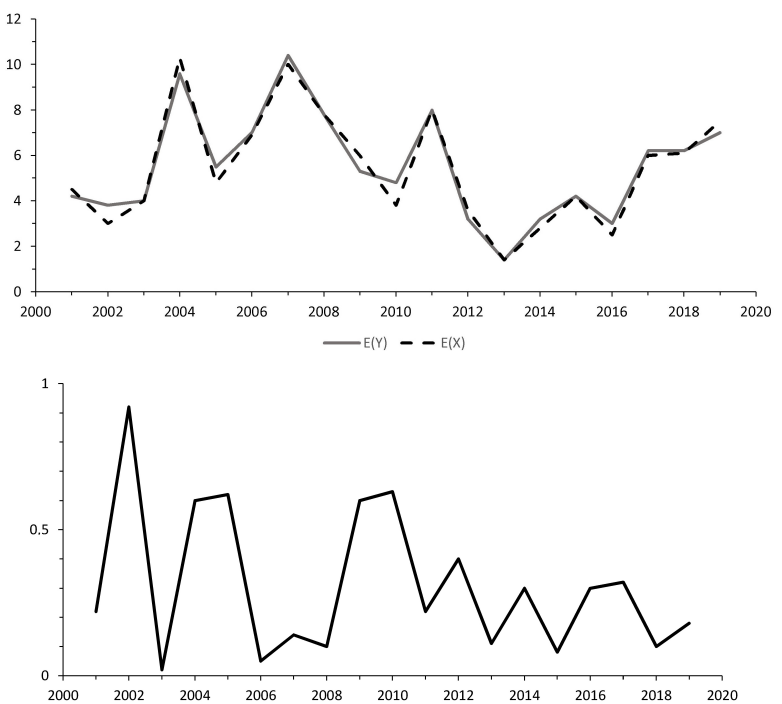


Figure 3: The expected value $E(Y_{i,t})$ and $E(X_t)$ - part of expected values, which is explained by observed data (first panel) and $|\,SE\,|$ (second panel).

From the formula (6) it is possible to determine the unobservable part of the explanatory variables:

$$\mid SE \mid = \mid \gamma_1 s_t^P + \gamma_2 s_t^Z - (\gamma_1 P_{.t} + \gamma_2 Z_{.t}) \mid = \mid \gamma_1 R_t^1 + \gamma_2 R_t^2 \mid .$$

The analysis indicates that the absolute value of the difference (denoted as $\mid SE \mid$) between the expected value $E(Y_{i,t})$ (obtained from the MUM model) and its part explained by the observed components $E(X_{i,t})$ reaches about 0.5% of GDP. The largest difference (1% of GDP) is observed only in the first half of the period. Thus, it can be concluded that in the production approach the data are correctly balanced with respect to the existence of the shadow economy.

The first panel of Figure 3 shows the expected GDP growth as the average of the growth of the voivodeships (denoted as $E(Y)$) and its part explained by the known and observed production components, denoted as $E(X)$. The second panel of Figure 3 shows the absolute value of their difference, $\mid SE \mid$, which may correspond to the existence of a shadow economy.

3.2.2 GDP - expenditure approach

The second approach to calculating GDP is based on expenditures. It takes into account household consumption, investment, government expenditures, and net exports. In addition, the model also takes into account the expenditure of the voivodeship budget. Hence, the MUM takes the form of

$$C_{i,t} = s_t^C + \varepsilon_{it}^C$$
$$I_{i,t} = s_t^I + \varepsilon_{it}^I$$
$$B_{i,t} = s_{i,t}^B + \varepsilon_{it}^B$$

and

$$Y_{i,t} = \gamma_1 s_{i,t}^C + \gamma_2 s_{i,t}^I + \gamma_3 s_{i,t}^B + \gamma_4 G_t + \gamma_5 E_t + \eta_{i,t}. \tag{8}$$

For comparison purposes, the ML estimates of the MUM are also compared with the OLS estimates. The results are shown in Table 2.

Table 2: Estimates of the model parameters of the relationship between GDP and its expenditure components, obtained from the OLS and MUM model

Method	γ_1	γ_2	γ_3	γ_4	γ_5	*MSE*
OLS	0.164	0.223^{**}	-0.050	0.635^{***}	0.208^{*}	1.738
MUM	0.126^{*}	0.605^{***}	-0.150	0.406^{***}	0.739^{**}	0.126

The MSE comparison of the two models shows that the MUM model is much more effective in the explanation of GDP than the model that does not take into account observation errors (OLS model). For the MUM the mean square error is 0.126, while for the OLS model it is as much as 1.738. Following a similar line of reasoning as in the previous section, the expected value of $Y_{i,t}$ is divided into two parts

$$E(Y_{i,t}) = \gamma_1 s_t^C + \gamma_2 s_t^I + \gamma_3 s_t^B + \gamma_4 G_{.,t} + \gamma_5 E_{.,t} =$$

$$\left(\gamma_1 C_{.,t} + \gamma_2 I_{.,t}^I + \gamma_3 B_{.,t}^B + \gamma_4 G_{.,t} + \gamma_5 E_{.,t}\right) - \left(\gamma_1 R_t^1 + \gamma_2 R_t^2 + \gamma_3 R_t^3\right)$$

In this case the level of the shadow economy $\mid SE \mid$ related to consumption, investment and expenditures of voivodeships, is estimated as:

$$\mid SE \mid = \mid \gamma_1 R_t^1 + \gamma_2 R_t^2 + \gamma_3 R_t^3 \mid .$$

It should be noted that the SE related to export and import is still unknown (no data to replicate). The first panel of Figure 4 shows the expected GDP growth as the average of the growth of the voivodeships (denoted as $E(Y)$) and its part explained by the expected value (denoted as $E(X_t)$) of the observed and known expenditure components. The second panel of Figure 4 shows the absolute value of their difference, denoted as $\mid SE \mid$, and it may also correspond to the existence of the shadow economy.

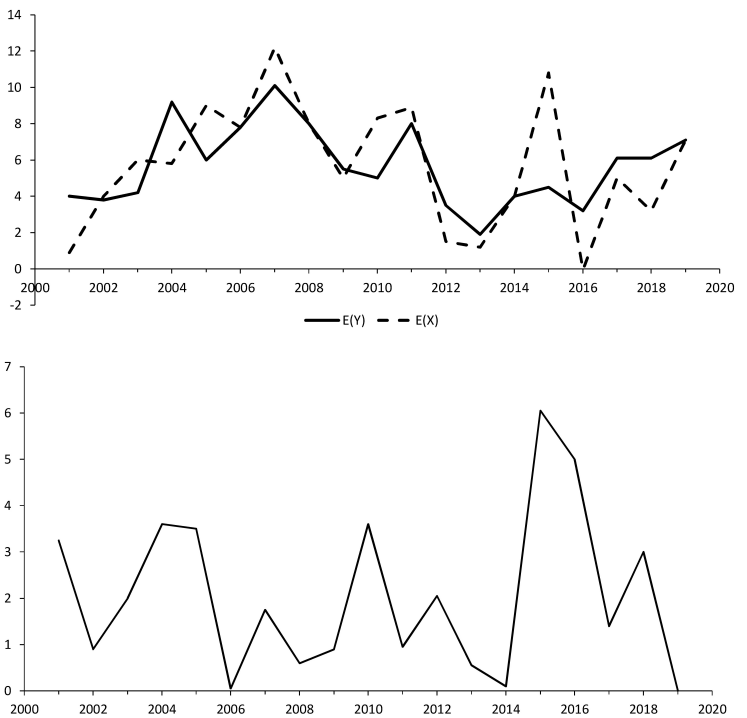


Figure 4: The expected value $E(Y_t)$ and $E(X_t)$ - part of expected values, which is explained by observed data (first panel) and $\mid SE \mid$ (second panel).

Analyzing Figure 4, we notice that the difference between the estimated expected value $E(Y_{i,t})$ and its part which is related to the observed data is relatively large. The average $\mid SE \mid$ estimate is about 3% of GDP. The highest $\mid SE \mid$ value (6% of GDP) is observed in 2016.

4. Conclusion

In the article, we presented an ultrastructural linear model (MUM) in which both the dependent and independent variables are subject to measurement error. We also developed an algorithm to estimate the parameters of this model based on the maximum likelihood method.

The MUM was used to estimate the relationship between GDP and its components, taking into account the existence of the shadow economy. In order to obtain consistent estimates of the model parameters, we adopted the voivodeship data as replications required in the MUM. However, the macroeconomic data used in the study are from Statistics Poland, which already takes into account the shadow economy. Therefore, the proposed method was considered as a way to validate or support the method of GDP adjustment in case of underestimation of the SE.

The results show that the MUM model is much more effective in explaining GDP than a model that does not take observation errors into account (OLS model). In addition, MUM provides estimates of the difference between published GDP and its components on the production side (about 0.5% of GDP) and on the expenditure side (about 3% of GDP).

The proposed methodology has several limitations. The main limitation of this method is the proper selection of variables and the method of creating their replication. In addition, the use of already aggregated data in the study leads to the loss of some information. Original survey data will be more appropriate not only to confirm the balance of GDP calculation methods, but also to estimate the size of the shadow economy. Therefore, the use of survey data in the MUM model to estimate the shadow economy (assuming that production and expenditure are balanced) will be the subject of further research.

References

Adair, P., (2021). Non-Observed Economy vs. Shadow Economy and Informal Employment in Poland: A Range of Mismatching Estimates. in: W. Andreff (Ed.). *Comparative Economic Studies in Europe. Studies in Economic Transition*. Palgrave Macmillan, Cham.

Adair, P., (2017). Non-Observed Economy vs. the Shadow Economy in the EU: The Accuracy of Measurements Methods and Estimates revisited. *4 th OBEGEF Interdisciplinary Insights on Fraud and Corruption*, Porto, Portugal.

Błasiak, Z. A., (2018). Przydatność metod ekonometrycznych w badaniach nad szarą strefą. *Roczniki Ekonomii i Zarządzania*, Towarzystwo Naukowe Katolickiego Uniwersytetu Lubelskiego Jana Pawła II, 10.

Breusch, T., (2005). Estimating the Underground Economy using MIMIC Models. *Working Paper*, Canberra, Australia.

Buszko, A., (2017). The level of shadow economy in Warmińsko-Mazurski and Kujawsko-Pomorski regions. *Copernican Journal of Finance & Accounting*, 6, pp. 9 –21.

Czapkiewicz, A., Brzozowska-Rup, K., (2021). Szacowanie rozmiar/'ow szarej strefy w Polsce. *Wiadomości Statystyczne. The Polish Statistician*, 66, pp. 7–24.

Chan, N. N., Mak, T.K. (1984). Heteroscedastic errors in a linear functional relationship. *Biometrika*, 71, pp. 212–215.

Dolby, G. R., (1976). The Ultrastructural Relation: A Synthesis of the Functional and Structural Relations. *Biometrika*, 63, pp. 39–50.

Dybka, P., Kowalczuk, M., Olesiński, B., Rozkrut, M., Torój, A., (2017). Currency demand and MIMIC models: towards a structured hybrid model-based estimation of the shadow economy size. *SGH KAE Working Papers Series*, 2017/030.

Dybka, P., Kowalczuk, M., Olesiński, B, Rozkrut, M., Torój, A., (2019). Currency demand and MIMIC models: towards a structured hybrid method of measuring the shadow economy. *Int Tax Public Finance*, 26, pp. 4–40.

European Commission, (2010). European system of account (ESA 2010).

Feige, E. L., (1997). Revisited Estimates of Underground Economy: Implications of US Currency Held Abroad, in O. Lippert and M. Walker (eds). *The Underground Economy, The Fraser Institute*, Canada.

Kirchgässner, G., (2016). On estimating the size of the shadow economy. *German Economic Review*, 18(1), pp. 99–111.

Kendall, M. G., Stuart A., (1973). *The Advanced Theory of Statistics Volume Two*. Charles Griffin and Co Ltd, London, Third edition.

Madzarevic-Sujster, S., Mikulic D., (2002). An Estimate of the Underground Economy via the National Accounts System. *Institute of Public Finance*, Zagreb, 26, pp. 31–56.

Malczewska, P., (2019). Szara strefa gospodarki. Determinanty i mechanizmy kształtowania. *Wydawnictwo Uniwersytetu Łódzkiego*.

Medina, L., Schneider, F. G., (2019). Shedding Light on the Shadow Economy: A Global Database and the Interaction with the Official One. *CESifo Working Paper*, 7981.

Schneider, F., (2005a). Shadow Economies around the World: What Do We Really Know? *European Journal of Political Economy*.

Schneider, F., Dellanno, R., (2005b). The Shadow Economy of Italy and other OECD Countries: What do we know? *Journal of Public Finance and Public Choice*.

Schneider, F., Buehn, A., (2016). Estimating the Size of the Shadow Economy: Methods, Problems and Open Questions. *IZA Discussion Paper*, 9820.

Schneider F., Medina L., (2018). Shadow Economies around the World: What Did We LEARN Over the Last 20 Years?. International Monetary Fund Working Paper, WP/18/17.

Trebicka, M., (2014). MIMIC model: A tool to estimate the shadow economy. *Academic Journal of Interdisciplinary Studies*, 3, pp. 295–300.

Wyżnikiewicz B., (2017). Produkt krajowy brutto jako przedmiot krytyki. *Wiadomości Statystyczne*, 3, pp. 5–15.

OECD, (2002). Measuring the Non-Observed Economy: A Handbook.

Raport 2015/2016, Przeciwdziałanie szarej strefie w Polsce. *The Global Compact Network Poland*.

Statistics Poland Local Data Bank, website: https://bdl.stat.gov.pl/BDL/.

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 83–102, https://doi.org/10.59170/stattrans-2024–029
Received – 07.04.2023; accepted – 25.04.2024

Extropy and entropy estimation based on progressive Type-I interval censoring

Huda H. Qubbaj[1], **Husam A. Bayoud**[2], **Hisham M. Hilow**[3]

Abstract

This paper proposes nonparametric estimates for the two information measures extropy and entropy when a progressively Type-I interval censored data is available. Different nonparametric approaches are used for deriving the estimates, including: moments of the empirical cumulative distribution function and linear regression. The performance of the proposed estimates is studied under various censoring schemes via simulation studies. Furthermore, different real data sets are analyzed for illustrative purposes.The estimates based on linear approximation $\hat{J}_2$ and $\hat{H}_2$ outperform the other estimate in the majority of studied cases.
Key words: entropy; extropy; mean square error; nonparametric statistics; Monte Carlo simulation; Type-I interval censoring.

1. Introduction

[Shannon,1948] defined entropy of a random variable (r.v.) X whose probability density function (pdf) $f(x)$ and cumulative distribution function (cdf) $F(x)$ as:

$$H(X) = -\int_{R_X} f(x) log\,(f(x))\,dx, \tag{1}$$

where R_X denotes the support of the r.v. X.

For more details on entropy the reader can see [Renyi,1961], [Awad,1987], [Tsallis,1988], [Rao et al.,2004] and [Kittaneh et al.,2016].

The differential extropy of X is defined by [Lad et al.,2015] as:

$$J(X) = -\frac{1}{2}\int f^2(x)dx. \tag{2}$$

Important properties of the extropy measure have been discussed in the literature since 2015. See [Qiu,2017] and [Qiu and Jia,2018a], who studied properties of the residual ex-

[1]Department of Mathematics, The University of Jordan, Jordan. E-mail: h.qubaj@ju.edu.jo; hudaqubbaj.ctc@gmail.com. ORCID: https://orcid.org/0009-0004-9364-5082.
[2]College of Sciences and Humanities, Fahad Bin Sultan University, Saudi Arabia. E-mail: hbayoud@fbsu.edu.sa;husam.awni@yahoo.com. ORCID: https://orcid.org/0000-0002-0066-7206.
[3]Department of Mathematics, The University of Jordan, Jordan. E-mail: hhilow@ju.edu.jo. ORCID: https://orcid.org/0000-0001-7291-3579.

tropy and the extropy of ordered statistics as well as the extropy of record values; that is for random variable X its residual extropy is defined as:

$$J_t(X) = -\frac{1}{2}\int_0^\infty f_t^2(x)dx = -\frac{1}{2\bar{F}^2(T)}\int_t^\infty f^2(x)dx, t \geq 0 \tag{3}$$

where,

$$f_t(X) = \frac{f(x+t)}{\bar{F}(T)}, x \geq 0, t \geq 0 \tag{4}$$

and $\bar{F}$ is the survival function of X.

[Raqab and Qiu,2019] studied properties of the extropy measure under ranked set sampling. The problem of estimating the extropy based on complete sample data has recently been considered by some authors, including: [Qiu and Jia,2018b] and [Noughabi and Jarrahiferiz, 2019]. [Hazeb et al.,2021a] and [Hazeb et al.,2021b] introduced nonparametric estimates for extropy and entropy based on progressively Type-II censored data. However, estimation of the extropy and entropy measures under progressively Type-I interval censored data have not been considered so far in the literature. Accordingly, our main objective in this paper is to develop different methods for estimating the extropy and entropy measures under the progressive Type-I censoring set-up.

Censoring schemes of statistical experiments arise naturally in survival, reliability and medical studies. [Cohen,1963] introduced progressive Type-I censoring as an extension of Type-I censoring, where in a progressively Type-I censored life test on n items, progressive censoring is carried out at the prefixed censoring times $t_1 < t_2 < ... < t_k$. That is, at the i^{th} censoring time t_i, R_i items are randomly removed from the experiment, $1 \leq i \leq k-1$, with the restriction $R_1 + ... + R_{k-1} \leq n - l$, $l \in \{0, 1, ..., n\}$. Then, at the k^{th} censoring time t_k, all remaining items are removed from the life test if there are any left. In many practical situations lifetimes of units placed on a test are observed within intervals, where this censoring scheme is called interval censoring.

[Aggarwala,2001] introduced progressive Type-I Interval censoring combining the two concepts: progressive Type-I censoring and Interval censoring, then developed statistical inference for the exponential distribution based on progressively type-I interval censored data. Under progressive type-I interval censoring, observations are only known within two consecutively pre-scheduled times, where items would be allowed to be withdrawn at pre-scheduled time points. [Ng and Wang,2009] introduced the concept of progressive type-I interval censoring using the Weibull distribution and compared many different estimation methods for two parameters of the Weibull distribution via simulation. [Lio et al.,2011] proposed parameter estimation for the generalized Rayleigh distribution under progressively Type-I interval censored data. Parameters of the inverse Weibull distribution under progressive type-I Interval censoring were estimated by [Singh and Tripathi,2016]. Also [Du et al.,2018] proposed Statistical Inference for the Entropy of the Log-Logistic Distribution under Progressive Type-I Interval Censoring Schemes. Recently, [Al otaibi et al.,2021] introduced Bayesian estimation for Dagum distribution based on progressive type -I Interval censoring.

The rest of the paper is organized as follows. Progressive Type-I interval censoring scheme with its properties are discussed in Section 2. Nonparametric estimates for extropy and entropy measures based on progressive Type-I interval censoring are developed in Section 3. Simulation experiments as well as analyses of real life data are performed in Section 4. Finally, we end the the paper in Section 5 with a conclusion.

2. Progressive Type-I Interval Censoring

Several censoring schemes have been proposed in the literature for saving time and cost of reliability experiments. Type I and type II censoring are probably the most commonly used methods in this regard. In type I censoring, failure times are recorded up to a specified time point where failures occurring afterward this fixed time are not recorded. On the other hand, in type II censoring a life test continues until a pre-specified number of failures have been recorded. These two censoring methods share one common feature that live units can be withdrawn only at the termination point of the test. In progressive censoring, items put on a test can be withdrawn during life experimentation, hence this censoring scheme is more flexible compared to the other two basic schemes. Progressive censoring has been widely studied in lifetime analysis by various researchers. One may refer to [Balakrishnan and Cramer ,2014] for an exhaustive list of references on this topic and also for a detailed discussion on applications of progressive censoring in life testing experiments. Sometimes it is difficult to record exact failure times of items under life testing due to lack of continuous monitoring of subjects or items under study. In such situations, observations are often recorded in intervals and the corresponding censoring is referred to as the interval censoring. [Aggarwala,2001] initially discussed progressive type I interval censoring in the literature and studied an exponential distribution using this censoring. Since then this censoring scheme has attracted attention among researchers. Progressive Type-I interval censoring can be briefly described as follows: Suppose n identical items are placed simultaneously on life testing at time $t_0 = 0$, where inspection is at m pre-fixed censoring times $t_1 < t_2 <.....< t_m$, and where t_m is the scheduled time to terminate the experiment and m is pre-fixed number of stops. For $i = 1,2,...,m$, let k_i be the number of failures in the interval $(t_{i-1},t_i]$. Let S_i be the number of the surviving items at t_i and R_i be the number of removed items at t_i. In this censoring scheme, k_i and S_i are random numbers while R_i is the number of remaining items, which is also a random number. At the 1^{st} inspection time t_1, we observe k_1 failures, then R_1 surviving items are randomly withdrawn from the remaining items $n-k_1$. One can see that after this step, the number of remaining items is $(n-k_1-R_1)$. Now, after time t_1 and at the 2^{nd} inspection time t_2, we observe k_2 failures where R_2 items are randomly removed from $(n-k_1-k_2-R_1)$ items. Lastly, at the m^{th} inspection time (the last inspection time), we observe k_m failures and all remaining $(n-\sum_{i=1}^{m} k_i-\sum_{i=1}^{m-1} R_i)$ items are immediately removed from the experiment. The observed progressive Type-I interval censored data can be represented as: $\{(k_i,R_i,t_i),i = 1,2,...,m\}$. The associated likelihood function under the progressive type-I interval censoring is given by:

$$L(\theta) \propto \prod_{i=1}^{m} [F(t_i;\theta)-F(t_{i-1};\theta)]^{k_i} [1-F(t_i;\theta)]^{R_i}. \tag{5}$$

Note that R_i should not be greater than S_i,where the values of R_i for $i = 1,2,...,m$ are determined based on pre-specified removal proportions $q_1, q_2, ..., q_{k-1}$ and $q_m = 1$, such that $R_i = [S_i q_i]$, for $i = 1,2,...,k-1$, where symbol $[b]$ is the greatest integer less than or equal to b. It can be easily seen that $n = \sum_{i=1}^{m}(R_i + k_i)$. If $R_i = 0$, for $i = 1,2,...,m-1$, then Progressive type-I interval censoring reduces to the conventional type-I censoring.

Theorem 2.1. *Let $U_{i:m:n} = F(t_i)$, $i = 1,2,...,m$ denote a progressively Type-I interval censoring sample obtained from the uniform $(0,1)$ distribution, assuming the sample size is n with progressive Type-I interval censored data $\{(k_i, R_i, t_i), i = 1,2,...,m\}$. Let*

$$U_{i:m:n} = 1 - \prod_{j=m-i+1}^{m} V_j,$$

where,

$$V_1 = \frac{1-U_{m:m:n}}{1-U_{m-1:m:n}}, V_2 = \frac{1-U_{m-1:m:n}}{1-U_{m-2:m:n}}, ..., V_m = 1 - U_{1:m:n}, \tag{6}$$

are all independent identically distributed (iid) r.v.'s. Then

$$V_i \overset{d}{=} Beta\left(i + \sum_{j=m-i+2}^{m} k_j + \sum_{j=m-i+1}^{m} R_j, k_{m-i+1} + 1\right), \qquad i = 1,2,...,m. \tag{7}$$

Proof. For simplicity, we denote $U_{i:m:n}$ by U_i.

Since U_i follows $U(0,1)$, then the pdf and cdf of U_i are $f_{U_i}(u) = 1$ and $F_{U_i}(u) = u$, for $0 \le u \le 1$, respectively. So, using Eq.(5), the joint pdf of $U_1 \le U_2 \le ... \le U_m$ is obtained as follows:

$$f_{U_{l:m:n}, U_{2:m:n}, ..., U_{m:m:n}}(u_1, u_2, .., u_m) \propto \prod_{i=1}^{m} (u_i - u_{i-1})^{k_i} (1-u_i)^{R_i}. \tag{8}$$

Since $U_i = 1 - \prod_{j=m-i+1}^{m} V_j$, one can see that:

$$\begin{aligned} U_1 &= 1 - V_m \\ U_2 &= 1 - V_{m-1}V_m \\ U_3 &= 1 - V_{m-2}V_{m-1}V_m \\ &\vdots \\ U_m &= 1 - V_1V_2V_3...V_m \end{aligned} \tag{9}$$

The Jacobian matrix of this transformation is given by the following lower triangular matrix:

$$J = \frac{\partial}{\partial V} U = \begin{pmatrix} 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & -V_m & -V_{m-1} \\ \vdots & \vdots & \ddots & \vdots & \\ -V_2 V_3 ... V_m & -V_1 V_3 ... V_m & \cdots & -V_1 V_2 ... V_{m-2} V_m & -V_1 V_2 ... V_{m-1} \end{pmatrix}$$

Therefore,

$$|J| = \left| \frac{\partial}{\partial V} U \right| = \prod_{j=2}^{m} V_j^{j-1}. \tag{10}$$

Now, assuming $U_0 = 0$, we have:

$$\begin{aligned} U_1 - U_0 &= 1 - V_m \\ U_2 - U_1 &= V_m (1 - V_{m-1}) \\ U_3 - U_2 &= V_{m-1} V_m (1 - V_{m-2}) \\ &\vdots \\ U_m - U_{m-1} &= V_2 V_3 ... V_m (1 - V_1) \end{aligned} \tag{11}$$

Now, in order to derive the joint pdf of $V_1, V_2, ..., V_m$, we simplify the terms of Eq.(8) separately.

Using Eq.(10), the first term in Eq.(8) is simplified to:

$$\prod_{i=1}^{m} (U_i - U_{i-1})^{k_i} = \prod_{i=1}^{m} (1 - V_{m-i+1})^{k_i} \prod_{i=1}^{m} V_{m-i+1}^{\sum_{j=i+1}^{m} k_j}. \tag{12}$$

Using Eq.(9), the second term in Eq.(8) is simplified to:

$$\prod_{i=1}^{m} (1 - U_i)^{R_i} = \prod_{i=1}^{m} V_{m-i+1}^{\sum_{j=i}^{m} R_j}. \tag{13}$$

Accordingly, the joint pdf of $V_1, ..., V_m$ is obtained as follows:

$$f_{V_l, V_2, ..., V_m}(v_1, v_2, .., v_m) \propto \prod_{i=1}^{m} (1 - V_i)^{k_m + i - 1} \prod_{i=2}^{m} V_i^{\sum_{j=m-i+2}^{m} k_j} \prod_{i=1}^{m} V_i^{\sum_{j=m-i+1}^{m} R_j} \prod_{i=2}^{m} V_i^{i-1}. \tag{14}$$

Since $\prod_{j=2}^{m} V_j^{j-1} = \prod_{j=1}^{m} V_j^{j-1}$ and $\prod_{i=2}^{m} V_i^{\sum_{j=m-i+2}^{m} k_j} = \prod_{i=1}^{m} V_i^{\sum_{j=m-i+2}^{m} k_j}$.

Then, Eq.(14), can be simplified to:

$$f_{V_l,V_2,...,V_m}(v_1,v_2,..,v_m) \propto \prod_{i=1}^{m}(1-V_i)^{k_m+i-1} V_i^{\sum_{j=m-i+2}^{m} k_j+\sum_{j=m-i+1}^{m} R_j+i-1} \tag{15}$$

where $0 < V_i < 1$.

By factorization theorem, we see that $V_1, V_2, ..., V_m$ are independent and

$$V_i \overset{d}{=} Beta\left(i+\sum_{j=m-i+2}^{m} k_j+\sum_{j=m-i+1}^{m} R_j, k_{m-i+1}+1\right), \qquad i=1,2,...,m.$$

□

Corollary 2.2. *As a result of Theorem (2.1), we find*

$$E(U_{i:m:n}) = 1-\prod_{j=m-i+1}^{m}\gamma_j, \tag{16}$$

where,

$$\gamma_i = \frac{i+\sum_{j=m-i+2}^{m} k_j+\sum_{j=m-i+1}^{m} R_j}{1+i+\sum_{j=m-i+1}^{m} k_j+R_j},$$

such that $\gamma_j = \gamma_1$ *if* $j \le 1$ *and* $\gamma_j = \gamma_m$ *if* $j \ge m$ *provided that* $\sum_{j=m+1}^{m} k_j = 0$.

3. Nonparametric Extropy and Entropy Estimates

This section develops non-parametric estimates for the extropy and entropy measures based on progressively Type-I interval censored samples. It is of importance here to mention that for a random variable (r.v.) T, extropy and entropy measures $J(T)$ and $H(T)$ are expressed as:

$$J(T) = -\frac{1}{2}\int_0^1 \left(\frac{d}{dp}F^{-1}(p)\right)^{-1} dp \tag{17}$$

and

$$H(T) = \int_0^1 log\left(\frac{d}{dp}F^{-1}(p)\right) dp \tag{18}$$

3.1. Moments Approximation Method

The first entropy and extropy estimate will be obtained by using the difference operator that was proposed by Vasicek (1976) for estimating the entropy. This method is based on the following fact:

$$\frac{d}{dp}F^{-1}(p) \approx \frac{T_{i+w:m:n}-T_{i-w:m:n}}{F(T_{i+w:m:n})-F(T_{i-w:m:n})}, \tag{19}$$

where the window size $w \leq m/2$, also $T_{i:m:n} = T_0 = 0$ if $i < 1$, while $T_{i:m:n} = T_{m:m:n}$ if $i > m$. Then $T_0 = 0 \leq T_{1:m:n} \leq T_{2:m:n} \leq ... \leq T_{m:m:n}$ are progressively Type-I interval censoring times of size m, which are pre-fixed. In the interval $[T_{i-1}, T_i)$, we observe a random number of failures, say k_i, then R_i surviving units are immediately removed from the remaining $(n - \sum_{j=1}^{i} k_i - \sum_{j=1}^{i-1} R_i)$ items.

Next, it can be seen that (17) is approximately equal to:

$$J(T) \approx -\frac{1}{2m} \sum_{i=1}^{m} \frac{F(T_{i+w:m:n}) - F(T_{i-w:m:n})}{T_{i+w:m:n} - T_{i-w:m:n}}. \tag{20}$$

Notice that $U_{i+w:m:n} = F(T_{i+w:m:n})$ and $U_{i-w:m:n} = F(T_{i-w:m:n})$, the moments-based estimate is proposed by replacing $U_{i+w:m:n}$ and $U_{i-w:m:n}$ by their moments, i.e. their expected values $E(U_{i+w:m:n})$ and $E(U_{i-w:m:n})$, respectively, which are given in (16). Consequently, an estimate of $J(T)$ is

$$\hat{J}_1 = -\frac{1}{2m} \sum_{i=1}^{m} \frac{\prod_{j=m-(i-w)+1}^{m} \gamma_j - \prod_{j=m-(i+w)+1}^{m} \gamma_j}{T_{i+w:m:n} - T_{i-w:m:n}}. \tag{21}$$

On the lines of Vasicek (1976), an estimate of the entropy is

$$\hat{H}_1 = \frac{1}{m} \sum_{i=1}^{m} log \left(\frac{T_{i+w:m:n} - T_{i-w:m:n}}{\prod_{j=m-(i-w)+1}^{m} \gamma_j - \prod_{j=m-(i+w)+1}^{m} \gamma_j} \right). \tag{22}$$

Proposition 3.1.1. *Let $Y = aT + b$, $a > 0$. Then $\hat{J}_1{}^Y = \frac{1}{a}\hat{J}_1{}^T$ and $\hat{H}_1{}^Y = log(a) + \hat{H}_1{}^T$.*

Proof. The result follows since for all i, $\gamma_i^Y = \gamma_i^T$. □

Proposition 3.1.2. *Estimates $\hat{J}_1$ and $\hat{H}_1$ are consistent estimates for J and H respectively, i.e.*

$$\hat{J}_1 \xrightarrow{p} J,$$

and

$$\hat{H}_1 \xrightarrow{p} H.$$

as $n \to \infty$, $m \to \infty$, and $m/n \to 0$.

Proof. Since when $m \to n$, the progressively Type-I interval censored sample becomes the complete sample. Then the estimates, $\hat{J}_1$ and $\hat{H}_1$ converge to the estimates proposed by [Qiu and Jia,2018b] and [Vasicek,1976], which are consistent estimates of J and H. □

3.2. Linear Approximation Method

The second estimate for extropy $J(T)$ in (2) is proposed following the steps of Correa (1995) by noticing that the quantity:

$$\frac{F(T_{i+w:m:n}) - F(T_{i-w:m:n})}{T_{i+w:m:n} - T_{i-w:m:n}}, \tag{23}$$

represents the slope of a straight line joining the following two points

$$(T_{i-w:m:n}, F(T_{i-w:m:n})) \text{ and } (T_{i+w:m:n}, F(T_{i+w:m:n})).$$

This estimation approach is based on estimating the function $F(T_j)$ by a local linear model on $(T_{i-w:m:n}, T_{i+w:m:n})$ by using $2w+1$ ordered pairs:

$$F(T_j) = U_j = a + bT_j + \varepsilon, j = i-w, ..., i+w. \tag{24}$$

On the other hand, slope in (23) can be approximated by b in (22), which can be estimated by the least squares method using $(2w+1)$ ordered pairs as follows:

$$b = \frac{S_{TU}}{S_T^2} = \frac{\sum_{j=i-w}^{i+w}(T_{j:m:n} - \bar{T}_{(i)})(U_{j:m:n} - \bar{U}_{(i)})}{\sum_{j=i-w}^{i+w}(T_{j:m:n} - \bar{T}_{(i)})^2}, \tag{25}$$

where

$$\bar{T}_{(i)} = \frac{1}{2w+1}\sum_{j=i-w}^{i+w} T_{j:m:n}, \text{ and } \bar{U}_{(i)} = \frac{1}{2w+1}\sum_{j=i-w}^{i+w} U_{j:m:n}.$$

Now, by replacing $U_{j:m:n}$ by its expected value, we get

$$\hat{\bar{U}}_{(i)} = \frac{1}{2w+1}\sum_{j=i-w}^{i+w}\left(1 - \prod_{k=m-j+1}^{m}\gamma_k\right).$$

Consequently, a second estimate of $J(T)$ using the slope b in Eq.(25) and replacing $F(T_{j:m:n}) = U_{j:m:n}$ by its respective expected value in terms of γ_j in Eq.(21) is as follows:

$$\hat{J}_2 = -\frac{1}{2m}\sum_{i=1}^{m}\frac{\sum_{j=i-w}^{i+w}(T_{j:m:n} - \bar{T}_{(i)})(\frac{\sum_{j=i-w}^{i+w}\prod_{k=m-j+1}^{m}\gamma_k}{2w+1} - \prod_{k=m-j+1}^{m}\gamma_k)}{\sum_{j=i-w}^{i+w}(T_{j:m:n} - \bar{T}_{(i)})^2}. \tag{26}$$

Following the same argument proposed for $\hat{J}_2$ in Eq.(26), we consider the slope of the linear regression of T on F as follows:

$$b_h = \frac{S_{TU}}{S_U^2} = \frac{\sum_{j=i-w}^{i+w}(T_{j:m:n} - \bar{T}_{(i)})(U_{j:m:n} - \bar{U}_{(i)})}{\sum_{j=i-w}^{i+w}(U_{j:m:n} - \bar{U}_{(i)})^2}, \tag{27}$$

Using Eq(18), Eq(19) and Eq(27), a second estimate for $H(T)$ in (1) can be introduced as

$$\hat{H}_2 = \frac{1}{m}\sum_{i=1}^{m} log\left[\frac{\sum_{j=i-w}^{i+w}(T_{j:m:n} - \bar{T}_{(i)})(\frac{\sum_{j=i-w}^{i+w}\prod_{k=m-j+1}^{m}\gamma_k}{2w+1} - \prod_{k=m-j+1}^{m}\gamma_k)}{\sum_{j=i-w}^{i+w}(\frac{\sum_{j=i-w}^{i+w}\prod_{k=m-j+1}^{m}\gamma_k}{2w+1} - \prod_{k=m-j+1}^{m}\gamma_k)^2}\right]. \tag{28}$$

Proposition 3.2.1. *Let $Y = aT + b$, $a > 0$. Then $\hat{J}_2^{\,Y} = \frac{1}{a}\hat{J}_2^{\,T}$ and $\hat{H}_2^{\,Y} = log(a) + \hat{H}_2^{\,T}$.*

Proof. The result follows since for all i, $\gamma_i^Y = \gamma_i^T$. □

Proposition 3.2.2. *Estimates $\hat{J}_2$ and $\hat{H}_2$ are consistent estimates for J and H respectively, i.e.*

$$\hat{J}_2 \xrightarrow{p} J,$$

and

$$\hat{H}_2 \xrightarrow{p} H.$$

as $n \to \infty$, $m \to \infty$ and $m/n \to 0$.

Proof. Proof is obvious by [Correa,1995] and [Qiu and Jia,2018b] and so it is omitted. □

4. Simulation study and data analysis

Here, we perform an extensive simulation study to compare the act of the proposed extropy and entropy estimates. Moreover, a real data set is discussed for illustrative purposes.

4.1. Simulation study

In this subsection, we carry out a Monte Carlo simulation to analyze the behavior of our proposed estimators of entropy and extropy. In order to perform the simulation process, we consider different sample sizes, i.e. $n = 10, 20, 30, 50, 100$ with different inspection times, i.e. $m = 3, 4, 5$. Next, we generate 1000 progressive type-I interval censoring data sets in each experimentation case. Also, we work with different withdrawal (removal) schemes as detailed in Table 1. We consider the uniform distribution $U(0, \theta)$ with $\theta = 1$, the exponential distribution $Exp(\beta)$ with $\beta = 1$ and normal distribution $N(\mu, \sigma)$ with $\mu = 0, \sigma = 1$, which are commonly used.

Table 1: Progressive interval censoring schemes used in the Monte Carlo simulation study

Scheme No.	m	$(t_1, ..., t_m)$	$(q_1, ..., q_m)$
1	3	$0.1, 0.5, 0.9$	$0.25, 0.1, 1$
2	3	$0.1, 0.5, 0.9$	$0.5, 0.1, 1$
3	3	$0.1, 0.5, 0.9$	$0, 0.25, 1$
4	3	$0.1, 0.5, 0.9$	$0.1, 0.1, 1$
5	4	$0.1, 0.5, 0.7, 0.9$	$0, 0, 0.25, 1$
6	4	$0.1, 0.5, 0.7, 0.9$	$0, 0.25, 0.25, 1$
7	4	$0.1, 0.5, 0.7, 0.9$	$0.25, 0, 0, 1$
8	4	$0.1, 0.5, 0.7, 0.9$	$0.2, 0.2, 0.2, 1$
9	5	$0.1, 0.3, 0.5, 0.7, 0.9$	$0.25, 0, 0, 0, 1$
10	5	$0.1, 0.3, 0.5, 0.7, 0.9$	$0, 0, 0, 0, 1$
11	5	$0.1, 0.3, 0.5, 0.7, 0.9$	$0.25, 0.25, 0, 0, 1$
12	5	$0.1, 0.3, 0.5, 0.7, 0.9$	$0.1, 0.1, 0.2, 0.2, 1$

For each generated data set, we compute the average estimator and the corresponding mean squared error (MSE) of the proposed entropy and extropy estimators over 1000 simulations. Simulation results are shown in Tables (2-10), where the bold type in these tables indicates the estimator achieving the minimal MSE. Comments on simulation results are provided after these tables. Here, we will detail the withdrawal schemes used through Tables 2 to 10.

For schemes 1, 2, 3 and 4, we specify three inspection (stopping) times at which surviving units are progressively removed at different proportions as detailed in Tables 2, 5 and 8. In scheme 1, censoring by removal is heavier at the end of the first interval and lighter at the end of the second interval. In scheme 2, the same censoring approach is applied but with different proportions. In scheme 3, the reverse situation is applied, i.e. the censoring by removal is lighter at the end of the first interval and heavier at the end of the second interval. In scheme 4, equal (uniform) proportions are used.

Similarly, for schemes 5, 6, 7 and 8, we specify the proportions of the surviving units to be removed at four inspection times as presented in Tables 3, 6 and 9. In scheme 5, censoring by removal is lighter at the end of the first and second intervals and heavier at the end of the third interval. In scheme 6, censoring by removal is lighter at the end of the first interval and heavier at the end of the second and third intervals. In scheme 7, only left-most and right-most removal is applied. In scheme 8, equal proportions (uniform) are used.

At last, for schemes 9, 10, 11 and 12, we specify the proportions of the surviving units to be removed at five inspection times as shown in Tables 4, 7 and 10. In scheme 9, only left-most and right-most removal is applied. In scheme 10, a conventional interval censoring scheme is employed, i.e. all removal is carried out following the final time interval, immediately prior to experiment termination. In scheme 11, censoring by removal is heavier at the end of the first and second intervals and lighter at the end of the third and fourth intervals. In scheme 12, censoring by removal is lighter at the end of the first and second intervals and heavier at the end of the third and fourth intervals.

Table 2: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for uniform $(0,1)$ assuming $m = 3$

Scheme Number	n	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
1	10	**0.0015**	0.0098	**0.0098**	0.0160
	20	**0.0003**	0.0066	**0.0018**	0.0050
	30	**0.0003**	0.0054	**0.0013**	0.0040
	50	**0.0002**	0.0046	**0.0005**	0.0025
	100	**0.0001**	0.0038	**0.0003**	0.0015
2	10	**0.0021**	0.0087	**0.0125**	0.0214
	20	**0.0009**	0.0050	**0.0044**	0.0088
	30	**0.0007**	0.0036	**0.0028**	0.0051
	50	**0.0006**	0.0025	**0.0023**	0.0033
	100	**0.0006**	0.0015	**0.0021**	0.0013
3	10	**0.0012**	0.0100	**0.0077**	0.0144
	20	**0.0006**	0.0074	**0.0033**	0.0084
	30	**0.0003**	0.0062	**0.0015**	0.0055
	50	**0.0001**	0.0053	**0.0006**	0.0039
	100	**0.0001**	0.0047	**0.0003**	0.0031
4	10	**0.0009**	0.0115	**0.0068**	0.0099
	20	**0.0005**	0.0088	**0.0031**	0.0066
	30	**0.0002**	0.0074	**0.0014**	0.0039
	50	**0.0001**	0.0066	**0.0008**	0.0031
	100	**0.0000**	0.0057	**0.0003**	0.0020

Table 3: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for uniform $(0,1)$ distribution assuming $m = 4$

Scheme Number	n	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
5	10	**0.0037**	0.0076	**0.0246**	0.0281
	20	**0.0030**	0.0065	**0.0206**	0.0235
	30	**0.0019**	0.0048	**0.0120**	0.0143
	50	**0.0013**	0.0038	**0.0081**	0.0098
	100	**0.0006**	0.0027	**0.0035**	0.0048
6	10	**0.0041**	0.0077	**0.0278**	0.0311
	20	**0.0038**	0.0068	**0.0277**	0.0302
	30	**0.0028**	0.0054	**0.0207**	0.0226
	50	**0.0020**	0.0042	**0.0141**	0.0154
	100	**0.0010**	0.0028	**0.0064**	0.0075
7	10	**0.0036**	0.0070	**0.0233**	0.0271
	20	**0.0027**	0.0051	**0.0186**	0.0210
	30	**0.0020**	0.0038	**0.0132**	0.0151
	50	**0.0011**	0.0023	**0.0060**	0.0071
	100	**0.0006**	0.0012	**0.0024**	0.0027
8	10	**0.0039**	0.0072	**0.0257**	0.0295
	20	**0.0031**	0.0057	**0.0206**	0.0234
	30	**0.0030**	0.0052	**0.0200**	0.0221
	50	**0.0019**	0.0034	**0.0110**	0.0120
	100	**0.0009**	0.0020	**0.0053**	0.0059

Table 4: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for uniform $(0,1)$ distribution assuming $m = 5$

Scheme Number	n	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
9	10	0.0039	**0.0038**	**0.0145**	0.0158
	20	0.0038	**0.0031**	**0.0136**	0.0142
	30	0.0037	**0.0025**	**0.0123**	**0.0123**
	50	0.0030	**0.0014**	0.0063	**0.0056**
	100	0.0026	**0.0007**	0.0042	**0.0026**
10	10	**0.0038**	**0.0038**	**0.0155**	0.0163
	20	0.0030	**0.0025**	**0.0101**	0.0104
	30	0.0028	**0.0018**	**0.0075**	**0.0075**
	50	0.0023	**0.0009**	0.0035	**0.0030**
	100	0.0021	**0.0003**	0.0023	**0.0014**
11	10	**0.0045**	0.0048	**0.0178**	0.0201
	20	0.0042	**0.0035**	**0.0155**	0.0162
	30	0.0040	**0.0030**	0.0145	**0.0143**
	50	0.0036	**0.0021**	0.0107	**0.0096**
	100	0.0030	**0.0011**	0.0075	**0.0049**
12	10	**0.0043**	0.0045	**0.0183**	0.0199
	20	**0.0036**	0.0039	**0.0175**	0.0184
	30	**0.0029**	0.0031	**0.0134**	0.0138
	50	**0.0024**	0.0025	0.0118	**0.0117**
	100	0.0015	**0.0013**	0.0057	**0.0050**

Table 5: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for exponential distribution with $\theta = 1$ assuming $m = 3$

Scheme Number	n	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
1	10	0.0247	**0.0095**	0.5494	**0.4789**
	20	0.0220	**0.0082**	0.5288	**0.4469**
	30	0.0204	**0.0073**	0.5154	**0.4305**
	50	0.0181	**0.0059**	0.4977	**0.4079**
	100	0.0163	**0.0047**	0.4853	**0.3922**
2	10	0.0262	**0.0113**	0.5742	**0.4888**
	20	0.0241	**0.0107**	0.5411	**0.4580**
	30	0.0221	**0.0096**	0.5206	**0.4328**
	50	0.0195	**0.0080**	0.4973	**0.4076**
	100	0.0178	**0.0066**	0.4892	**0.3949**
3	10	0.0209	**0.0080**	0.4945	**0.4214**
	20	0.0187	**0.0067**	0.4817	**0.4020**
	30	0.0176	**0.0057**	0.4805	**0.3958**
	50	0.0166	**0.0050**	0.4744	**0.3866**
	100	0.0146	**0.0037**	0.4614	**0.3708**
4	10	0.0228	**0.0078**	0.5300	**0.4692**
	20	0.0193	**0.0063**	0.4963	**0.4214**
	30	0.0189	**0.0060**	0.4995	**0.4205**
	50	0.0176	**0.0051**	0.4928	**0.4094**
	100	0.0160	**0.0041**	0.4841	**0.3967**

Table 6: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for exponential distribution with $\theta = 1$ assuming $m = 4$

Scheme Number	n		$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
5	10	0.0169	**0.0101**	0.4120	**0.3790**	
	20	0.0122	**0.0076**	0.3306	**0.3036**	
	30	0.0096	**0.0065**	0.2832	**0.2602**	
	50	0.0071	**0.0048**	0.2488	**0.2273**	
	100	0.0040	**0.0031**	0.1767	**0.1592**	
6	10	0.0211	**0.0127**	0.4892	**0.4473**	
	20	0.0139	**0.0090**	0.3453	**0.3147**	
	30	0.0120	**0.0083**	0.3106	**0.2842**	
	50	0.0083	**0.0057**	0.2547	**0.2321**	
	100	0.0049	**0.0040**	0.1823	**0.1653**	
7	10	0.0194	**0.0119**	0.4544	**0.4148**	
	20	0.0144	**0.0092**	0.3654	**0.3298**	
	30	0.0122	**0.0080**	0.3303	**0.2973**	
	50	0.0093	**0.0059**	0.3006	**0.2681**	
	100	0.0064	**0.0037**	0.2589	**0.2262**	
8	10	0.0191	**0.0115**	0.4568	**0.4149**	
	20	0.0147	**0.0095**	0.3602	**0.3261**	
	30	0.0125	**0.0081**	0.3317	**0.2999**	
	50	0.0096	**0.0065**	0.2888	**0.2612**	
	100	0.0059	**0.0041**	0.2182	**0.1953**	

Table 7: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for exponential distribution with $\theta = 1$ assuming $m = 5$

Scheme Number	n	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
9	10	0.0283	**0.0180**	0.5597	**0.5208**
	20	0.0259	**0.0170**	0.5053	**0.4699**
	30	0.0215	**0.0139**	0.4658	**0.4303**
	50	0.0164	**0.0101**	0.4143	**0.3791**
	100	0.0127	**0.0073**	0.3744	**0.3391**
10	10	0.0300	**0.0189**	0.5563	**0.5235**
	20	0.0232	**0.0144**	0.4737	**0.4436**
	30	0.0194	**0.0117**	0.4436	**0.4132**
	50	0.0154	**0.0090**	0.3993	**0.3695**
	100	0.0114	**0.0058**	0.3599	**0.3289**
11	10	0.0359	**0.0238**	0.6342	**0.5922**
	20	0.0254	**0.0171**	0.4983	**0.4576**
	30	0.0236	**0.0158**	0.4919	**0.4500**
	50	0.0191	**0.0126**	0.4443	**0.4039**
	100	0.0131	**0.0081**	0.3733	**0.3347**
12	10	0.0251	**0.0159**	0.4953	**0.4621**
	20	0.0212	**0.0137**	0.4481	**0.4146**
	30	0.0186	**0.0121**	0.4180	**0.3857**
	50	0.0133	**0.0088**	0.3463	**0.3189**
	100	0.0079	**0.0048**	0.2723	**0.2477**

Table 8: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for normal distribution with $\mu = 0, \sigma = 1$ assuming $m = 3$

Scheme Number	n	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
1	10	0.0845	**0.0520**	1.5360	**1.4058**
	20	0.0821	**0.0515**	1.5007	**1.3609**
	30	0.0796	**0.0499**	1.4879	**1.3401**
	50	0.0771	**0.0487**	1.4700	**1.3076**
	100	0.0754	**0.0472**	1.4700	**1.3060**
2	10	0.0867	**0.0551**	1.5636	**1.4262**
	20	0.0832	**0.0541**	1.5046	**1.3688**
	30	0.0815	**0.0535**	1.4869	**1.3474**
	50	0.0801	**0.0526**	1.4803	**1.3360**
	100	0.0762	**0.0498**	1.4576	**1.3037**
3	10	0.0800	**0.0486**	1.4779	**1.3399**
	20	0.0768	**0.0478**	1.4434	**1.2884**
	30	0.0755	**0.0471**	1.4392	**1.2781**
	50	0.0735	**0.0457**	1.4330	**1.2654**
	100	0.0717	**0.0443**	1.4324	**1.2590**
4	10	0.0825	**0.0508**	1.5122	**1.3729**
	20	0.0804	**0.0491**	1.4912	**1.3513**
	30	0.0786	**0.0481**	1.4801	**1.3315**
	50	0.0763	**0.0467**	1.4733	**1.3145**
	100	0.0742	**0.0454**	1.4634	**1.2965**

Table 9: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for normal distribution with $\mu = 0, \sigma = 1$ assuming $m = 4$

Scheme Number	n	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
5	10	0.0735	**0.0555**	1.4046	**1.3288**
	20	0.0638	**0.0482**	1.2529	**1.1836**
	30	0.0553	**0.0411**	1.1298	**1.0654**
	50	0.0484	**0.0352**	1.0534	**0.9925**
	100	0.0406	**0.0284**	0.9559	**0.9005**
6	10	0.0768	**0.0586**	1.4402	**1.3612**
	20	0.0665	**0.0507**	1.2799	**1.2066**
	30	0.0590	**0.0445**	1.1722	**1.1058**
	50	0.0486	**0.0358**	1.0230	**0.9669**
	100	0.0427	**0.0304**	0.9581	**0.9077**
7	10	0.0799	**0.0615**	1.4823	**1.4038**
	20	0.0677	**0.0521**	1.3037	**1.2296**
	30	0.0622	**0.0477**	1.2309	**1.1578**
	50	0.0568	**0.0431**	1.1638	**1.0907**
	100	0.0511	**0.0380**	1.1216	**1.0472**
8	10	0.0770	**0.0589**	1.4454	**1.3674**
	20	0.0682	**0.0525**	1.3116	**1.2354**
	30	0.0629	**0.0484**	1.2217	**1.1512**
	50	0.0533	**0.0402**	1.0890	**1.0263**
	100	0.0462	**0.0337**	1.0106	**0.9537**

Table 10: MSEs of the estimates of extropy $J(Y)$ and entropy $H(Y)$ for normal distribution with $\mu = 0, \sigma = 1$ assuming $m = 5$

Scheme Number	n	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
9	10	0.1021	**0.0810**	1.6604	**1.5913**
	20	0.0925	**0.0740**	1.5344	**1.4681**
	30	0.0848	**0.0678**	1.4609	**1.3933**
	50	0.0787	**0.0627**	1.4157	**1.3462**
	100	0.0716	**0.0565**	1.3670	**1.2956**
10	10	0.1002	**0.0788**	1.6234	**1.5584**
	20	0.0863	**0.0678**	1.4751	**1.4082**
	30	0.0806	**0.0632**	1.4172	**1.3496**
	50	0.0743	**0.0578**	1.3672	**1.2976**
	100	0.0686	**0.0528**	1.3462	**1.2732**
11	10	0.1047	**0.0836**	1.6928	**1.6201**
	20	0.0944	**0.0761**	1.5693	**1.4960**
	30	0.0871	**0.0704**	1.4807	**1.4087**
	50	0.0791	**0.0637**	1.3985	**1.3251**
	100	0.0690	**0.0549**	1.3175	**1.2431**
12	10	0.0990	**0.0779**	1.6263	**1.5586**
	20	0.0833	**0.0659**	1.4533	**1.3839**
	30	0.0773	**0.0611**	1.3866	**1.3168**
	50	0.0656	**0.0512**	1.2499	**1.1834**
	100	0.0554	**0.0424**	1.1382	**1.0757**

It can be seen from Tables 2-10 that the performance of the proposed estimates is affected by the distribution of the sample under study. For uniform distribution, the obtained simulation results are given in Tables 2-4. We observe that extropy estimates $\hat{J}_1$ and $\hat{J}_2$ perform satisfactorily, as shown in Tables 2 and 3. Moreover, $\hat{J}_1$ dominates $\hat{J}_2$ under all censoring schemes for all sample sizes. It is also obvious from Tables 2 and 3 that both entropy estimates $\hat{H}_1$ and $\hat{H}_2$ perform well. Moreover, $\hat{H}_1$ dominates $\hat{H}_2$ under all censoring schemes for all sample sizes. Table 4 shows that for scheme 9, $\hat{J}_1$ and $\hat{J}_2$ perform well and $\hat{J}_2$ dominates $\hat{J}_1$ for all sample sizes; while for scheme 10 we see that the MSEs of $\hat{J}_2$ are always smaller than those of $\hat{J}_1$ except when $n = 10$, where $\hat{J}_1$ and $\hat{J}_2$ are equal. As for scheme 11, we see that MSEs of $\hat{J}_2$ are always smaller than those of $\hat{J}_1$ except when $n = 10$. On the other hand, in scheme 12, we see that MSEs of$\hat{J}_1$ are always smaller than those of $\hat{J}_2$ except when $n = 100$. Furthermore, from Table 4, we observe that for schemes 9 and 10, estimates of $\hat{H}_1$ and $\hat{H}_2$ perform satisfactorily and we also see that MSEs of $\hat{H}_1$ are always smaller than those of $\hat{H}_2$ except when $n = 50$ and $n = 100$ and they are equal when $n = 30$. As for scheme 11, we see that MSEs of $\hat{H}_2$ are always smaller than those of $\hat{H}_1$ except when $n = 10$ and $n = 20$. On the other hand, in scheme 12 we see that MSEs of $\hat{H}_1$ are always smaller than those of $\hat{H}_2$ except when $n = 50$ and $n = 100$.

For exponential distribution, results obtained are displayed in Tables 5-7. It is clear from Tables 5, 6 and 7 that extropy estimates $\hat{J}_1$ and $\hat{J}_2$ perform satisfactorily where $\hat{J}_2$ dominates $\hat{J}_1$ under all censoring schemes for all sample sizes. Moreover, it is clear from Tables 9, 11 and 13, entropy estimates $\hat{H}_1$ and $\hat{H}_2$ perform well where $\hat{H}_2$ dominates $\hat{H}_1$ under all censoring schemes for all sample sizes.

For normal distribution, the obtained results are arranged in Tables 8-10. We observe that $\hat{J}_1$ and $\hat{J}_2$ perform satisfactorily, as shown in Tables 8, 9 and 10, where $\hat{J}_2$ dominates $\hat{J}_1$ under all censoring schemes for all sample sizes. It is also obvious from Tables 8, 9 and 10 that both $\hat{H}_1$ and $\hat{H}_2$ estimators perform well, where $\hat{H}_2$ dominates $\hat{H}_1$ under all censoring schemes for all sample sizes.

As expected, MSE decreases as the sample size n increases. It is worthwhile to point that the extropy and entropy estimates are affected by the sample size, censoring schemes and the type of distribution of data. For example, we notice that the MSEs generated from uniform distribution seems to outperform their parallel MSEs generated from Exponential distribution under all censoring schemes and for all sample sizes.

4.2. Real data analysis

In this subsection, we present an example to show the behavior of the proposed entropy and extropy estimates in real case. Here we consider the insulating fluid example from [Nelson, 1982, P.105].

Example : The following data, represents failure times (in minutes) for an insulating fluid between two electrodes, subject to a voltage of 34 kV.:

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

It is well known that the exponential distribution, $Exp(\theta)$ with pdf $f(x) = \theta \exp(-\theta x)$is appropriate fitting model for this dataset. The MLE of θ based on the complete sample is equal to $1/\bar{X}$ which equals 0.0696. Therefore, assuming $\theta = 0.0696$, the extropy and entropy of X in this case are obtained to be

$$J(X) = -\frac{\theta}{4} = -0.0174, \text{ and } H(X) = 1 - \log\ \theta = 3.6644,$$

respectively.

Next, we study the behavior of the proposed estimators based on the following progressive Type-I Interval censoring schemes that are presented in Table 11. In this study, we consider $m = 4$ and $w = 2$.

Table 11: Progressive censoring schemes used in this real data example

Censoring scheme No.	(T_1, T_2, T_3, T_4)	(q_1, q_2, q_3, q_4)
1	$1, 5, 15, 25$	$0.25, 0, 0.25, 1$
2	$2.5, 5, 10, 35$	$0.1, 0.2, 0.3, 1$
3	$1, 6, 18, 30$	$0.05, 0.05, 0.05, 1$
4	$0.5, 4, 10, 30$	$0, 0, 0, 1$
5	$0.9, 4, 12, 40$	$0.3, 0.2, 0.1, 1$

Continuing with the exploration of progressive Type-I Interval censoring under this lifetime model, the following censored data are observed according to the applied censoring scheme on the insulation data. The generated censored data are summarized in Table 12.

Table 12: The observed censored data from this real data example

Scheme No.	$\{k_1,k_2,k_3,k_4\}$	$\{R_1,R_2,R_3,R_4\}$
Scheme 1	$\{3,4,4,0\}$	$\{4,0,1,3\}$
Scheme 2	$\{4,4,3,2\}$	$\{2,2,1,1\}$
Scheme 3	$\{3,6,4,0\}$	$\{1,0,0,5\}$
Scheme 4	$\{1,5,7,1\}$	$\{0,0,0,5\}$
Scheme 5	$\{2,3,5,1\}$	$\{5,2,0,1\}$

Results of estimation for the extropy and entropy measures using the proposed estimates under the above censoring schemes are shown in Table 13.

Table 13: Extropy and entropy estimates for this real data example

Scheme Number	$\hat{J}_1$	$\hat{J}_2$	$\hat{H}_1$	$\hat{H}_2$
1	-0.0158	-0.0141	3.5088	3.5066
2	-0.0192	-0.0155	3.4825	3.5296
3	-0.0125	-0.0112	3.7624	3.7314
4	-0.0175	-0.0155	3.4866	3.4543
5	-0.0157	-0.0137	3.6187	3.6018

From Table 13, we can clearly see, for the real data set, that all estimates $\hat{J}_1$, $\hat{J}_2$, $\hat{H}_1$ and $\hat{H}_2$ provide close results to those obtained using the complete sample with slight differences. The proposed extropy estimates $\hat{J}_1$ and $\hat{J}_2$ perform satisfactorily when comparing their values to $J(X) = -0.0174$. On the other hand, the estimates $\hat{H}_1$ and $\hat{H}_2$ perform well under all suggested censoring schemes when comparing their values to $H(X) = 3.6644$. This confirms that these results are in agreement with what have been concluded from the simulation studies.

5. Conclusions

In this paper, we have considered the estimation problem of the extropy and entropy measures based on progressive Type-I interval censoring samples. Nonparametric-based methods involving moments approximation and linear approximation have been proposed for estimating the extropy and entropy measures. The performance of the proposed estimates have been studied via simulation studies and real data sets considering various censoring schemes and three probability distributions, namely uniform, exponential and normal distributions. It has been observed that the proposed estimates of the extropy and entropy measures are affected by the sample size, censoring schemes and the type of distribution

of data. The Monte Carlo simulations show that both estimates perform well in terms of the MSE. Yet, the estimates based on linear approximation $\hat{J}_2$ and $\hat{H}_2$ outperform the other estimate in the majority of studied cases.

Acknowledgements

The authors would like to thank the associate editor and anonymous referees for their constructive comments which improved the quality and presentation of our results.

References

Aggarwala, R., (2001). Progressive interval censoring: some mathematical results with applications to inference. *Commun Stat: Theory Methods*, 30, pp. 1921–1935.

Alotaibi, R., Rezk, H., Dey, S., and Okasha, H., (2021). Bayesian estimation for Dagum distribution based on progressive type I interval censoring. *PLOS ONE*, 16(6), DOI: 10.1371/journal.pone.0252556

Awad, A. M. and Alawneh, A., (1987). Application of entropy of a life time model. *IMA J. Math. Control Inf.*, 4, pp. 143–147.

Balakrishnan, N., Cramer, E., (2014). *The Art of Progressive Censoring: Applications to Reliability and Quality*. Boston.

Cohen, A. C., (1963). Progressively censored sample in life testing. *Technometrics*, 5, pp. 327–339.

Correa, J. C., (1995). A new estimator of entropy. *Communications in Statistics Theory and Methods*, 24, pp. 2439–2449.

Du Y., Guo Y. and Gui W., (2018). Statistical Inference for the Information Entropy of the Log-Logistic Distribution under Progressive Type-I Interval Censoring Schemes. *Symmetry*, 10, p. 445.

Hazeb, R., Raqab, M. and Bayoud, H., (2021a). Non-parametric estimation of the extropy and the entropy measures based on progressive type-II censored data with testing uniformity. *Journal of Statistical Computation and Simulation*, 91 (11), pp. 2178–2210.

Hazeb, R., Bayoud, H. and Raqab, M., (2021b). Kernel and CDF-Based Estimation of Extropy and Entropy from Progressively Type-II Censoring with Application for Goodness of Fit Problems. *Stochastic and Quality Control*, 36 (1), pp. 73–83.

Kittaneh, O. A., Khan, M. A., Akbar, M., and Bayoud, H. A. (2016). Average entropy: a new uncertainty measure with application to image segmentation. *The American Statistician*, 70 (1), pp. 18–24.

Lad, F., Sanfilippo, G. and Agro, G., (2015). Extropy: Complementary Dual of Entropy. *Statistical Science*, 30, pp. 40–58.

Lio, YL., Chen, D. G. and Tasi, T. R., (2011). Parameter estimations for generalized exponential distribution under progressive type-I interval censoring.*Comput. Stat. Data Anal.*, 54, pp. 1581–1591.

Nelson, W., (1982). *Applied Life Data Analysis*, Wiley, New York.

Ng, H. K. T., Wang, Z., (2009). Statistical estimation for the parameters of Weibulldistribution based on progressively Type-I interval censored sample.*J.Stat.Comput.Simulat*, 79(2), pp. 145–159.

Noughabi, H. A., Jarrahiferiz, J., (2019). On the estimation of extropy. *Journal of Nonparametric Statistics*, 31(1), pp. 88–99, DOI: 10.1080/10485252.2018.1533133.

Qiu, G., (2017). The Extropy of Order Statistics and Record values. *Statistics and Probability Letters*, 120, pp. 52–60.

Qiu, G. and Jia, K., (2018a). The Residual Extropy of Order statistics. *Statistics and Probability Letters*, 133, pp. 15–22.

Qiu, G. and Jia, K., (2018b). Extropy Estimators with Applications in Testing Uniformity. *Journal of Nonparametric Statistics*, 30(1), pp. 182–196, DOI: 10.1080/1048252.2017.1404063.

Rao, M., Chen, Y., Vemuri, B. C., and Wang, F., (2004). Cumulative residual entropy: a new measure of information. *IEEE transactions on Information Theory*, 50 (6), pp. 1220–1228.

Raqab, M. Z. and Qiu, G., (2019). On extropy properties of ranked set sampling. *Statistics*, 53(1), pp. 210–226, DOI: 10.1080/02331888.2018.1533963.

Renyi, A., (1961). On measures of entropy and information. *Stat. and Prob.*, 1, pp. 547–561.

Shannon, C. E., (1948). A mathematical theory of communications. *Bell System Tech. J.*, 27(3), pp. 379–423.

Singh, S, Tripathi, Y. M., (2016). Estimating the parameters of an inverse Weibull distribution under progressive type-I interval censoring. *Stat. Pap.*, 59, pp. 21–56.

Tsallis, C., (1988). Possible generalization of Boltzmann Gibbs statistics. *J. Stat. Phys.*, 52, pp. 470–487.

Vasicek, O., (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society B*, 38, pp. 54–59.

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 103–122, https://doi.org/10.59170/stattrans-2024-030
Received – 08.06.2022; accepted – 11.01.2024

Improved estimation of the mean through regressed exponential estimators based on sub-sampling non-respondents

R. R. Sinha[1], **Bharti**[2]

Abstract

The present study concerns the issue of estimating the population mean and presents novel and improved regressed exponential estimators using different parameters of an auxiliary character based on sub-sampling non-respondents. The bias and mean square error *(MSE)* of the proposed estimators for the most pragmatic simple random sampling without replacement (SRSWOR) scheme have been derived up to the first order of approximation (i.e. the expression containing errors up to the power of two so that the expectation comes only in terms of the mean, variance and covariance). The optimum value of the MSE of the estimators is found, along with the necessary conditions for optimising the *MSE*. The effectiveness of the suggested estimators, outperforming the existing ones in terms of their MSE, has been studied theoretically, while the empirical as well as the simulation studies have confirmed these findings.

Key words: population mean, bias, mean square error, auxiliary character.

1. Introduction

The history of optimal use of auxiliary information to increase the efficiency of the estimators has been established by a variety of research articles in surveys sampling [see Cochran (1940), Tripathi et al. (1994), Khare (2003)]. But practically in a factual scenario, auxiliary information is not only available in the form of a variable but also in the form of an attribute, such as less or more fertility of soil, high or low breed of animals, gender (male or female), tall or short height of person, etc. So, when the auxiliary information is available in the form of attributes, several authors have taken the advantage of point bi-serial correlation coefficient between the study character $'y'$

[1] Dr B R Ambedkar, NIT, Jalandhar, India. E-mail: raghawraman@gmail.com.
ORCID: https://orcid.org/0000-0001-6386-1973
[2] Dr B R Ambedkar, NIT, Jalandhar, India. E-mail: bhartikhanna_512@yahoo.com.
ORCID: https://orcid.org/0009-0009-5787-4298.

and the auxiliary attribute $'\varphi'$ and improvised conventional estimators for estimating the parameters which have been reviewed by Singh et al. (2019).

Recently, using information from auxiliary attributes, Zaman and Kadilar (2019) and Zaman (2020) proposed novel classes of exponential estimators, while Zaman and Kadilar (2021a, b) proposed two phase exponential ratio and product type estimators and class of exponential estimators for estimating population mean. The effectiveness of ratio-type estimators for estimating population has been further improved by Yadav and Zaman (2021) using some conventional and non-conventional parameters.

These days, researchers of various fields are facing problems to reduce non-random bias in the estimation of parameters due to incomplete information on units selected in the sample. One of the main reasons is that nowadays most of the surveys related to different issues of human beings are based on an internet-oriented online program in which respondents are reluctant to reply specially on critical or sensitive matters. In this way, non-response is a massive challenge which creates bias and reduces the exactitude of estimates of parameters. Hansen and Hurwitz (1946) were first to suggest an unbiased estimator by initiating a method of sub-sampling from non-respondents to estimate the population mean.

Following Hansen and Hurwitz's (1946) sub-sampling methodology of non-respondents with known and unknown population means of auxiliary character(s), Rao (1986, 1990), Khare and Srivastava (1993, 1995, 1997, 2000), Khare and Sinha (2009), Singh and Kumar (2009), and Sinha and Kumar (2011, 2014) have made contributions to the estimation of the population by suggesting conventional and alternative ratio, product, regression estimators, generalized and classes of estimators. Furthermore, Khare and Sinha (2002) attempted to estimate the ratio of two population means using an auxiliary character with an unknown population mean. Meanwhile, Sinha and Kumar (2013) and Sinha and Bharti (2021, 2022) suggested some improved estimators using an auxiliary attribute and non-conventional auxiliary parameters to estimate the population mean in the presence of non-response.

Now, we propose a new family of estimators of population mean when the non-response problem occurs – not only in the case of target variable but also in the case of auxiliary attributes expressed usually by relevant categorical variables. It is assumed that an additional feature expressed by a binary variable is investigated and that some part of respondents has not provided some or all data (i.e. that item non-response or unit non-response occur) concerning target or auxiliary variables. The suggested estimators combine a regression estimator with an exponential function of auxiliary information that has two optimizing constants for two distinct non-response scenarios. Their efficiency is verified using empirical data from 1981 Census in India and a simulation study based on some population data in the same country.

2. Preliminary Sample Selection and Literature Review

Consider a finite population of size N from which a simple random sample of size n is taken without replacement. In surveys of human populations, it happens frequently that n_1 of the units respond on the first try to the questions being asked, while the remaining $n_2\ (= n - n_1)$ units do not respond at all. Hansen and Hurwitz (1946) considered a double sampling strategy for estimating population mean consisting of the steps outlined when non-response occurs in the initial attempt. A simple random sample of size n is chosen, and the survey is mailed to the sample units. A subsample of size $n_\omega (= n_2\omega^{-1};\ \omega > 1)$ from the n_2 units that did not respond in the initial attempt is then contacted and information is obtained through personal interviews. For the purposes of this procedure, consider a population of size N that is split into two non-overlapping responding (N_1 units) and non-responding (N_2 units) groups with population means of $\bar{Y}_{(1)}$ and $\bar{Y}_{(2)}$ respectively. Although the proportional weights of the response $W_1 = N_1N^{-1}$ and the non-response $W_2 = N_2N^{-1}$ are not known, they can be estimated by $w_1 = n_1n^{-1}$ and $w_2 = n_2n^{-1}$, respectively. On the basis of readily available data for $(n_1 + n_\omega)$ units, Hansen and Hurwitz (1946) proposed an unbiased estimator for estimating the population mean $(\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2)$ that is given by

$$T_{HH} = \bar{y}^{\#} = w_1\bar{y}_1 + w_2\bar{y}_{(n_\omega)} \tag{1}$$

Its variance up to the first order of approximation $[O(n^{-1})]$ is given by

$$V(T_{HH}) = \bar{Y}^2\left[\pi C_y^2 + \pi^{\#}C_{y(2)}^2\right], \tag{2}$$

where $\pi = (n^{-1} - N^{-1})$, $\pi^{\#} = N_2(\omega - 1)(Nn)^{-1}$, $C_y^2\left(= \frac{S_y^2}{\bar{Y}^2}\right)$ and $C_{y(2)}^2\left(= \frac{S_{y(2)}^2}{\bar{Y}_{(2)}^2}\right)$ are the coefficients of variation while S_y^2 and $S_{y(2)}^2$ are the population mean squares of y for entire and non-responding parts of the population. $\bar{y}_1$ and $\bar{y}_{(n_\omega)}$ are sample means of the study variate depending upon n_1 and n_ω units respectively.

Suppose the population is dichotomous with respect to presence and absence of an attribute $'\varphi'$ which assumes only two values $'1'$ for possessing attribute and $'0'$ otherwise. Let the observations of study character and auxiliary attribute for $i^{th}(i = 1,2,3, \dots, N)$ population unit be denoted by y_i and φ_i.

Let the total number of units possessing the attribute $'\varphi'$ in the population and sample be $T_N = \sum_{i=1}^{N}\varphi_i$ and $T_n = \sum_i^n \varphi_i$ respectively. Let $P\left(= \frac{T_N}{N}\right)$ and $p\left(= \frac{T_n}{n}\right)$ be the proportion of units in the population and sample while $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{y} = \frac{1}{n}\sum_i^n y_i$ be the population mean and sample mean of study variable.

In this manuscript, the unit non-response case is considered, which is the phenomenon described by Bethlehem et al. (2011), in which the questionnaire remains empty for some elements in the sample. Therefore, under the assumption of population division

between responding and non-responding groups, let $\bar{P}_{(1)}$ and $\bar{P}_{(2)}$ be the population proportions of the units possessing the attribute for the responding and non-responding groups of the population respectively, even though they are unknown. If the $\bar{p}_1$ and $\bar{p}_{(n_\omega)}$ are the sample proportions of the units possessing φ for the n_1 and n_ω units respectively, then an unbiased estimator for estimating P is given by

$$\bar{p}^{\#} = w_1\bar{p}_1 + w_2\bar{p}_{(n_\omega)} \tag{3}$$

The variance of $\bar{p}^{\#}$ up to $[O(n^{-1})]$ is given by

$$V(\bar{p}^{\#}) = P^2\left[\pi C_p^2 + \pi^{\#} C_{p(2)}^2\right], \tag{4}$$

where $C_p^2\left(=\frac{S_\varphi^2}{P^2}\right)$ and $C_{p(2)}^2\left(=\frac{S_{\varphi(2)}^2}{P_{(2)}^2}\right)$ are the coefficients of variation while S_φ^2 and $S_{\varphi(2)}^2$ are the population mean squares of units possessing the attribute φ for entire and non-responding groups of the population.

Under the supposition of unit non-response, Rao (1986) and Khare and Srivastava (1995, 1997, 2000) envisaged ratio, product, and generalized estimators to estimate the mean of the study variable y using the auxiliary variable x. Adopting them, the ratio, product, and generalized estimators are suggested for estimating the population mean $\bar{Y}$ using the known population proportion (P) if non-response only pertains to the study character as follows:

$$T_{r1}^{\#} = \bar{y}^{\#}\frac{P}{p}, \qquad \text{[Ratio estimator]} \tag{5}$$

$$T_{p1}^{\#} = \bar{y}^{\#}\frac{p}{P}, \qquad \text{[Product estimator]} \tag{6}$$

and $$T_{g1}^{\#} = \bar{y}^{\#}\left(\frac{p}{P}\right)^{\gamma_1}, \qquad \text{[Generalized estimator]}, \tag{7}$$

where γ_1 is an optimizing constant for mean square error.

Furthermore, Riaz and Darda (2016) adopted a regression estimator to estimate the population mean using an auxiliary attribute under the non-response on study character with a known population proportion P, which is given as

$$T_{reg1}^{\#} = \bar{y}^{\#} + \beta_1(P - p) \qquad \text{[Regression estimator]}. \tag{8}$$

Under large sample approximation, the $MSEs$ of all the above estimators up to the order of n^{-1} are given by

$$MSE(T_{r1}^{\#}) = \bar{Y}^2\{\pi {C_p}^2 + \left(\pi C_y^2 + \pi^{\#} C_{y(2)}^2\right) - 2\pi\rho_{yp}C_yC_p\}, \tag{9}$$

$$MSE(T_{p1}^{\#}) = \bar{Y}^2\{\pi {C_p}^2 + \left(\pi C_y^2 + \pi^{\#} C_{y(2)}^2\right) + 2\pi\rho_{yp}C_yC_p\}, \tag{10}$$

$$\left[MSE(T_{g1}^{\#})\right]_{min} = V(\bar{y}^{\#}) - \pi\rho_{yp}^2 S_y^2 \text{ at } \gamma_{1_{opt}} = -\rho_{yp}\frac{C_y}{C_p} \tag{11}$$

and $$\left[MSE(T_{reg1}^{\#})\right]_{min} = V(\bar{y}^{\#}) - \pi\rho_{yp}^2 S_y^2 \tag{12}$$

where ρ_{yp} is the point bi-serial correlation coefficient between y and p.

Advocating the prior discussed notable contributions, the conventional ratio, product, generalized and regression estimators for estimating the population mean

with known proportion of auxiliary variable under unit non-response on study as well as auxiliary variates may respectively be adopted and define as

$$T_{r2}^{\#} = \bar{y}^{\#}\frac{P}{p^{\#}}, \tag{13}$$

$$T_{p2}^{\#} = \bar{y}^{\#}\frac{p^{\#}}{P}, \tag{14}$$

$$T_{g2}^{\#} = \bar{y}^{\#}\left(\frac{p^{\#}}{P}\right)^{\gamma_2}, \text{ where } \gamma_2 \text{ is an arbitrary constant,} \tag{15}$$

and $$T_{reg2}^{\#} = \bar{y}^{\#} + \beta_2(P - p^{\#}). \tag{16}$$

The $MSEs$ of the estimators $T_{r2}^{\#}$, $T_{p2}^{\#}$, $T_{reg2}^{\#}$ and $T_{g2}^{\#}$ up to the order of n^{-1} are given as

$$MSE(T_{r2}^{\#}) = \bar{Y}^2\left[\pi\left(C_y^2 + C_p^2 - 2\rho_{yp}C_yC_p\right) + \pi^{\#}\left(C_{y(2)}^2 + C_{p(2)}^2 - 2\rho_{yp(2)}C_{y(2)}C_{p(2)}\right)\right], \tag{17}$$

$$MSE\left(T_{p2}^{\#}\right) = \bar{Y}^2\left[\pi\left(C_y^2 + C_p^2 + 2\rho_{yp}C_yC_p\right) + \pi^{\#}\left(C_{y(2)}^2 + C_{p(2)}^2 + 2\rho_{yp(2)}C_{y(2)}C_{p(2)}\right)\right], \tag{18}$$

$$\left[MSE(T_{g2}^{\#})\right]_{min} = V(\bar{y}^{\#}) - \frac{\left\{\pi\rho_{yp}S_yS_p+\pi^{\#}\rho_{yp(2)}S_{y(2)}S_{p(2)}\right\}^2}{\left\{\pi S_p^2+\pi^{\#}S_{p(2)}^2\right\}} \tag{19}$$

$$\text{at } (\gamma_2)_{opt} = -\frac{\left\{\pi\rho_{yp}C_yC_p+\pi^{\#}\rho_{yp(2)}C_{y(2)}C_{p(2)}\right\}}{\left\{\pi C_p^2+\pi^{\#}C_{p(2)}^2\right\}}$$

and $$\left[MSE(T_{reg2}^{\#})\right]_{min} = V(\bar{y}^{\#}) - \frac{\left\{\pi\rho_{yp}S_yS_p+\pi^{\#}\rho_{yp(2)}S_{y(2)}S_{p(2)}\right\}^2}{\left\{\pi S_p^2+\pi^{\#}S_{p(2)}^2\right\}}. \tag{20}$$

Here, $\rho_{yp(2)}$ is the point bi-serial correlation coefficient between y and p for the non-responding part of the population.

In this sequence, exponential estimators for estimation of population mean using auxiliary attribute have been proposed by Kumar and Kumar (2019) in both the cases of non-response discussed earlier. Exponential ratio, exponential product and generalized estimators for the case when non-response occurs only on study variable are defined as follows:

$$T_{KK1(r)}^{\#} = \bar{y}^{\#}\exp\left(\frac{P-p}{P+p}\right), \tag{21}$$

$$T_{KK1(p)}^{\#} = \bar{y}^{\#}\exp\left(\frac{p-P}{p+P}\right) \tag{22}$$

and $$T_{KK1(g)}^{\#} = \bar{y}^{\#}\exp\left(\alpha_1\frac{P-p}{P+p}\right). \tag{23}$$

The mean square errors of these exponential estimators up to $[O(n^{-1})]$ are derived as

$$MSE\left(T_{KK1(r)}^{\#}\right) = V(\bar{y}^{\#}) + \bar{Y}^2\pi\left\{\frac{C_p^2}{4} - \rho_{yp}C_yC_p\right\} \tag{24}$$

$$MSE\left(T_{KK1(p)}^{\#}\right) = V(\bar{y}^{\#}) + \bar{Y}^2\pi\left\{\frac{C_p^2}{4} + \rho_{yp}C_yC_p\right\} \tag{25}$$

and $$\left[MSE(T_{KK1(g)}^{\#})\right]_{min} = V(\bar{y}^{\#}) - \pi\rho_{yp}^2S_y^2\,. \tag{26}$$

Moreover, Kumar and Kumar (2019) suggested the ratio, product, and generalized exponential estimators, which are provided below along with their mean square errors up to $[O(n^{-1})]$ in the case of non-response on both the study variable and the auxiliary attribute:

$$T_{KK2(r)}^{\#} = \bar{y}^{\#}\exp\left(\frac{P-p^{\#}}{P+p^{\#}}\right), \tag{27}$$

$$T_{KK2(p)}^{\#} = \bar{y}^{\#}\exp\left(\frac{p^{\#}-P}{p^{\#}+P}\right) \tag{28}$$

$$T_{KK2(g)}^{\#} = \bar{y}^{\#}\exp\left(\alpha_1\frac{P-p^{\#}}{P+p^{\#}}\right), \tag{29}$$

$$MSE\left(T_{KK2(r)}^{\#}\right) = V(\bar{y}^{\#}) + \bar{Y}^2\left[\pi\left\{\frac{C_p^2}{4} - \rho_{yp}C_yC_p\right\} + \pi^{\#}\left\{\frac{C_{p(2)}^2}{4} - \rho_{yp(2)}C_{y(2)}C_{p(2)}\right\}\right], \tag{30}$$

$$MSE\left(T_{KK2(p)}^{\#}\right) = V(\bar{y}^{\#}) + \bar{Y}^2\left[\pi\left\{\frac{C_p^2}{4} + \rho_{yp}C_yC_p\right\} + \pi^{\#}\left\{\frac{C_{p(2)}^2}{4} + \rho_{yp(2)}C_{y(2)}C_{p(2)}\right\}\right] \tag{31}$$

and $$\left[MSE(T_{KK2(g)}^{\#})\right]_{min} = V(\bar{y}^{\#}) - \frac{\left\{\pi\rho_{yp}S_yS_p+\pi^{\#}\rho_{yp(2)}S_{y(2)}S_{p(2)}\right\}^2}{\left\{\pi S_p^2+\pi^{\#}S_{p(2)}^2\right\}}. \tag{32}$$

The following conclusions have been drawn when comparing the efficacy of the aforementioned distinct estimators in terms of their MSEs:

(i) from (11), (12) and (26)

$$\left[MSE(T_{KK1(g)}^{\#})\right]_{min} = \left[MSE(T_{reg1}^{\#})\right]_{min} = \left[MSE(T_{g1}^{\#})\right]_{min} \tag{33}$$

and (ii) from (19), (20) and (32)

$$\left[MSE(T_{KK2(g)}^{\#})\right]_{min} = \left[MSE(T_{reg2}^{\#})\right]_{min} = \left[MSE(T_{g2}^{\#})\right]_{min} \tag{34}$$

3. Proposed Estimators

Influenced by the methodology of Koyuncu (2012) regression-cum-ratio class estimator and Singh and Solanki (2012) generalized class of estimator, novel ratio and product type improved regressed exponential estimators to estimate the population mean using known proportion of the auxiliary variable for two different cases are proposed as follows:

Case I: Unit non-response observed only on study variable - the proposed estimators for this circumstance are as follows:

$$T_{1prop}^{r} = \mathcal{A}_1\bar{y}^{\#} + \mathcal{B}_1(P-p)\exp\left[\frac{(\kappa P-\mathcal{L})-(\kappa p-\mathcal{L})}{(\kappa P-\mathcal{L})+(\kappa p-\mathcal{L})}\right] \quad \text{[Ratio type]} \tag{35}$$

and $$T_{1prop}^{p} = \mathcal{A}_2\bar{y}^{\#} + \mathcal{B}_2(p-P)\exp\left[\frac{(\kappa p-\mathcal{L})-(\kappa P-\mathcal{L})}{(\kappa p-\mathcal{L})+(\kappa P-\mathcal{L})}\right] \quad \text{[Product type]} \tag{36}$$

where κ and $\mathcal{L}$ are known constants and $\mathcal{A}_1, \mathcal{A}_2, \mathcal{B}_1, \mathcal{B}_2$ are the arbitrary constants.

Furthermore, in accordance with Singh and Taylor (2003), Kadilar and Cingi (2004), Singh et al. (2019), certain members of $T_{1_{prop}}^{r}$ and $T_{1_{prop}}^{p}$ are suggested by giving specific values to κ and $\mathcal{L}$, as shown below.

κ, $\mathcal{L}$	**Ratio Type Estimators**	**Product Type Estimators**
$\kappa = 1$ $\mathcal{L} = \rho$	$T_{1_{prop}}^{r(1)} = \mathcal{A}_1\bar{y}^{\#} + \mathcal{B}_1(P-p)\exp\left\{\frac{(P-p)}{P+p-2\rho}\right\}$	$T_{1_{prop}}^{p(1)} = \mathcal{A}_2\bar{y}^{\#} + \mathcal{B}_2(p-P)\exp\left\{\frac{(p-P)}{p+P-2\rho}\right\}$
$\kappa = C_p$ $\mathcal{L} = \rho$	$T_{1_{prop}}^{r(2)} = \mathcal{A}_1\bar{y}^{\#} + \mathcal{B}_1(P-p)\exp\left\{\frac{C_p(P-p)}{P+p-2\rho}\right\}$	$T_{1_{prop}}^{p(2)} = \mathcal{A}_2\bar{y}^{\#} + \mathcal{B}_2(p-P)\exp\left\{\frac{C_p(p-P)}{p+P-2\rho}\right\}$
$\kappa = \beta_1$ $\mathcal{L} = \rho$	$T_{1_{prop}}^{r(3)} = \mathcal{A}_1\bar{y}^{\#} + \mathcal{B}_1(P-p)\exp\left\{\frac{\beta_1(P-p)}{P+p-2\rho}\right\}$	$T_{1_{prop}}^{p(3)} = \mathcal{A}_2\bar{y}^{\#} + \mathcal{B}_2(p-P)\exp\left\{\frac{\beta_1(p-P)}{p+P-2\rho}\right\}$
$\kappa = \beta_2$ $\mathcal{L} = C_p$	$T_{1_{prop}}^{r(4)} = \mathcal{A}_1\bar{y}^{\#} + \mathcal{B}_1(P-p)\exp\left\{\frac{\beta_2(P-p)}{P+p-2C_p}\right\}$	$T_{1_{prop}}^{p(4)} = \mathcal{A}_2\bar{y}^{\#} + \mathcal{B}_2(p-P)\exp\left\{\frac{\beta_2(p-P)}{p+P-2C_p}\right\}$
$\kappa = \beta_2$ $\mathcal{L} = \rho$	$T_{1_{prop}}^{r(5)} = \mathcal{A}_1\bar{y}^{\#} + \mathcal{B}_1(P-p)\exp\left\{\frac{\beta_2(P-p)}{P+p-2\rho}\right\}$	$T_{1_{prop}}^{p(5)} = \mathcal{A}_2\bar{y}^{\#} + \mathcal{B}_2(p-P)\exp\left\{\frac{\beta_2(p-P)}{p+P-2\rho}\right\}$
$\kappa = \beta_2$ $\mathcal{L} = \beta_1$	$T_{1_{prop}}^{r(6)} = \mathcal{A}_1\bar{y}^{\#} + \mathcal{B}_1(P-p)\exp\left\{\frac{\beta_2(P-p)}{P+p-2\beta_1}\right\}$	$T_{1_{prop}}^{p(6)} = \mathcal{A}_2\bar{y}^{\#} + \mathcal{B}_2(p-P)\exp\left\{\frac{\beta_2(p-P)}{p+P-2\beta_1}\right\}$

The following approximations under large sample have been assumed to calculate $Bias$ and MSE of the proposed estimators:

$$\frac{\bar{y}^{\#}-\bar{Y}}{\bar{Y}} = \varepsilon_0, \qquad \frac{p-P}{P} = \varepsilon_2, \qquad \text{such that } E(\varepsilon_0) = E(\varepsilon_2) = 0$$

and $\quad E(\varepsilon_0^2) = \pi C_y^2 + \pi^{\#} C_{y(2)}^2, \;\; E(\varepsilon_2^2) = \pi C_p^2, \qquad E(\varepsilon_0\varepsilon_2) = \pi\rho_{yp}C_yC_p.$

Using these approximations, the estimators given in (35) and (36) are reduced to

$$T_{1_{prop}}^{r} = \mathcal{A}_1\bar{Y}(1+\varepsilon_0) - \mathcal{B}_1P(\varepsilon_2 - \theta\varepsilon_2^2) \tag{37}$$

and $\quad T_{1_{prop}}^{p} = \mathcal{A}_2\bar{Y}(1+\varepsilon_0) + \mathcal{B}_2P(\varepsilon_2 + \theta\varepsilon_2^2)\,,$ (38)

where $\theta = \frac{\kappa P}{2(\kappa P - \mathcal{L})}$.

Taking expectation on both sides of (37) and (38) and subtracting $\bar{Y}$ from them, the expressions of $Bias$ of $T_{1_{prop}}^{r}$and $T_{1_{prop}}^{p}$ up to the first order of approximation are as follows:

$$Bias\left(T_{1_{prop}}^{r}\right) = (\mathcal{A}_1 - 1)\bar{Y} + \mathcal{B}_1P\theta v_p \tag{39}$$

and $\quad Bias\left(T_{1_{prop}}^{p}\right) = (\mathcal{A}_2 - 1)\bar{Y} + \mathcal{B}_2P\theta v_p\,,$ (40)

where $v_p = \pi C_p^2$.

The MSE of $T_{1_{prop}}^{r}$ and $T_{1_{prop}}^{p}$ are calculated up to the $[O(n^{-1})]$ as

$$MSE\left(T_{1_{prop}}^{r}\right) = E[\{\mathcal{A}_1\bar{Y}(1+\varepsilon_0) - \mathcal{B}_1P(\varepsilon_2 - \theta\varepsilon_2^2)\} - \bar{Y}]^2$$

and $\quad MSE\left(T_{1_{prop}}^{p}\right) = E[\{\mathcal{A}_2\bar{Y}(1+\varepsilon_0) + \mathcal{B}_2P(\varepsilon_2 + \theta\varepsilon_2^2)\} - \bar{Y}]^2.$

After simplifying up to the first order of approximation, the expressions of MSE of $T_{1_{prop}}^{r}$ and $T_{1_{prop}}^{p}$ are as follows:

$$MSE\left(T_{1_{prop}}^{r}\right) = (\mathcal{A}_1 - 1)^2\bar{Y}^2 + \mathcal{A}_1^2\bar{Y}^2V_y + \mathcal{B}_1^2P^2v_p + 2(\mathcal{A}_1 - 1)\mathcal{B}_1P\bar{Y}\theta v_p - 2\mathcal{A}_1\mathcal{B}_1P\bar{Y}c_{yp}, \tag{41}$$

and

$$MSE\left(T_{1_{prop}}^{p}\right) = (\mathcal{A}_2 - 1)^2\bar{Y}^2 + \mathcal{A}_2^2\bar{Y}^2V_y + \mathcal{B}_2^2P^2v_p + 2(\mathcal{A}_2 - 1)\mathcal{B}_2P\bar{Y}\theta v_p + 2\mathcal{A}_2\mathcal{B}_2P\bar{Y}c_{yp}, \tag{42}$$

where $V_y = \pi C_y^2 + \pi^{\#}C_{y(2)}^2$, $\quad c_{yp} = \pi\rho_{yp}C_yC_p$.

To obtain the optimum $MSEs$ of $T_{1_{prop}}^{r}$ and $T_{1_{prop}}^{p}$, partially differentiating (41) with respect to $(\mathcal{A}_1, \mathcal{B}_1)$ and (42) with respect to $(\mathcal{A}_2, \mathcal{B}_2)$ and equating them to zero, the optimum values of $\mathcal{A}_i, \mathcal{B}_i; i = 1, 2$ are

$$\mathcal{A}_{1(o)} = \frac{v_p - \theta^2v_p^2 + \theta v_pc_{yp}}{v_p + v_pV_y - \theta^2v_p^2 + 2\theta v_pc_{yp} - c_{yp}^2},$$

$$\mathcal{B}_{1(o)} = \frac{\bar{Y}(c_{yp} + \theta v_pV_y)}{P(v_p + v_pV_y - \theta^2v_p^2 + 2\theta v_pc_{yp} - c_{yp}^2)},$$

$$\mathcal{A}_{2(o)} = \frac{v_p - \theta^2v_p^2 - \theta v_pc_{yp}}{v_p + v_pV_y - \theta^2v_p^2 - 2\theta v_pc_{yp} - c_{yp}^2}$$

and $\quad \mathcal{B}_{2(o)} = \frac{\bar{Y}(-c_{yp} + \theta v_pV_y)}{P(v_p + v_pV_y - \theta^2v_p^2 - 2\theta v_pc_{yp} - c_{yp}^2)}.$

Substituting the values of $\mathcal{A}_{1(o)}$ and $\mathcal{B}_{1(0)}$ in (41) and $\mathcal{A}_{2(o)}$ and $\mathcal{B}_{2(o)}$ in (42), we get the optimum value of MSE of the proposed estimators as

$$\left[MSE\left(T_{1_{prop}}^{r}\right)\right]_{min} = \frac{\bar{Y}^2\{v_pV_y - c_{yp}^2 - \theta^2v_p^2V_y\}}{v_p + v_pV_y - \theta^2v_p^2 + 2\theta v_pc_{yp} - c_{yp}^2}, \tag{43}$$

and

$$\left[MSE\left(T_{1_{prop}}^{p}\right)\right]_{min} = \frac{\bar{Y}^2\{v_pV_y - c_{yp}^2 - \theta^2v_p^2V_y\}}{v_p + v_pV_y - \theta^2v_p^2 - 2\theta v_pc_{yp} - c_{yp}^2}. \tag{44}$$

Case II: Unit non-response observed on both study and auxiliary variables - the proposed estimators for this occurrence are as follows:

$$T_{2_{prop}}^{r} = \mathcal{A}_3\bar{y}^{\#} + \mathcal{B}_3(P - p^{\#})\exp\left[\frac{(\kappa P - \mathcal{L}) - (\kappa p^{\#} - \mathcal{L})}{(\kappa P - \mathcal{L}) + (\kappa p^{\#} - \mathcal{L})}\right] \quad \text{[Ratio type]} \tag{45}$$

and

$$T_{2_{prop}}^{p} = \mathcal{A}_4\bar{y}^{\#} + \mathcal{B}_4(p^{\#} - P)\exp\left[\frac{(\kappa p^{\#} - \mathcal{L}) - (\kappa P - \mathcal{L})}{(\kappa p^{\#} - \mathcal{L}) + (\kappa P - \mathcal{L})}\right] \quad \text{[Product type]} \tag{46}$$

where κ and $\mathcal{L}$ are known constants and $\mathcal{A}_3$, $\mathcal{A}_4$, $\mathcal{B}_3$, $\mathcal{B}_4$ are the arbitrary constants.

Proceeding in the same manner as for case I, different members of $T_{2_{prop}}^{r}$ and $T_{2_{prop}}^{p}$ have been suggested by assigning different values to κ and $\mathcal{L}$ as

κ, $\mathcal{L}$	**Ratio type Estimators**	**Product type Estimators**
$\kappa = 1$ $\mathcal{L} = \rho$	$T_{2_{prop}}^{r(1)} = \mathcal{A}_3\bar{y}^{\#} + \mathcal{B}_3(P - p^{\#})\exp\left\{\frac{(P-p^{\#})}{P+p^{\#}-2\rho}\right\}$	$T_{2_{prop}}^{p(1)} = \mathcal{A}_4\bar{y}^{\#} + \mathcal{B}_4(p^{\#} - P)\exp\left\{\frac{(p^{\#}-P)}{p^{\#}+P-2\rho}\right\}$
$\kappa = C_p$ $\mathcal{L} = \rho$	$T_{2_{prop}}^{r(2)} = \mathcal{A}_3\bar{y}^{\#} + \mathcal{B}_3(P - p^{\#})\exp\left\{\frac{C_p(P-p^{\#})}{C_p(P+p^{\#})-2\rho}\right\}$	$T_{2_{prop}}^{p(2)} = \mathcal{A}_4\bar{y}^{\#} + \mathcal{B}_4(p^{\#} - P)\exp\left\{\frac{C_p(p^{\#}-P)}{C_p(p^{\#}+P)-2\rho}\right\}$
$\kappa = \beta_1$ $\mathcal{L} = \rho$	$T_{2_{prop}}^{r(3)} = \mathcal{A}_3\bar{y}^{\#} + \mathcal{B}_3(P - p^{\#})\exp\left\{\frac{\beta_1(P-p^{\#})}{\beta_1(P+p^{\#})-2\rho}\right\}$	$T_{2_{prop}}^{p(3)} = \mathcal{A}_4\bar{y}^{\#} + \mathcal{B}_4(p^{\#} - P)\exp\left\{\frac{\beta_1(p^{\#}-P)}{\beta_1(p^{\#}+P)-2\rho}\right\}$
$\kappa = \beta_2$ $\mathcal{L} = C_p$	$T_{2_{prop}}^{r(4)} = \mathcal{A}_3\bar{y}^{\#} + \mathcal{B}_3(P - p^{\#})\exp\left\{\frac{\beta_2(P-p^{\#})}{\beta_2(P+p^{\#})-2C_p}\right\}$	$T_{2_{prop}}^{p(4)} = \mathcal{A}_4\bar{y}^{\#} + \mathcal{B}_4(p^{\#} - P)\exp\left\{\frac{\beta_2(p^{\#}-P)}{\beta_2(p^{\#}+P)-2C_p}\right\}$
$\kappa = \beta_2$ $\mathcal{L} = \rho$	$T_{2_{prop}}^{r(5)} = \mathcal{A}_3\bar{y}^{\#} + \mathcal{B}_3(P - p^{\#})\exp\left\{\frac{\beta_2(P-p^{\#})}{\beta_2(P+p^{\#})-2\rho}\right\}$	$T_{2_{prop}}^{p(5)} = \mathcal{A}_4\bar{y}^{\#} + \mathcal{B}_4(p^{\#} - P)\exp\left\{\frac{\beta_2(p^{\#}-P)}{\beta_2(p^{\#}+P)-2\rho}\right\}$
$\kappa = \beta_2$ $\mathcal{L} = \beta_1$	$T_{2_{prop}}^{r(6)} = \mathcal{A}_3\bar{y}^{\#} + \mathcal{B}_3(P - p^{\#})\exp\left\{\frac{\beta_2(P-p^{\#})}{\beta_2(P+p^{\#})-2\beta_1}\right\}$	$T_{2_{prop}}^{p(6)} = \mathcal{A}_4\bar{y}^{\#} + \mathcal{B}_4(p^{\#} - P)\exp\left\{\frac{\beta_2(p^{\#}-P)}{\beta_2(p^{\#}+P)-2\beta_1}\right\}$

In continuation to the approximations assumed in Case I, another large sample approximation for proportion of auxiliary variable is considered as

$$\frac{p^{\#}-P}{P} = \varepsilon_1 \text{ such that } E(\varepsilon_1) = 0,$$
$$E(\varepsilon_1^2) = \pi C_p^2 + \pi^{\#} C_{p(2)}^2 \quad \text{and} \quad E(\varepsilon_0\varepsilon_1) = \pi\rho_{yp}C_yC_p + \pi^{\#}\rho_{yp(2)}C_{y(2)}C_{p(2)}.$$

Now, the estimators given in (45) and (46) can be reduced in terms of ε_0 and ε_1 as

$$T_{2_{prop}}^{r} = \mathcal{A}_3\bar{Y}(1+\varepsilon_0) - \mathcal{B}_3P(\varepsilon_1 - \theta\varepsilon_1^2) \tag{47}$$

and $$T_{2_{prop}}^{p} = \mathcal{A}_4\bar{Y}(1+\varepsilon_0) + \mathcal{B}_4P(\varepsilon_1 + \theta\varepsilon_1^2) \tag{48}$$

The $bias$ and MSE of the estimators $T_{2_{prop}}^{r}$ and $T_{2_{prop}}^{p}$ up to $O(n^{-1})$ can be given as follows:

$$Bias\left(T_{2_{prop}}^{r}\right) = (\mathcal{A}_3 - 1)\bar{Y} + \mathcal{B}_3P\theta V_p, \tag{49}$$

$$Bias\left(T_{2_{prop}}^{p}\right) = (\mathcal{A}_4 - 1)\bar{Y} + \mathcal{B}_4P\theta V_p\,, \tag{50}$$

$$MSE\left(T_{2_{prop}}^{r}\right) = (\mathcal{A}_3 - 1)^2\bar{Y}^2 + \mathcal{A}_3^2\bar{Y}^2V_y + \mathcal{B}_3^2P^2V_p + 2(\mathcal{A}_3 - 1)\mathcal{B}_3P\bar{Y}\theta V_p - 2\mathcal{A}_3\mathcal{B}_3P\bar{Y}C_{yp} \tag{51}$$

and $$MSE\left(T_{2_{prop}}^{p}\right) = (\mathcal{A}_4 - 1)^2\bar{Y}^2 + \mathcal{A}_4^2\bar{Y}^2V_y + \mathcal{B}_4^2P^2V_p + 2(\mathcal{A}_4 - 1)\mathcal{B}_4P\bar{Y}\theta V_p + 2\mathcal{A}_4\mathcal{B}_4P\bar{Y}C_{yp}. \tag{52}$$

where $V_p = \pi C_p^2 + \pi^{\#} C_{p(2)}^2$, $V_y = \pi C_y^2 + \pi^{\#} C_{y(2)}^2$,
$C_{yp} = \pi\rho_{yp}C_yC_p + \pi^{\#}\rho_{yp(2)}C_{y(2)}C_{p(2)}$.

The optimum values of $\mathcal{A}_i$, $\mathcal{B}_i$; $i = 3, 4$ to optimize the MSE of $T_{2_{prop}}^{r}$ and $T_{2_{prop}}^{p}$, can be obtained by partially differentiating (51) with respect to $(\mathcal{A}_3, \mathcal{B}_3)$ and (52) with respect to $(\mathcal{A}_4, \mathcal{B}_4)$, and equating them to zero we get

$$\mathcal{A}_{3(o)} = \frac{V_p - \theta^2 V_p^2 + \theta V_p C_{yp}}{V_p + V_p V_y - \theta^2 V_p^2 + 2\theta V_p C_{yp} - C_{yp}^2},$$

$$\mathcal{B}_{3(o)} = \frac{\bar{Y}(C_{yp} + \theta V_p V_y)}{P(V_p + V_p V_y - \theta^2 V_p^2 + 2\theta V_p C_{yp} - C_{yp}^2)},$$

$$\mathcal{A}_{4(o)} = \frac{V_p - \theta^2 V_p^2 - \theta V_p C_{yp}}{V_p + V_p V_y - \theta^2 V_p^2 - 2\theta V_p C_{yp} - C_{yp}^2}$$

and $$\mathcal{B}_{4(o)} = \frac{\bar{Y}(-C_{yp} + \theta V_p V_y)}{P(V_p + V_p V_y - \theta^2 V_p^2 - 2\theta V_p C_{yp} - C_{yp}^2)}.$$

Putting the values of $\mathcal{A}_{3(o)}$ and $\mathcal{B}_{3(o)}$ in (51) and $\mathcal{A}_{4(o)}$ and $\mathcal{B}_{4(o)}$ in (52), we get the optimum value of MSEs of the proposed estimators as

$$\left[MSE\left(T_{2_{prop}}^{r}\right)\right]_{min} = \frac{\bar{Y}^2\{V_p V_y - C_{yp}^2 - \theta^2 V_p^2 V_y\}}{V_p + V_p V_y - \theta^2 V_p^2 + 2\theta V_p C_{yp} - C_{yp}^2} \tag{53}$$

and $$\left[MSE\left(T_{2_{prop}}^{p}\right)\right]_{min} = \frac{\bar{Y}^2\{V_p V_y - C_{yp}^2 - \theta^2 V_p^2 V_y\}}{V_p + V_p V_y - \theta^2 V_p^2 - 2\theta V_p C_{yp} - C_{yp}^2}. \tag{54}$$

It would be remarkable to mention here that the optimum values of the constants $\mathcal{A}_{i(o)}$, $\mathcal{B}_i(o)$; $i = 1,2,3\ \&\ 4$ involved in optimizing the MSE of the suggested estimators depend upon unknown population parameters like v_p, c_{yp}, V_y, V_p and C_{yp}, which may be practically obtained from the supposition value based on prior information accessible from past data/pilot survey or replaced with their estimated values [for instance see Reddy (1978) and Srivastava and Jhajj (1983)].

4. Efficiency Comparisons

To show the efficiency of the proposed estimators with respect to the relevant estimators, mathematical conditions are derived by comparing their mean square errors, which are as follows:

(i) From (9), (10), (11), (12) and (2)

$MSE(T_{r1}^{\#}) \leq V(T_{HH})$ if $\quad v_p \leq 2c_{yp}$, where $v_p = \pi C_p^2$ and $c_{yp} = \pi \rho_{yp} C_y C_p$

$MSE(T_{p1}^{\#}) \leq V(T_{HH})$ if $\quad v_p \geq -2c_{yp}$

$\left[MSE(T_{g1}^{\#})\right]_{min} = \left[MSE(T_{reg1}^{\#})\right]_{min} \leq V(T_{HH})$ if $\quad c_{yp}^2 \geq 0$, $\quad$ always true.

(ii) From (9), (10) and (12)

$MSE\left(T_{r1}^{\#}\right) - \left[MSE(T_{reg1}^{\#})\right]_{min} = \left(v_p - c_{yp}\right)^2 \geq 0$, $\quad$ always true.

$MSE\left(T_{p1}^{\#}\right) - \left[MSE(T_{reg1}^{\#})\right]_{min} = \left(v_p + c_{yp}\right)^2 \geq 0$, $\quad$ always true.

Accordingly, using the aforementioned findings with (33), we have

$$\left[MSE(T^{\#}_{KK1(g)})\right]_{min} = \left[MSE(T^{\#}_{g1})\right]_{min} = \left[MSE(T^{\#}_{reg1})\right]_{min} \leq MSE(T^{\#}_{r1})$$

$$\left[MSE(T^{\#}_{KK1(g)})\right]_{min} = \left[MSE(T^{\#}_{g1})\right]_{min} = \left[MSE(T^{\#}_{reg1})\right]_{min} \leq MSE(T^{\#}_{p1}).$$

(iii) Proceeding as in (ii), we have the next two comparisons

From (24) and (33)

$$\left[MSE(T^{\#}_{KK1(g)})\right]_{min} = \left[MSE(T^{\#}_{g1})\right]_{min} = \left[MSE(T^{\#}_{reg1})\right]_{min} \leq MSE\left(T^{\#}_{KK1(r)}\right)$$

as $MSE\left(T^{\#}_{KK1(r)}\right) - \left[MSE\left(T^{\#}_{reg1}\right)\right]_{min} = \left(v_p - 2c_{yp}\right)^2 \geq 0$, always true.

And from (25) and (33)

$$\left[MSE(T^{\#}_{KK1(g)})\right]_{min} = \left[MSE(T^{\#}_{g1})\right]_{min} = \left[MSE(T^{\#}_{reg1})\right]_{min} \leq MSE\left(T^{\#}_{KK1(p)}\right)$$

as $MSE\left(T^{\#}_{KK1(p)}\right) - \left[MSE\left(T^{\#}_{reg1}\right)\right]_{min} = \left(v_p + 2c_{yp}\right)^2 \geq 0$, always true.

(iv) From (17), (18), (19), (20) and (2)

$MSE(T^{\#}_{r2}) \leq V(T_{HH})$ if $V_p \leq 2C_{yp}$.

$MSE(T^{\#}_{p2}) \leq V(T_{HH})$ if $V_p \geq -2C_{yp}$.

$\left[MSE\left(T^{\#}_{g2}\right)\right]_{min} = \left[MSE\left(T^{\#}_{reg2}\right)\right]_{min} \leq V(T_{HH})$ if $C^2_{yp} \geq 0$, always true.

Following a similar path as the comparison in (iii) and (iv), we arrive at the results in (v) and (vi).

(v) From (17) and (34)

$$\left[MSE(T^{\#}_{KK2(g)})\right]_{min} = \left[MSE(T^{\#}_{g2})\right]_{min} = \left[MSE(T^{\#}_{reg2})\right]_{min} \leq MSE(T^{\#}_{r2})$$

as $MSE\left(T^{\#}_{r2}\right) - \left[MSE\left(T^{\#}_{reg2}\right)\right]_{min} = \left(V_p - C^2_{yp}\right)^2 \geq 0$, always true.

And, from (18) and (34)

$$\left[MSE(T^{\#}_{KK2(g)})\right]_{min} = \left[MSE(T^{\#}_{g2})\right]_{min} = \left[MSE(T^{\#}_{reg2})\right]_{min} \leq MSE(T^{\#}_{p2})$$

as $MSE\left(T^{\#}_{p2}\right) - \left[MSE\left(T^{\#}_{reg2}\right)\right]_{min} = \left(V_p + C^2_{yp}\right)^2 \geq 0$, always true.

(vi) From (30) and (34)

$$\left[MSE(T^{\#}_{KK2(g)})\right]_{min} = \left[MSE(T^{\#}_{g2})\right]_{min} = \left[MSE(T^{\#}_{reg2})\right]_{min} \leq MSE\left(T^{\#}_{KK2(r)}\right)$$

as $MSE\left(T^{\#}_{KK2(r)}\right) - \left[MSE\left(T_{reg2}{}^{\#}\right)\right]_{min} = \left(V_p - 2C^2_{yp}\right)^2 \geq 0$, always true.

From (31) and (34)

$$\left[MSE(T^{\#}_{KK2(g)})\right]_{min} = \left[MSE(T^{\#}_{g2})\right]_{min} = \left[MSE(T^{\#}_{reg2})\right]_{min} \leq MSE\left(T^{\#}_{KK2(p)}\right)$$

as $MSE\left(T^{\#}_{KK2(p)}\right) - \left[MSE\left(T^{\#}_{reg2}\right)\right]_{min} = \left(V_p + 2C^2_{yp}\right)^2 \geq 0$, always true.

(vii) From (11), (12) and (43)

$$\left[MSE\left(T_{1prop}^{r}\right)\right]_{min} \leq \left[MSE(T_{g1}^{\#})\right]_{min} = \left[MSE(T_{reg1}^{\#})\right]_{min} \quad \text{if} \quad \theta \geq \frac{-v_p V_y + c_{yp}^2}{v_p c_{yp}}.$$

(viii) From (11), (12) and (44)

$$\left[MSE\left(T_{1prop}^{p}\right)\right]_{min} \leq \left[MSE(T_{g1}^{\#})\right]_{min} = \left[MSE(T_{reg1}^{\#})\right]_{min} \quad \text{if} \quad \theta \geq \frac{v_p V_y - c_{yp}^2}{v_p\, c_{yp}}.$$

Following the same steps as in (vii) and (viii), we have:

(ix) From (19), (20) and (53)

$$\left[MSE\left(T_{2prop}^{r}\right)\right]_{min} \leq \left[MSE\left(T_{g2}^{\#}\right)\right]_{min} = \left[MSE\left(T_{reg2}^{\#}\right)\right]_{min} \quad \text{if} \quad \theta \geq \frac{-V_p V_y + C_{yp}^2}{V_p\, C_{yp}}.$$

(x) From (19), (20) and (54)

$$\left[MSE\left(T_{2prop}^{p}\right)\right]_{min} \leq \left[MSE\left(T_{g2}^{\#}\right)\right]_{min} = \left[MSE\left(T_{reg2}^{\#}\right)\right]_{min} \quad \text{if} \quad \theta \geq \frac{V_p V_y - C_{yp}^2}{V_p\, C_{yp}}.$$

From these theoretical comparisons, it has been observed that the proposed estimators would be more efficient than the mean unbiased estimator, ratio, product, regression, generalized and classes of estimators under the specified conditions.

A momentous remark in the overall comparison is that all the members of suggested ratio estimators will be more efficient than corresponding product estimators in respective cases if either $\theta \geq 0$ and $C_{yp} \geq 0$ or $\theta \leq 0$ and $C_{yp} \leq 0$ otherwise results will be reverse.

5. Empirical Study

An empirical study using the real data set has been conducted to show evidence of theoretical comparison and result derivation. The purpose of the study's data set is merely to provide illustrations; analysis is not intended for these data.

Data Description-

We have taken into consideration the Census Data -1981, released by the Government of India of Orissa, Police Station - Baria, Tahsil - Champua. This data set includes the number of agricultural laborers and occupied houses in 109 villages under the jurisdiction of the Baria police station. Data representing upper 25% of all villages (i.e. 27 villages) are taken into account for the population's unit non-respondents.

The study variable (y) is the number of agricultural laborers employed in a village; the auxiliary variable (x) is the number of houses that are occupied in that village. Because the number of occupied homes varies from village to village, villages have been classified as either big or small based on the number of occupied houses. In this instance, a village receives the attribute (φ) of a big village if it has more than 70 occupied houses; if not, it is categorized as a small village.

The parameters for this data are:

$N = 109$	$n = 30$	$\bar{Y} = 41.2385$	$P = 0.5229$
$\lambda_n = 0.02416$	$W_2 = 0.2477$	$\bar{Y}_2 = 51.7037$	$P_{(2)}$=0.7037
$S_y = 46.64779$	$S_p = 0.50178$	$\rho_{yp} = 0.426$	$S_{y(2)} = 38.42857$
$S_{p_{(2)}} = 0.46532$	$\rho_{yp(2)} = 0.227$	$\beta_1 = 2.4103$	$\beta_2 = 6.912$
	$C_y = 1.1312$	$C_p = 0.9596$	

To show the efficiency of the proposed estimators, minimum mean square errors are calculated along with the relevant existing estimators. The percentage relative efficiency (PRE) of the proposed ($prop$) and relevant existing (ex) estimators with respect to conventional mean per unit unbiased estimator (T_{HH}) is calculated by the formula:

$$PRE\left(T_{ex/prop}\right) = \frac{V(T_{HH})}{\left[MSE(T_{ex/prop})\right]_{min}} \times 100.$$

For cases I and II under the considered data set, the minimum MSE and PRE of the proposed and existing estimators are obtained and provided in Tables 1 and 2.

Table 1: ***MSE*** and ***PRE*** of estimators for different values of $\mathbf{1/\omega}$ (for Case I)

Estimator	***MSE*(*PRE*) and constants**			
	$\mathbf{1/\omega = 1/5}$	$\mathbf{1/\omega = 1/4}$	$\mathbf{1/\omega = 1/3}$	$\mathbf{1/\omega = 1/2}$
T_{HH}	101.345(100%)	89.152(100%)	76.959(100%)	64.7656(100%)
$T_{r1}^{\#}$	101.181(100.2%)	88.988(100.2%)	76.795(100.2%)	64.602(100.2%)
$T_{p1}^{\#}$	177.178(57.2%)	164.985(54%)	152.792(50.4%)	140.599(46.1%)
$T_{g1}^{\#}$	91.804(110.4%)	79.611(112%)	67.418(114.2%)	55.225(117.3%)
$T_{reg1}^{\#}$	91.804(110.4%)	79.611(112%)	67.418(114.2%)	55.225(117.3%)
$T_{KK1(r)}^{\#}$	91.804(110.4%)	79.611(112%)	67.418(114.2%)	55.225(117.3%)
$T_{KK1(p)}^{\#}$	129.803(78.1%)	117.611(75.8%)	105.417(73%)	93.224(69.5%)
$T_{KK1(g)}^{\#}$	91.804(110.4%)	79.611(112%)	67.418(114.2%)	55.225(117.3%)
$\boldsymbol{T_{1prop}^{r(1)}}$	**64.254(157.7%)**	**57.631(154.9%)**	**50.944(151.1%)**	**44.191(146.6%)**
$\boldsymbol{T_{1prop}^{r(2)}}$	**61.432(165.0%)**	**55.006(162.1%)**	**48.512(158.6%)**	**41.950(154.4%)**
$T_{1prop}^{r(3)}$	69.672(145.5%)	62.635(142.3%)	55.537(138.6%)	48.376(133.9%)
$T_{1prop}^{r(4)}$	69.748(145.3%)	62.705(142.2%)	55.600(138.4%)	48.432(133.7%)
$T_{1prop}^{r(5)}$	70.006(144.8%)	62.941(141.6%)	55.813(137.9%)	48.623(133.2%)
$T_{1prop}^{r(6)}$	66.501(152.4%)	59.716(149.3%)	52.867(145.6%)	45.955(140.9%)
$T_{1prop}^{p(1)}$	73.686(137.5%)	66.139(134.8%)	58.507(131.5%)	50.788(127.5%)
$T_{1prop}^{p(2)}$	73.958(137.0%)	66.292(134.5%)	58.530(131.5%)	50.667(127.8%)
$T_{1prop}^{p(3)}$	71.991(140.8%)	64.730(137.7%)	57.402(134.1%)	50.008(129.5%)
$T_{1prop}^{p(4)}$	71.943(140.9%)	64.687(137.8%)	57.366(134.2%)	49.977(129.6%)
$T_{1prop}^{p(5)}$	71.771(141.2%)	64.534(138.1%)	57.233(134.5%)	49.865(129.9%)
$T_{1prop}^{p(6)}$	73.261(138.3%)	65.816(135.4%)	58.295(132.0%)	50.698(127.7%)

Source: Own work.

Table 2: $\boldsymbol{MSE}$ and $\boldsymbol{PRE}$ of estimators for different values of $^1/_{\omega}$ (for Case II)

Estimator	$\boldsymbol{MSE(PRE)}$ and constants			
	$^1/_{\omega} = {}^1/_5$	$^1/_{\omega} = {}^1/_4$	$^1/_{\omega} = {}^1/_3$	$^1/_{\omega} = {}^1/_2$
T_{HH}	101.345(100%)	89.1518(100%)	76.959(100%)	64.7656(100%)
$T_{r2}^{\#}$	124.513(81.4%)	106.487(83.7%)	88.461(87.0%)	70.4351(92.0%)
$T_{p2}^{\#}$	242.800(41.7%)	214.202(41.6%)	185.603(41.5%)	157.004(43.2%)
$T_{g2}^{\#}$	90.721(111.7%)	78.966(112.9%)	67.141(114.6%)	55.198(117.3%)
$T_{reg2}^{\#}$	90.721(111.7%)	78.966(112.9%)	67.141(114.6%)	55.198(117.3%)
$T_{KK2(r)}^{\#}$	92.351(109.7%)	80.021(111.4%)	67.692(113.7%)	55.362(117.0%)
$T_{KK2(p)}^{\#}$	151.495(66.9%)	133.879(66.6%)	116.26(66.2%)	98.646(65.6%)
$T_{KK2(g)}^{\#}$	90.721(111.7%)	78.966(112.9%)	67.141(114.6%)	55.198(117.3%)
$\boldsymbol{T_{2prop}^{r(1)}}$	**59.611(170.0%)**	**54.512(163.5%)**	**49.108(156.7%)**	**43.398(149.2%)**
$\boldsymbol{T_{2prop}^{r(2)}}$	**53.758(188.5%)**	**49.918(178.6%)**	**45.542(169.0%)**	**43.668(148.3%)**
$T_{2prop}^{r(3)}$	68.557(147.8%)	61.8956(144%)	55.120(139.6%)	48.214(134.3%)
$T_{2prop}^{r(4)}$	68.666(147.6%)	61.9872(143.6%)	55.196(139.4%)	48.276(134.2%)
$T_{2prop}^{r(5)}$	69.028(146.8%)	62.2948(143.1%)	55.453(138.8%)	48.486(133.6%)
$T_{2prop}^{r(6)}$	63.643(159.2%)	57.7873(154.3%)	51.734(148.8%)	45.472(142.4%)
$T_{2prop}^{p(1)}$	73.346(138.2%)	65.8557(135.8%)	58.310(132.0%)	50.694(127.8%)
$T_{2prop}^{p(2)}$	73.254(138.3%)	65.6986(135.7%)	58.101(132.4%)	50.444(128.4%)
$T_{2prop}^{p(3)}$	71.635(141.5%)	64.5046(138.2%)	58.213(132.2%)	49.981(129.6%)
$T_{2prop}^{p(4)}$	71.576(141.6%)	64.455(138.3%)	57.251(134.4%)	49.949(129.7%)
$T_{2prop}^{p(5)}$	71.364(142.0%)	64.276(138.7%)	57.104(134.8%)	49.840(130.0%)
$T_{2prop}^{p(6)}$	73.042(138.7%)	65.653(135.8%)	58.199(132.2%)	50.663(127.8%)

Source: Own work.

The bias of estimators has been calculated and is displayed in Table 3 in order to better support the comparison regarding the efficiency of the suggested estimators.

Table 3: $\boldsymbol{Bias}$ of estimators for different values of $^1/_{\omega}$ (for Case I)

Estimator	Bias			
	$^1/_{\omega} = {}^1/_5$	$^1/_{\omega} = {}^1/_4$	$^1/_{\omega} = {}^1/_3$	$^1/_{\omega} = {}^1/_2$
$\boldsymbol{T_{1prop}^{r(1)}}$	**−1.558**	**−1.398**	**−1.235**	**−1.072**
$\boldsymbol{T_{1prop}^{r(2)}}$	**−1.490**	**−1.334**	**−1.176**	**−1.017**
$T_{1prop}^{r(3)}$	−1.689	−1.519	−1.347	−1.173
$T_{1prop}^{r(4)}$	−1.691	−1.520	−1.348	−1.174
$T_{1prop}^{r(5)}$	−1.698	−1.526	−1.353	−1.179
$T_{1prop}^{r(6)}$	−1.613	−1.448	−1.282	−1.114
$T_{1prop}^{p(1)}$	−1.787	−1.604	−1.419	−1.232
$T_{1prop}^{p(2)}$	−1.793	−1.608	−1.419	−1.229
$T_{1prop}^{p(3)}$	−1.746	−1.570	−1.392	−1.213
$T_{1prop}^{p(4)}$	−1.744	−1.569	−1.391	−1.212
$T_{1prop}^{p(5)}$	−1.740	−1.565	−1.388	−1.209
$T_{1prop}^{p(6)}$	−1.776	−1.596	−1.414	−1.299

Source: Own work.

Table 4: $\boldsymbol{Bias}$ of estimators for different values of $\boldsymbol{1/\omega}$ (for Case II)

Estimator	Bias			
	$1/\omega = 1/5$	$1/\omega = 1/4$	$1/\omega = 1/3$	$1/\omega = 1/2$
$\boldsymbol{T_{2prop}^{r(1)}}$	$\boldsymbol{-1.446}$	$\boldsymbol{-1.322}$	$\boldsymbol{-1.191}$	$\boldsymbol{-1.052}$
$\boldsymbol{T_{2prop}^{r(2)}}$	$\boldsymbol{-1.304}$	$\boldsymbol{-1.210}$	$\boldsymbol{-1.104}$	$\boldsymbol{-0.986}$
$T_{2prop}^{r(3)}$	−1.662	−1.501	−1.337	−1.169
$T_{2prop}^{r(4)}$	−1.665	−1.503	−1.338	−1.171
$T_{2prop}^{r(5)}$	−1.674	−1.511	−1.345	−1.176
$T_{2prop}^{r(6)}$	−1.543	−1.401	−1.254	−1.103
$T_{2prop}^{p(1)}$	−1.778	−1.597	−1.414	−1.229
$T_{2prop}^{p(2)}$	−1.776	−1.593	−1.409	−1.223
$T_{2prop}^{p(3)}$	−1.737	−1.564	−1.389	−1.212
$T_{2prop}^{p(4)}$	−1.736	−1.563	−1.388	−1.211
$T_{2prop}^{p(5)}$	−1.730	−1.559	−1.385	−1.208
$T_{2prop}^{p(6)}$	−1.771	−1.592	−1.411	−1.228

Source: Own work.

Tables 1 and 2 show that in the two distinct cases of non-response, the estimators for regression ($T_{reg1}^{\#}$ and $T_{reg2}^{\#}$), generalized ($T_{g1}^{\#}$ and $T_{g2}^{\#}$), and Kumar and Kumar ($T_{KK1(g)}^{\#}$ and $T_{KK2(g)}^{\#}$) exhibited equal efficiency among all the predominating existing estimators. Tables 1 and 2 also show that in both non-response scenarios, every member of the suggested estimators is more efficient than every member of the pre-defined estimators currently in use at every level of sub-sampling fraction (ω^{-1}). Furthermore, the bias of each member of suggested estimators under the two distinct non-response cases is presented in Tables 3 and 4, where it is evident that the estimators ($T_{1prop}^{r(1)}$ and $T_{1prop}^{r(2)}$) and ($T_{2prop}^{r(1)}$ and $T_{2prop}^{r(2)}$) achieve the lowest bias value among all suggested members of the proposed estimators for all values of ω^{-1}.

6. Simulation Study

A simulation study has been carried out to provide the reliability of the comparison of the efficacy of the suggested estimators by real data. According to the District Census Handbook from 1981, 96 villages in the rural area under Police Station Singur in the District of Hooghly, West Bengal, have been taken into consideration for the simulation study [Source: Khare and Sinha (2011)]. The first 25% of the villages, or 24 villages, have been deemed the population's non-respondent group.

Here, the village's population is used as the study character (y), and its area is used as an auxiliary character (x_1). In this case, if a village has an area larger than 80 hectares,

it is given the attribute (φ) of a big area village; otherwise, it is classified as a small area village. The parameters of this study are:

$N = 96$	$n = 40$	$\bar{Y} = 1993.3$	$P = 0.7292$
$\lambda_n = 0.5833$	$W_2 = 0.3958$	$\bar{Y}_2 = 2394.8$	$P_{(2)} = 0.8158$
$S_y = 2308.3484$	$S_p = 0.4467$	$\rho_{yp} = 0.341$	$S_{y(2)} = 2971.6196$
$S_{p_{(2)}} = 0.3929$	$\rho_{yp(2)} = 0.251$	$\beta_1 = 1.0642$	$\beta_2 = 2.0640$
	$C_y = 1.1581$	$C_p = 0.6126$	

Through the use of R software, a random sample of size 40 is drawn from this population. The estimators' values $(T_{ex/prop})$ have been calculated using 3000 replications, and their $MSEs$ have been calculated using the following formula:

$$MSE(T_{ex/prop}) = \frac{1}{3000}\sum_i^{3000}\left(T_{ex/prop} - \bar{Y}\right)^2.$$

The minimum MSE and PRE in conjunction with the constants involved in proposed and existing estimators for case I and II are given in Tables 5 and 6 respectively.

Table 5: ***MSE*** and ***PRE*** of estimators for different values of $^1/_{\omega}$ (for Case I)

Estimator	***MSE*(*PRE*) and constants**			
	$^1/_{\omega} = {}^1/_5$	$^1/_{\omega} = {}^1/_4$	$^1/_{\omega} = {}^1/_3$	$^1/_{\omega} = {}^1/_2$
T_{HH}	341376.2(100%)	288854.3(100%)	180370.2(100%)	142235.2(100%)
$T_{r1}^{\#}$	332198.3(102.8%)	276059.6(104.6%)	173184.9(104.2%)	136618.4(100.2%)
$T_{p1}^{\#}$	404432.0(84.4%)	338662.3(85.3%)	229098.1(78.7%)	186186.0(76.4%)
$T_{g1}^{\#}$	330045.3(103.4%)	276379.4(104.5%)	170828.3(105.6%)	134345.2(105.9%)
$T_{reg1}^{\#}$	331010.3(103.1%)	276112.0(104.6%)	171045.0(105.4%)	133980.4(106.2%)
$T_{KK1(r)}^{\#}$	338842.9(100.7%)	276908.7(104.3%)	172245.4(104.7%)	140798.5(101.0%)
$T_{KK1(p)}^{\#}$	437661.8(78.0%)	361235.8(80.0%)	23507.3(76.7%)	182300.0(78.0%)
$T_{KK1(g)}^{\#}$	331235.2(103.1%)	275929.2(104.7%)	170223.0(106.0%)	134031.2(106.1%)
$T_{1prop}^{r(1)}$	324032.0(105.4%)	265591.3(108.8%)	169100.8(106.7%)	132441.1(107.4%)
$T_{1prop}^{r(2)}$	320142.5(106.6%)	262760.4(109.9%)	169776.6(106.2%)	133163.0(106.8%)
$T_{1prop}^{r(3)}$	317643.7(107.5%)	256724.2(108.7%)	169147.1(106.6%)	132469.8(107.4%)
$T_{1prop}^{r(4)}$	326438.5(104.6%)	265860.8(108.6%)	169196.6(106.6%)	132500.6(107.4%)
$T_{1prop}^{r(5)}$	329327.6(103.6%)	266404.8(108.4%)	169409.5(106.5%)	132636.4(107.2%)
$T_{1prop}^{r(6)}$	334343.7(102.1%)	263559.4(109.6%)	168842.9(106.8%)	132364.3(107.5%)
$T_{1prop}^{p(1)}$	330475.0(103.3%)	270408.2(106.8%)	171445.0(105.2%)	133951.5(106.2%)
$T_{1prop}^{p(2)}$	334892.5(101.9%)	273329.1(105.7%)	173222.7(104.1%)	135141.1(105.2%)
$T_{1prop}^{p(3)}$	330489.3(103.3%)	270290.4(106.9%)	171378.9(105.2%)	133908.1(106.2%)
$T_{1prop}^{p(4)}$	328424.3(103.9%)	270169.0(106.9%)	171310.9(105.3%)	133862.6(106.2%)
$T_{1prop}^{p(5)}$	329475.3(103.6%)	269680.2(107.1%)	171039.5(105.5%)	133682.9(106.4%)
$T_{1prop}^{p(6)}$	334325.6(102.1%)	272319.7(106.1%)	172579.0(104.5%)	134709.5(105.6%)

Source: Own work.

Table 6: ***MSE*** and ***PRE*** of estimators for different values of $^1/_{\omega}$ (for Case II)

Estimator	***MSE*(*PRE*) and constants**			
	$^1/_{\omega} = {}^1/_5$	$^1/_{\omega} = {}^1/_4$	$^1/_{\omega} = {}^1/_3$	$^1/_{\omega} = {}^1/_2$
T_{HH}	341376.2(100%)	288854.3(100%)	180370.2(100%)	142235.2(100%)
$T_{r2}^{\#}$	322053.0(106%)	257895.6(112.0%)	167848.1(107.5%)	135439.3(105.0%)
$T_{p2}^{\#}$	474133.6(72.0%)	409292.1(70.6%)	264000.0(68.4%)	205304.1(69.3%)
$T_{g2}^{\#}$	318754.2(107.1%)	258343.2(111.8%)	164000.0(110.0%)	131069.2(108.5%)
$T_{reg2}^{\#}$	318754.2(107.1%)	258343.2(111.8%)	164000.0(110.0%)	131069.2(108.5%)
$T_{KK2(r)}^{\#}$	297043.2(114.9%)	259437.3(111.3%)	165035.0(109.3%)	131980.4(107.8%)
$T_{KK2(p)}^{\#}$	502623.8(67.9%)	431075.2(67.0%)	265045.3(68.0%)	206043.5(69.0%)
$T_{KK2(g)}^{\#}$	313147.8(109.0%)	258354.7(111.8%)	164864.3(109.4%)	131145.0(108.4%)
$T_{2prop}^{r(1)}$	305618.8(111.7%)	249933.3(115.6%)	163543.6(110.3%)	128772.2(110.4%)
$T_{2prop}^{r(2)}$	304876.2(112.0s%)	277232.5(104.2%)	175000.0(103.0%)	132906.2(107.0%)
$T_{2prop}^{r(3)}$	289765.2(117.8%)	250023.1(115.5%)	163588.8(110.3%)	128799.6(110.4%)
$T_{2prop}^{r(4)}$	288475.3(118.3%)	250163.1(115.5%)	164000.0(110.0%)	128833.3(110.4%)
$T_{2prop}^{r(5)}$	289043.6(118.1%)	251007.9(115.1%)	163871.8(110.1%)	129008.9(110.2%)
$T_{2prop}^{r(6)}$	292345.2(116.8%)	259563.5(111.3%)	167000.0(108.0%)	129823.9(109.6%)
$T_{2prop}^{p(1)}$	326445.0(104.5%)	260355.6(111.0%)	167000.0(108.0%)	131215.4(108.4%)
$T_{2prop}^{p(2)}$	301732.3(112.1%)	269485.9(107.2%)	171469.2(105.2%)	133572.9(106.5%)
$T_{2prop}^{p(3)}$	307049.2(111.2%)	260028.0(111.1%)	167000.0(108.0%)	131134.5(108.5%)
$T_{2prop}^{p(4)}$	298321.3(114.4%)	259694.5(111.2%)	167188.7(107.9%)	131052.5(108.5%)
$T_{2prop}^{p(5)}$	295464.8(115.5%)	258392.9(111.8%)	167000.0(108.0%)	130734.1(108.8%)
$T_{2prop}^{p(6)}$	303234.7(112.6%)	266000.0(109.0%)	169968.7(106.1%)	132701.8(107.2%)

Source: Own work.

The simulation study results shown in Tables 5 and 6 validate the theoretical findings about the MSE and estimator efficiency calculated using real data and displayed in Tables 1 and 2. Replications have, however, resulted in nominal changes in the MSE and PRE of the estimators ($T_{g1}^{\#}$, $T_{reg1}^{\#}$ and $T_{KK1(g)}^{\#}$) and ($T_{g2}^{\#}$, $T_{reg2}^{\#}$ and $T_{KK2(g)}^{\#}$).

7. Conclusions

From the analytical study of empirical data, it is clear for both the cases I and II that the proposed estimators are more efficient than all the existing estimators. For case I, when non-response occurs only on study variable, $T_{1prop}^{r(2)}$ and $T_{1prop}^{r(1)}$ are more efficient than all other relevant proposed ratio type estimators while in the category of product type estimators, $T_{1prop}^{p(5)}$ and $T_{1prop}^{p(4)}$ are found to be more efficient. For case II, when non-response occurs on both study variable as well as auxiliary attribute, the proposed

estimators $\left(T_{2_{prop}}^{r(2)},\ T_{2_{prop}}^{r(1)} \right)$ and $\left(T_{2_{prop}}^{p(5)},\ T_{2_{prop}}^{p(4)} \right)$ are more efficient among all the members of the proposed ratio and product type estimators respectively. Further, it has also been observed that MSE and PRE both decrease when the values of sub-sampling fraction (ω^{-1}) increase. The reason of the decreasing PRE of the proposed estimators is the faster rate of decrease of the variance of T_{HH} compared to the proposed estimators.

A simulation study confirms and reveals that the efficiency of the proposed estimators is significantly higher than all the relevant estimators at every level of sub-sampling fractions (ω^{-1}), however some estimators have average efficiency as the value of the coefficient of skewness is very small.

Therefore, on the basis of theoretical, empirical and simulation studies, the proposed estimators may be recommended for the improved estimation of mean subject to the condition of availability of the suggested constants of auxiliary variable to increase the precision. It means that one can use any available known parameter of the auxiliary variable among the suggested ones to obtain the efficient estimate, since all members of the proposed estimators are efficient with less $\boldsymbol{MSE}$ compared to all conventional adopted as well as predominating existing estimators.

Acknowledgements

The authors express their gratitude to the Editor-in-Chief and two eminent referees for their valuable comments/suggestions, which greatly helped in bringing the paper to its present form.

References

Bethlehem, J., Cobben, F., Schouten, B., (2011). Handbook of non-response in household surveys. John Wiley and Sons.

Cochran, W. G., (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce, *Jour. Agri. Sci.*, pp. 262–275.

Hansen, M. H., Hurwitz, W. N., (1946). The problem of non-response in sample surveys, *Jour. Amer. Stat. Assoc.*, Vol. 41, pp. 517–529.

Kadilar, C., Cingi, H., (2004). Ratio estimators in simple random sampling, *Appl. Math. Comp.*, Vol. 151, pp. 893–902.

Khare, B. B., Srivastava, S., (1993). Estimation of population mean using auxiliary character in presence of non-response, *Nat. Acad. Sci. Lett.*, Vol. 16, pp. 111–114.

Khare, B. B., Srivastava, S., (1995). Study of conventional and alternative two phase sampling ratio, product and regression estimators in presence of non-response, *Proc. Nat. Acad. Sci.*, Vol. 65(A) II, pp. 195–203.

Khare, B. B., Srivastava, S., (1997). Transformed ratio type estimators for the population mean in the presence of non-response, *Comm. Stat. –Theo. Meth.*, Vol. 26(7), pp. 1779–1791.

Khare, B. B., Srivastava, S., (2000). Generalized estimators for population mean in presence of non-response, *Inter. Jour. Math. Stat.*, Vol. 9, pp. 75–87.

Khare, B. B., Sinha, R. R., (2002). Estimation of the ratio of two population means using auxiliary character with unknown population mean in presence of non-response, *Prog. Math. BHU*, Vol. 36, pp. 337–348.

Khare, B. B., (2003). Use of auxiliary information in sample surveys up to 2000–A review, *Proc. Bio. Sci. India: M/S Centre of Bio-Mathematical Studies*, pp. 76–87.

Khare, B. B., Sinha, R. R., (2009). On class of estimators for population mean using multi-auxiliary characters in the presence of non-response, *Stat. Trans. new series*, Vol. 10(1), pp. 3–14.

Khare, B. B., Sinha, R. R., (2011). Estimation of population mean using multi-auxiliary characters with sub sampling the non-respondents, *Stat. Trans. new series*, Vol. 12(1), pp. 45–56.

Koyuncu, N., (2012). Efficient estimators of population mean using auxiliary attributes, App. Math. Comp., Vol. 218(22), pp. 10900–10905.

Kumar, K., Kumar, A., (2019). Estimation of population mean using auxiliary attribute in the presence of non-response, *Int. J. Comp. Theo. Stat.*, Vol. 6(1), pp. 43–49.

Rao, P. S. R. S., (1986). Ratio estimation with sub-sampling the non-respondents, *Survey Methodology*, Vol. 12, pp. 217–230.

Rao, P. S. R. S., (1990). Regression estimators with sub-sampling of non-respondents, In-Data Quality Control, Theory and Pragmatics, (Eds.) Gunar E. Liepins and V.R.R. Uppuluri, Marcel Dekker, New York, pp. 191–208.

Reddy, V. N., (1978). A study on the use of prior knowledge on certain population parameters in estimation, *Sankhya C*, Vol. 40, pp. 29–37.

Riaz, S., Darda, M. A., (2016). Some classes of estimators in the presence of non-response using auxiliary attribute, *Springer Plus*, Vol. 5(1), pp. 1–14.

Singh, H. P., Tailor, R., (2003). Use of known correlation coefficient in estimating the finite population mean, *Stat. Trans. new series*, Vol. 6(4), pp. 555–560.

Singh, H. P., Kumar, S., (2009). A general class of estimators of the population mean in survey sampling using auxiliary information with sub-sampling the non-respondents, *Korean Jour. Appl. Stat.*, Vol. 22(2), pp. 387–402.

Singh, R., Mishra, P., Bouza, C., (2019). Estimation of Population Mean using information on auxiliary attribute: A Review, Ranked Set Sampling, Elsevier.

Singh, H. P., Solanki, R. S., (2012). Improved estimation of population mean in simple random sampling using information on auxiliary attribute, *App Math. Comp.*, Vol. 218, pp. 7798–7812.

Sinha, R. R., Bharti, (2021). Regress exponential estimators for estimating the population mean via auxiliary attribute, *Int. Jour. App. Math. Stat.*, Vol. 60(2), pp. 18–29.

Sinha, R. R., Bharti, (2022) Ameliorate estimation of mean using skewness and kurtosis of auxiliary character, *Jour. Stat. Manag. Sys.*, DOI: 10.1080/09720510.2021. 1966956.

Sinha, R. R., Kumar, V., (2011). Generalized estimators for population mean with sub sampling the non-respondents, *Aligarh Jour. Stat.*, Vol. 31, pp. 53–62.

Sinha, R. R., Kumar, V., (2013). Improved estimators for population mean using attributes and auxiliary characters under incomplete information, *Inter. Jour. Math. Stat.*, Vol. 14, pp. 43–54.

Sinha, R. R., Kumar, V., (2014). Improved classes of estimators for population mean using information on auxiliary character under double sampling the non-respondents, *Nat. Acad. Sci. Lett.*, Vol. 37(1), pp. 71–79.

Srivastava, S. K., Jhajj, H. S., (1983). A class of estimators of population mean using multi-auxiliary information, *Cal. Stat. Assoc. Bull.*, Vol. 32, pp. 47–56.

Tripathi, T. P., Das, A. K., Khare, B. B., (1994). Use of auxiliary information in sample surveys – A review, *Aligarh Jour. Stat.*, Vol. 14, pp. 79–134.

Yadav, S. K., Zaman, T., (2021). Use of some conventional and non-conventional parameters for improving the efficiency of ratio-type estimators. *Jour. Stat. Manag. Sys.*, Vol. 24(5), pp. 1077–1100.

Zaman, T., (2020). Generalized exponential estimators for the finite population mean. *Statistics in Transition. New Series*, Vol. 21(1), pp. 159–168.

Zaman, T., Kadilar, C., (2019). Novel family of exponential estimators using information of auxiliary attribute. *Jour. Stat. Manag. Sys.*, Vol. 22(8), pp. 1499–1509.

Zaman, T., Kadilar, C., (2021 a). Exponential ratio and product type estimators of the mean in stratified two-phase sampling. *AIMS Mathematics*, Vol. 6(5), pp. 4265–4279.

Zaman, T., Kadilar, C., (2021 b). New class of exponential estimators for finite population mean in two-phase sampling. *Comm. Stat. –Theo. Meth.*, Vol. 50(4), pp. 874–889.

STATISTICS IN TRANSITION new series, September 2024
Vol. 24, No. 3, pp. 123–140, https://doi.org/10.59170/stattrans-2024-031
Received – 02.03.2023; accepted – 30.04.2024

Forecasts of the mortality risk of COVID-19 using the Markov-switching autoregressive model: a case study of Nigeria (2020–2022)

Idowu Oluwasayo Ayodeji[1]

Abstract

The global pandemic due to SARS-Cov-2 ravaged the world and killed more than 6 million people globally within two years. Studies predicting future occurrences are essential to effectively combat the virus. This study modeled daily fatality rate in Nigeria from March 23, 2020 to March 19, 2022 and forecast future occurrences using Markov switching model (MSM). MSM estimates segmented fatality rates into three states of low-, medium- and high-risks. Further, estimates revealed that as at 19th March, 2022, Nigeria remained at the low-risk regime in which 1 (95%CI: 0, 1) person, on the average, died of coronavirus daily; however, the most probable scenario in the nearest future was the medium-risk state in which an average of 4 (95%CI: 2, 5) persons would die daily with 48.7% probability. The study concluded that Nigerian COVID mortality risks followed a switching pattern which fluctuated within low-, medium- and high-risks; however, the medium-risk state was most likely in the future. Our results indicated that the quarantine measures adopted by the governments yielded positive results. It also underscored the need for governments and individuals to intensify efforts to ensure that the country remained at the low-risk zone till the virus would be eventually eradicated.

Key words: Nigeria, hidden Markov, autoregressive, coronavirus death rate.

1. Introduction

The novel coronavirus pandemic began in Wuhan, China on 8th December, 2019 (Ihekweazu, 2020) and sent shock waves around the world. Owing to its severity and rate of spread, the World Health Organization (WHO) declared it a public health emergency of international concern on the 30th January, 2020 (Ogundokun et al., 2020). None of the health institutions around the world was well prepared for its invasion as they were overwhelmed within few weeks of its arrival. Going by the way

[1] Obafemi Awolowo University, Ile-Ife, Nigeria. E-mail: ioayodeji@oauife.edu.ng, idowu.sayo@yahoo.com. ORCID: https://orcid.org/0000-0001-8671-5637.

it affected the developed countries which were better placed in terms of health infrastructure, the pundits had speculated that Africa would not be able to withstand the effect of the pandemic (Ibrahim et al., 2021). Nigeria, being the most populous nation in Africa was expected to be the worst hit given the mode of transmission of the virus and the fact that several settlements within the country had no decent access to health facilities (Marbot, 2020). Worse still, the poor level of infrastructure in the country posed serious challenges to effective testing and treatment of coronavirus patients (Adekunle et al., 2020). Accordingly, the country was categorized by WHO as one of the 13 high-risk priority countries in Africa (Ihekweazu, 2020).

In order to effectively combat the virus, an accurate description and prediction of the associated risk would be essential. It could provide individuals and health practitioners with understanding of the fatality level and assist in anticipating its progression in the future. It could also offer useful insights to policy makers to make evidence-based decisions and strengthen their efforts in combating the spread. In addition, it could be used to assess the level of success of health' interventions made by governments and health organizations so far and assist in future projections.

The major objective of this study was to formulate a suitable model for COVID mortality rates in Nigeria which would be used to forecast future occurrences. Various methods had been used in previous studies to model coronavirus fatalities; the most commonly-used included the compartmental models (See Adekunle et al. (2020), Bagal et al. (2020) and Carcione et al. (2020), among others), and the time series processes (See Anne (2020), Khan and Lounis (2021), Pourghasemi et al. (2020), Singh et al. (2020), among others).

The time series specification was also predominant in existing prediction studies (Didi et al., 2021; Ibrahim and Oladipo, 2020; Ibrahim et al., 2021; Khan and Lounis, 2021; Li et al., 2022; Odekina et al., 2022). In addition to being able to describe the dynamics of epidemics, time series models could reveal the underlying data-generating process which could be used to forecast future patterns associated with the epidemics. However, relative to modeling studies, existing prediction studies on coronavirus were few and were mostly limited to short-term forecasts due to the type of time series models adopted.

A new variant of the time series model, namely the Markov-switching specification, had emerged in the late nineties which could be used to make accurate forecasts for long-term horizons because its forecasts were based on probabilities rather than some average values. Markov-switching model (MSM) was most appropriate to model and forecast series of ***time-varying*** nature such as the coronavirus fatalities owing to its switching features. What do we mean by time-varying nature? Consider Figure 1 which referred to coronavirus mortality rates in Nigeria from 23rd March, 2020-19th March, 2022. One obvious fact to be observed from the Figure was that the mortality risk was

not constant over time but appeared to switch infrequently between the different states of low, medium and high risks. For instance, cases of medium and high mortality risk could be found under labels 1, 2, 4 and 5. Label 3, especially, corresponded to high mortality risk period, while the unlabeled sections appeared to be the low risk periods in Nigeria.

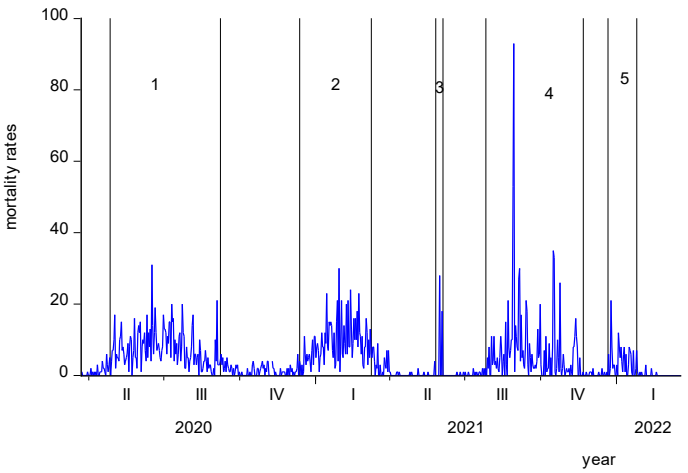


Figure 1: Coronavirus mortality rates in Nigeria (23rd March, 2020 – 19th March, 2022).

Owing to the peaks and troughs in Figure 1 therefore, the appropriate model for measuring and forecasting coronavirus mortality risk in Nigeria was the Markov-switching model (MSM) proposed by Hamilton (1989). In this context, MSM may be used to characterize mortality risks into different states that agreed with the realities in Figure 1. By design, MSM incorporates the Markov process into an autoregressive model such that the average mortality rate may vary across three (low, medium and high risks) unobserved states that evolved according to a first order Markov transition process.

In simple terms, MSM provided that the varying mortality rates seen in Figure 1 be modeled with a mixture of time series specifications whose transition was governed by a Markov process. This was in contrast with previous studies which employed a single model in forecasting. Ahlburg (1995) earlier noted that a combination of forecast models improve accuracy than a single model. Ibrahim et al. (2021) also provided empirical evidence in favour of multiple models.

MSM is especially applicable in the treatment of epidemics as it generates probabilities which can be used to measure the risks associated with the outbreak, as opposed to other time series variants which provide forecasts in form of some mean values. The probabilities obtained from MSM can be used to segment Nigerian coronavirus data into low-, medium- or high-risk state in line with the categorizations

of WHO (Ihekweazu, 2020). It is noteworthy that few application studies had applied MSM to model COVID cases in United States (Oliveira et al., 2021) and South Africa (Mthethwa et al., 2022). They found that MSM outperformed other time series and growth models.

Overview of the paper was as follows: Section 2 contained literature review; in Section 3 we presented data and statistical techniques; Section 4 concerned results and discussions; and lastly, Section 5 contained conclusions and recommendations.

2. Literature Review

2.1. Evolution of Coronavirus in Nigeria and Governments' Interventions

The novel coronavirus was first discovered in Wuhan China on 8th December, 2019 (Ihekweazu, 2020). It was declared a public health emergency of international concern by WHO on 30th January, 2020 after 118,000 people had been infected globally (Ogundokun et al., 2020). The Nigeria Centre for Disease Control (NCDC) not being unaware of the enormous challenges involved immediately constituted rapid response teams across the 36 states and concluded their trainings in December 2019. NCDC also subscribed to plans to work with 22 state governors to establish emergency operation centers whose activities will be coordinated at the national centre (Ihekweazu, 2020). In addition, Nigeria also intensified surveillance at the international airports to prevent importation of the virus. However, despite all these efforts, its first index case from Italy was reported on the 27th February, 2020, and its first death on 23rd March, 2020 (Ileyemi, 2021). As at 24th March, 2022, there were 255,244 confirmed cases out of which 249,486 recovered and 3,142 died (Worldometer, 2022).

On 29th March, 2020, the Federal government imposed lockdown in Lagos state, being the epicenter of the virus. Ogun state was also included in the curfew arrangement because it shared border with Lagos. The lockdown also extended to the Federal capital territory (FCT), Abuja, being the region with second highest number of confirmed cases. Movement restrictions were also enforced by state governments of some other states which were not included in the federal-government-imposed lockdown (Odekina et al., 2022). By 23rd April, 2020, movement restrictions had been enforced in all thirty-six states of the federation and the FCT (Jacobs and Okeke, 2022).

In addition to the lockdown order, the Federal government also imposed restrictions on influx of people from thirteen countries which included China, Italy, United States, Iran, South Korea, United Kingdom, Spain, Netherlands, Japan, France, Norway, Switzerland and Germany on 8th March, 2020 (Ibrahim and Oladipo, 2020). 13 days later, the Abuja and Lagos international airports were totally shut down. As part of the safety measures, transport by rail was also suspended on 23rd March, 2020

(Ibrahim and Oladipo, 2020). Due to economic considerations, the Federal government began to ease lockdown and travel restrictions gradually on 4th May, 2020. However, other measures such as social distancing, contact tracing, source control, quarantine, administration of COVID vaccines and self-isolation were sustained (Jacobs and Okeke, 2022).

The quarantine measures and safety protocols notwithstanding, Nigeria continued to record casualties on daily basis. On Saturday, 28th August 2021, NCDC reported a record death of 53 Nigerians, the highest ever since the first confirmed case in February; and barely 24 hours later, 93 more people died of the same cause (Ileyemi, 2021). Though NCDC noted that the high number recorded was due to backlog of fatalities from Lagos state, however, most people were of the opinion that it was the sign of the third wave, and this had sparked fears afresh in the minds of the citizens (Ileyemi, 2021).

2.2. Review of Related Studies

The theory behind coronavirus infection and prognosis was multi-faceted. The occurrence of the virus had been linked with the black swan theory in the literature because it reflected the three characteristics of the theory; (i) it was an unexpected event; (ii) it had extreme impact; and (iii) it could not be predicted in advance, but its occurrence could be explained after it has occurred (Taleb, 2007). Accordingly, researchers had made several attempts to explain its occurrence since it was made public.

Sharifi et al. (2022) presented coronavirus as a syndemic which affected various aspects of human lives. The effect of COVID on people living with non-communicable diseases may differ from others who were not due to different precautionary regimens that were imposed during the pandemic. For instance, movement restrictions reduced physical activities and exercises which were crucial in the management of diabetes, obesity, and so on. Further, isolation of aged people may also increase loneliness and mental health issues such as depression.

Sociological theory and implications of COVID pandemic were presented in Bello and Amzat (2021). The study leaned heavily on the assumptions of George Simmel, Auguste Comte and Herbert Spencer to explain the effect of the pandemic on people's way of life. Based on Comte's theory, the study considered COVID-19 a cause of social instability and disruption to social and economic well-being. Quarantine measures such as social distancing and lockdown would cause a breakdown in social order and interactions; however, they would eventually make life better in the long run.

On the other hand, the decision to ease lockdown protocols at some point despite the pandemic followed Spencer's theory of "survival of the fittest". It is of economic and social advantage to accept that individuals would adapt and learn to live with the virus

than to "lock them down" indefinitely. The consequences of long-term social and economic disorder outweighed that of coronavirus (Bello and Amzat, 2021).

Simmel's theory was applicable to interpersonal relationship among individuals. Personal relationships among individuals had been restricted to virtual interactions through social media. Telemedicine replaced routine hospital visits in advanced countries (Sharifi et al., 2022) and people spent more time online than with one another. A different perspective to coronavirus incidence is contained in Simmel's theories of secrecy and conflict, and how they affected social relationships during COVID (Bello and Amzat, 2021).

From the empirical viewpoint, most studies measuring the rate of coronavirus fatalities employed compartmental models (Adekunle et al., 2020; Bagal et al., 2020; Carcione et al., 2020) and time series processes (Anne, 2020; Khan and Lounis, 2021; Pourghasemi et al., 2020; Singh et al., 2020). Studies which employed the Markov switching variants in particular included Oliveira et al. (2021) and Mthethwa et al. (2022). Time series investigations conducted with Nigerian data were presented in Abdulmajeed et al. (2020), Chigbu et al. (2021), Li et al. (2022) and Odekina et al. (2022).

The most commonly-used method of forecasting was time series. Predictions on coronavirus status varied by outcome and dates: Chigbu et al. (2021) predicted values for Nigerian fatality and recovery rates from 25th August, 2020 to 31st January, 2021. The study predicted a gradual decrease in infection and fatality rates and an increase in recovery rate over the period of forecast. Li et al. (2022), in contrast to Chigbu et al. (2021), predicted an increase in infection and fatality rates in all the countries considered, including Nigeria, within 1st and 27th March, 2021; and Nigeria was expected to have the lowest prevalence rate among selected countries. Lastly, Mthethwa et al. (2022) predicted in South Africa 322 fatalities for 27th August, 2021 and expected the number to decrease over the next 10 days to 41.

3. Data collection and analysis

3.1. Data and study area

Daily counts of coronavirus deaths were reported on daily basis at the official website of the National Centre for Disease Control (https://covid19.ncdc.gov.ng/report/). They covered 23rd March, 2020, which corresponded to the death of the first index case, Suleiman Achimugu, to 19th March, 2022. The sample size was 727 observations. Details on the data collection process and other descriptions can be found at the website.

3.1. The Markov Switching Model

Denote d_t the daily death rate of coronavirus in Nigeria. The Markov switching model can be written as (Hamilton, 1989):

$$d_t = \mu_{S_t} + \sum_{i=1}^{p} \beta_{i,S_t} d_{i,t-1} + \varepsilon_t,\ S_t = 1,2,\cdots,m;\ t = 1,2,\cdots,T. \tag{1}$$

For mathematical tractability, it was assumed that d_t was normally distributed with means μ and variances σ^2 in the different possible states. A similar assumption was made in Engel and Hamilton (1990). The variance σ^2 varied alongside the mean so that the hidden Markov model could capture the peaks and troughs in Figure 1. At each different states S_t of the coronavirus fatality rate, a variance was computed. The autoregressive term d_{t-1} was included to capture serial correlation that may exist in the data. Following WHO categorization (Ihekweazu, 2020), the study adopted $m = 3$ in the 3 different states, which, in this context, may be denoted $S_t = 1$ for low-risk fatality rate, $S_t = 2$ for medium-risk fatality rate, and $S_t = 3$ for high-risk fatality rate. The Markov process employed was such that the state at any time t was determined randomly and only depended on the state at time $t - 1$.

Transition from one state to the other was determined by the transition probabilities P_{ij}. The closer the value of P_{ij} to 1 the longer it took for the system to transit to the next state. The probability of being in state i tomorrow given that death rate was in state i today was $P_{ii} = \Pr(S_t = i|S_{t-1} = i), i = 1,2,\cdots,m$. In general, $P_{ij} = \Pr(S_t = j|S_{t-1} = i)$ and

$$\sum_{j=1}^{m} P_{ij} = 1; j = 1,2,\cdots,m \text{ and for all } i. \tag{2}$$

The loglikelihood equation corresponding to System (1) - (2) can be written as (Engel and Hamilton, 1990):

$$\log L = \sum_{t=1}^{T} \log f(d_t|S_t) \tag{3}$$

where

$$f(d_t|S_t) = \frac{1}{\sigma_{S_t}\sqrt{2\pi}} \exp\left\{\frac{1}{2\sigma_{S_t}^2}\left(d_t - \mu_{S_t}\right)^2\right\} \tag{4}$$

S_t was not directly observable hence following Engel and Hamilton's (1990) we re-wrote $f(d_t|S_t)$ as

$$f(d_t, S_t|\upsilon_{t-1}) = f(d_t|S_t, \upsilon_{t-1})P(S_t|\upsilon_{t-1}) \tag{5}$$

and

$$f(d_t|\upsilon_{t-1}) = \sum_{S_t=1}^{m} f(d_t|S_t, \upsilon_{t-1})P(S_t|\upsilon_{t-1}) \tag{6}$$

where υ_{t-1} was the information available up to time $t-1$.

The loglikelihood equation (3) was then updated as follows:

$$\log L = \sum_{t=1}^{T} \log \sum_{S_t}^{m} f(d_t|S_t, \upsilon_{t-1}) P(S_t|\upsilon_{t-1}) \quad (7)$$

Following Hamilton (1989), model parameters were estimated from Equation (7) using the maximum likelihood approach. We also drew probabilistic inference in form of a nonlinear iterative filter and smoother, popularly referred to as filtering and smoothing probabilities, respectively in econometrics literature. Expected length or duration of state i was computed as (Engel and Hamilton, 1990):

$$E(D) = \frac{1}{1-P_{ii}} \quad (8)$$

For more comprehensive details about the method, interested reader may refer to Hamilton (1989).

4. Results

4.1. Model fitting

Following our observations in Figure 1, we estimated System (1)–(2) using 3 states[2]. Subsequently we categorized Nigeria as low-, medium- or high-risk country. The simplest Markov-switching autoregressive specification AR(1) has been shown to perform well in forecasting (Engel and Hamilton, 1990) hence it was adopted here. Table 1 referred to maximum likelihood estimates of the parameters in System (1)–(2).

Table 1: Maximum likelihood estimates of MSM

Parameter	States		
	1	2	3
μ	8.160* (-4.308, 12.011)	0.163* (0.104, 0.221)	3.329* (2.611, 4.046)
σ	9.137*	0.376*	3.187*
P_{ii}	0.125	0.834	0.694

95% confidence interval in parentheses *significant at the 5% level.

From our result, State 1 was identified as the high risk state where an average of 9 ($p<.05$) deaths were recorded on daily basis; State 2 as the low risk state where an average of 1 ($p<.05$) death was recorded on daily basis; and State 3 as the medium risk state where an average of 4 ($p<.05$) deaths were recorded on daily basis.

[2] We also experimented with 1 and 2 states, however, result, not displayed here for lack of space, showed that the 3 states model described the realities in Figure 1 better than the two other alternatives. Note that for only 1 state, MSM reduced to the conventional autoregressive (AR) model.

The estimated standard deviations σ showed that the number of deaths recorded in the high risk state varied more than in the other states. This implied that the deaths in high risk state were more difficult to predict compared to the remaining states.

Lastly, from the P_{ii} row we observed that the probability of remaining in low-risk state once the system was in it was the highest followed by the medium-risk and lastly the high-risk. Consequently, the expected durations for the three states were computed as 2, 4, and 7 days for high-, medium- and low-risk states, respectively. This implied that the high risk state was estimated to last for only 2 days at a stretch while the low-risk period was expected to last for 7 days whenever the system transited into the state.

The remaining transition probabilities were shown in the matrix,

$$\hat{P} = \begin{pmatrix} 0.125 & 0.088 & 0.787 \\ 0.024 & 0.834 & 0.142 \\ 0.192 & 0.114 & 0.694 \end{pmatrix}$$

Since these probabilities were within (0,1), we can infer that the system was transitive however the likelihood of transiting from one state to the other varied. Out of the three states, the highest likelihood of transition (78.7%) was from the high-risk state directly to the medium-risk state; whereas there was only an 2.4% chance of moving to a high-risk regime at the expiration of a low-risk mortality period.

Again, following Engel and Hamilton (1990) we tested the null hypothesis that the three states did not differ from one another. Wald's test statistics and corresponding p-values were presented in Table 2. The results showed that the three states were different both in the means and variances. This confirmed our observations in Figure 1 that the mortality risks varied over time.

Table 2: Test of hypotheses

Null hypothesis	Wald's statistic
$\mu_1 = \mu_2 = \mu_3$	110.9974 (0.0000)
$\sigma_1^2 = \sigma_2^2 = \sigma_3^2$	1150.423 (0.0000)

p-values in parentheses

Figure 2 displayed the filtered probability plots of the three states (regimes) superimposed with the daily death rates. We followed the conventional assumption in econometrics literature (Engel and Hamilton, 1990) that the system had switched from one state to the other when the probability exceeds 0.5. It was evident from the

plot that Markov switching model efficiently captured all the high-, medium- and low-risk mortality periods between 23rd March 2020 and 19th March, 2022 in Nigeria. The fatality rate switched infrequently among the three risk states; it switched severally between low- and medium-risk zones in the periods between the third and fourth quarters in 2020 whereas the high-risk regime dominated the third quarter of 2021. We recalled that Nigeria recorded 53 and 93 deaths in 2 consecutive days within August 2021 in the third quarter of 2021.

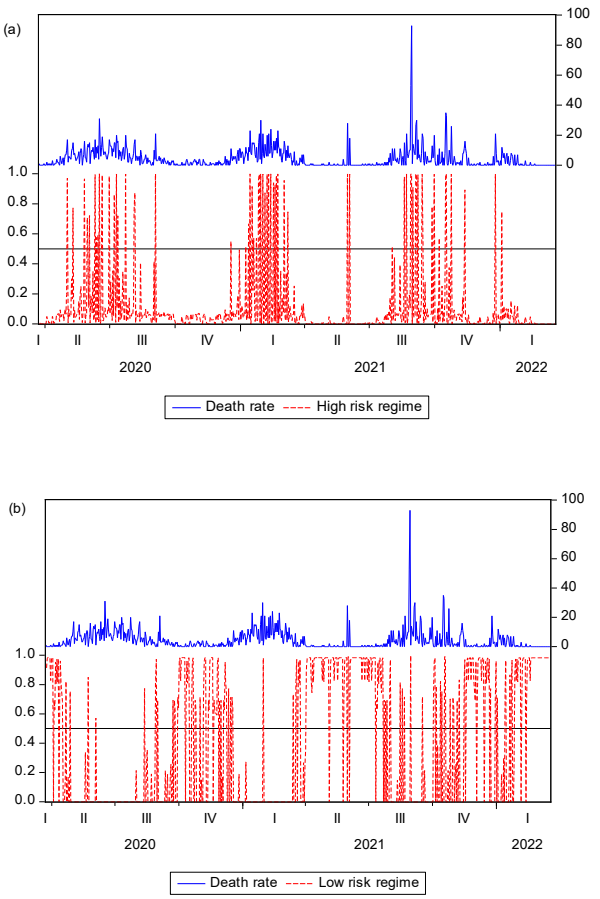


Figure 2: Filtered probability plots by states superimposed with death rates.

4.2. Forecast

We compared three specifications of MSM, namely the 1-state case, 2-state case and 3-state case using various metrics and graphics. The details were available in the

Appendix. The highlight of the result was that the 3-state model outperformed the other two in all the measures employed; hence, it was used for forecast.

Given

$$\widehat{\mathrm{P}} = \begin{pmatrix} 0.125 & 0.088 & 0.787 \\ 0.024 & 0.834 & 0.142 \\ 0.192 & 0.114 & 0.694 \end{pmatrix}$$

the probability P_{12} that if the system was in the high-risk zone today, it would transit to low-risk zone the following day was approximately 0.09. In other words, given the current realities, if 9 deaths, on the average, were recorded today, the probability that the fatality rate would reduce to 1, on the average, tomorrow was just 9%. In 2 days' time, the probability increased to 17.4%:

$$\widehat{\mathrm{P}}^2 = \begin{pmatrix} 0.169 & 0.174 & 0.657 \\ 0.050 & 0.714 & 0.236 \\ 0.160 & 0.191 & 0.649 \end{pmatrix}$$

In 3 days' time, it increased to 23.5%:

$$\widehat{\mathrm{P}}^3 = \begin{pmatrix} 0.152 & 0.235 & 0.614 \\ 0.069 & 0.627 & 0.305 \\ 0.149 & 0.247 & 0.604 \end{pmatrix}$$

The value continued to improve until it gradually converged to 39.6%, and this convergence only occurred after 30 days:

$$\lim_{\mathrm{n}\to\infty} \widehat{\mathrm{P}}^{\mathrm{n}} = \begin{pmatrix} 0.118 & 0.396 & 0.487 \\ 0.118 & 0.396 & 0.487 \\ 0.118 & 0.396 & 0.487 \end{pmatrix}$$

Thus, given the state of health facilities and all forms of government's interventions, if the system was in the high-risk zone today, it was expected that the fatality rate would decrease to the barest minimum within 30 days with probability 0.396. In the same vein, the probability that the fatality rate would remain high could be computed in like manner.

Figure 3 below displayed the forecast probabilities that the system would remain in the high-, medium- and low-risk zones; and that it would transit from high-risk state to low-risk. We inferred from the Figure that in the nearest future, the most probable scenario in Nigeria was the medium-risk zone in which 4 persons, on the average, died of coronavirus on daily basis with 48.7% probability, followed by the low-risk zone in which 1 Nigerian, on the average, died of coronavirus on daily basis with 39.6%, probability and lastly the high-risk regime in which 9 persons, on the average, died of coronavirus on daily basis with 11.8% probability.

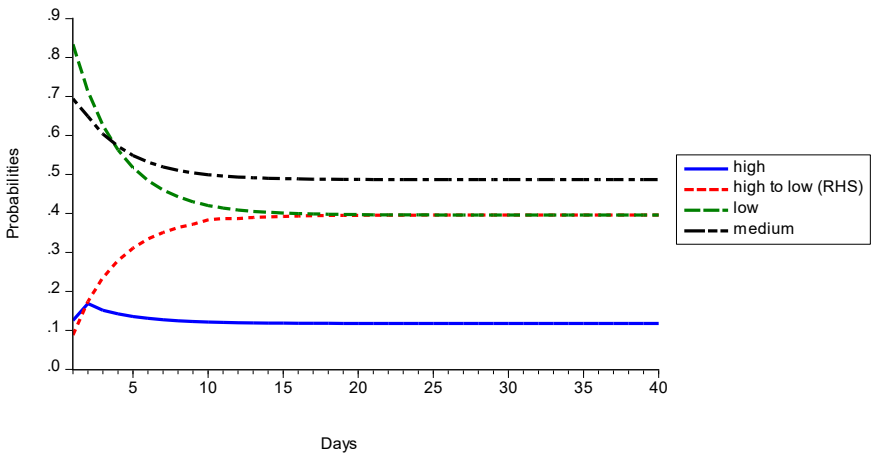


Figure 3: Forecast probabilities.

4.3. Discussion

We developed a Markov switching model for the coronavirus fatality rates in Nigeria. Our results showed that Nigerian death rates switched infrequently among three states of low-, medium- and high-risk within 2020 and 2022. In particular, the system remained in the medium-risk state in which an average of 4 (CI=2, 5) deaths/day were recorded most of the time in the fourth quarter of 2020 though pockets of low- and high-risk transitions were also recorded within the time. This finding agreed with Chigbu et al. (2021) who predicted an average of 3 deaths per day from August 25, 2020 to January 31, 2021.

Further, our results showed that the medium-risk regime was dominant from March 1, 2021 to March 27, 2021. This was in contrast with Li et al. (2022) who predicted an average of 14 (CI= 12, 16) deaths per day. Our results indicated that the system was in the medium-risk state in which an average daily mortality risk of 4 (CI=2, 5) was expected. It is noteworthy that the actual observations of new deaths reported by the NCDC within the stipulated period averaged 5 deaths per day, which supported our findings. .

We also observed that the high-risk regime dominated the third quarter of 2021 in which an estimated 9 persons were expected to die daily of coronavirus in Nigeria. Though Nigeria recorded its highest number of coronavirus casualties within the third quarter of 2021, these were significantly less than the number of fatalities predicted in South Africa within the same time frame; (see Mthethwa et al. (2022)). Thus South Africa was the indeed the epicenter of the virus in Africa.

As at 19th March, 2022, Nigeria no longer belonged to the high-risk group; our result showed that the country had transited to and settled in the low-risk regime in

which an average of 1 person was expected to die of coronavirus daily. This was an indication that individuals indeed adapted and learnt to live with the virus over time. We recall that Bello and Amzat (2021) had earlier observed that the government's decision to ease lockdown while the pandemic was still on followed Spencer's theory of "survival of the fittest". In addition, the fact that the mortality rate has significantly reduced was also an indication that the various government policies and interventions yielded the desired results. In addition, it also bore testament to the fact that quarantine measures could effectively flatten the curve as it did in developed countries such as Germany and South Korea (Jacobs and Okeke, 2022).

We noted earlier in section 2.2 that the virus caused huge disruptions to the social, economic and health aspects of human lives. It led to social order breakdown, limited human interpersonal relationships to virtual mode and also prevented people living with various non-communicable diseases to access treatments at their convenience. Ensuring that normalcy returned as soon as possible should be a priority, not for the government alone, but for everyone.

3. Conclusions

The Markov switching model was employed to model and forecast the fatality rate scenarios of coronavirus in Nigeria. The highlight of the results was that Nigeria, as at 19th March, 2022, which was earlier categorized by WHO as a high-risk country had transited from that state to a low-risk one in which 1 person was expected to die of coronavirus daily. This indicated that the various health intervention programs and policies instituted by the government to combat the virus yielded positive results; however, all things given, the most probable scenario in Nigeria was the medium-risk level in which an average of 4 persons would die daily; hence the need to sustain the fight against the virus.

The study concluded that Nigerian COVID mortality risks followed a switching pattern which fluctuated within low-, medium- and high-risks; however, the medium-risk state was most likely in the future. The findings from this study gave an accurate description of the fatality level which could assist in anticipating its progression in the future. It offered useful insights to aid policy makers in making evidence-based decisions and strengthen their efforts in reducing and eventually eradicating coronavirus deaths. In addition, it also assessed the level of success of health' interventions the government and health organizations have made so far and concluded that the efforts had yielded positive results. Our findings also underscored the need to sustain and intensify efforts with the quarantine measures that have been adopted and also embark on awareness programs for individuals to be more responsible for their safety so as to completely eradicate the virus.

References

Abdulmajeed, K., Adeleke, M., Popoola, L., (2020). Online forecasting of covid-19 cases in Nigeria using limited data. *Data in Brief* , Vol. 30, 105683.

Adekunle, A. I., Adegboye, O. A., Gayawan, E., McBryde, E. S., (2020). Is Nigeria really on top of COVID-19? Message from effective reproduction number. *Epidemiology and Infection*. Vol. 148, pp. 1–7, https://doi.org/10.1017/S0950268820001740.

Ahlburg, D., (1995). Simple versus complex models: evaluation, accuracy, and combining. *Math Popul Stud.*, Vol. 5, pp. 281–290, doi: 10.1080/08898489509525406.

Anne, W. R., (2020). ARIMA modelling of predicting COVID-19 infections, *medRxiv*, https://doi.org/10.1101/2020.04.18.20070631.

Bagal, D. K., Rath, A., Barua, A., Patnaik, D., (2020). Estimating the parameters of the susceptible-infected-recovered model of COVID-19 cases in India during lockdown periods. *Chaos Solitons Fractals*, https://doi.org/10.1016/j.chaos. 2020.110154.

Bello, B., Amzat, J., (2021). The Theory isn't dead: A Classical Sociological Gaze of COVID-19. *Tanzania Journal of Sociology*, Vol. 7, pp. 48–61.

Carcione, J. M., Santos, J. E., Bagaini, C., Ba, J., (2020). A simulation of a COVID-19 epidemic based on a deterministic SEIR model. *Front Public Health*, Vol. 8, https://doi.org/10.3389/fpubh. 2020.00230.

Chigbu, B. C., Edikpa, E. C., Onu, E. A., Nwabueze, A. I., Aneke, M. C., Vita-Agundu, U. C., Adepoju, E. B., (2021). Analysis and forecasting of confirmed, death, and recovered cases of COVID-19 infections in Nigeria: Implications for university administrators. *Medicine*, Vol. 100.

Didi, E. S., Kingdom, N., Harrison, E. E., (2021). ARIMA modeling and forecasting of COVID-19 daily confirmed/death cases: A case study of Nigeria. *Asian Journal of Probability and Statistics,* Vol. 12, pp. 59–80.

Engel, C., Hamilton, J. D., (1990). Long swings in the dollar: are they in the data and do markets know it? *The American economic review*, Vol. 80, pp. 689–713.

Hamilton, J. D., (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, Vol. 57, pp. 357–384.

Ibrahim, R. R., Oladipo, H. O., (2020). Forecasting the spread of COVID-19 in Nigeria using Box-Jenkins Modeling Procedure. *MedRxiv*, https://doi.org/10.1101/2020.05.05.20091686.

Ibrahim, S., Rasul, A., Ozigis, M. S., Adamu, B., (2021). Comparing the accuracies of forecasting models from the time series data of covid-19 infection in Nigeria. *European Journal of Public Health Studies*, Vol. 4, doi. 10.46827/ejphs.v4i2.106.

Ihekweazu, C., (2020). Steps Nigeria is taking to prepare for cases of coronavirus, http://theconversation.com/steps-nigeria-is-taking-to-prepare-for-cases-of-coronavirus-130704.

Ileyemi, M., (2021). COVID-19: Nigeria records 93 deaths, 362 new cases. Premium Times, Nigeria 30 August, 2021.

Jacobs, E., Okeke, M., (2022). A critical evaluation of Nigeria's response to the first wave of COVID-19. *Bull. Natl. Res. Cent.*, Vol. 46.

Khan, F., Lounis, M., (2021). Short‐term forecasting of daily infections, fatalities and recoveries about COVID‐19 in Algeria using statistical models. *Beni-Suef University Journal of Basic and Applied Sciences*, Vol. 10, https://doi.org/10.1186/s43088-021-00136-5.

Li, C., Sampene, A. K., Agyeman, F. O., Robert, B., Ayisi, A. L., (2022). Forecasting the Severity of COVID-19 Pandemic Amidst the Emerging SARS-CoV-2 Variants: Adoption of ARIMA Model. *Computational and Mathematical Methods in Medicine*, Vol. 2022, https://doi.org/10.1155/2022/3163854.

Marbot, O., (2020). Coronavirus Africa Map: Which Countries are Most at Risk? https://www.theafricareport.com/23948/coronavirus-africa-which-countries-are-most-at-risk/.

Mthethwa, N., Chifurira, C., Chinhamu, K., (2022). Estimating the risk of SARS‐CoV‐2 deaths using a Markov switching‐volatility model combined with heavy‐tailed distributions for South Africa. *BMC Public Health*, Vol. 22.

Odekina, G. O., Adedotun, A. F., Imaga, O. F., (2022). Modeling and Forecasting the Third wave of Covid-19 Incidence Rate in Nigeria Using Vector Autoregressive Model Approach. *Journal of the Nigerian Society of Physical Sciences*, Vol. 4, pp. 117–122.

Ogundokun, R. O., Lukman, A. F., Kibria, G. B., Awotunde, J. B., Aladeitan, B. B., (2020). Predictive modelling of COVID-19 confirmed cases in Nigeria. *Infectious Disease Modelling*, Vol. 5, pp. 543–548.

Oliveira, A., Binner, J., Mandal, A., Kelly, L., Power, G., (2021). Using GAM functions and Markov-Switching models in an evaluation framework to assess countries' performance in controlling the COVID-19 pandemic. *BMC Public Health*, Vol. 21.

Pourghasemi, H. R., Pouyan, S., Farajzadeh, Z., Sadhasivam, N., Heidari, B., Babaei, S., Tiefenbacher, J. P., (2020). Assessment of the outbreak risk, mapping and infestation behavior of COVID-19: application of the autoregressive integrated and moving average (ARIMA) and polynomial models, *PloSOne*, Vol. 15, https://doi.org/10.1371/journal.pone.0236238.

Sharifi, Y., Ebrahimpur, M., Payab, M., Larijani, B., (2022) The Syndemic Theory, the COVID-19 Pandemic, and The Epidemics of Non-Communicable Diseases (NCDs). *Medical Journal of the Islamic Republic of Iran*, Vol. 36.

Singh, R. K., Rani, M., Bhagavathula, A. S., Sah, R., Rodriguez-Morales, A. J., KalitaH, N. C., Sharma, S., Sharma, Y. D., Rabaan, A. A., (2020). Prediction of the covid-19 pandemic for the top 15 affected countries: advanced autoregressive integrated moving average (arima) model, *JMIR Public Health Surveill.*, Vol. 6, https://doi.org/10.2196/19115.

Taleb, N., (2007). The Black Swan: The Impact of the Highly Improbable. New York: Random House, vol. 2, https://innovation.cc/book-reviews/2008_13_3_14_gow_bk-rev_taleb_black-swan.

Worldometer, (2022) Nigeria COVID: Coronavirus Statistics - Worldometer, April 2022, https://www.worldometers.info/coronavirus/country/nigeria/.

Appendices

We compared three specifications of MSM, namely the 1-state case (which corresponded to the conventional autoregressive model), 2-state case and 3-state case using various statistical techniques including (i) information criteria; (ii) error measurements (the maximum absolute error and the mean square error); and (iii) visual aid.

Table A1 comparing the models by information criteria showed that the 3-state case performed better than the other two specifications as it had the lowest Akaike and Bayesian information criteria values.

Table A1: Model comparison by information criteria

Model	AIC	BIC
1-state	6.445	6.458
2-states	5.025	5.075
3-states	4.663	4.757

Further we compared the three models by prediction errors from in-sample we assumed that the system is in the high-risk state whenever P_{11} exceeded 0.5, at such times, coronavirus death rates in Nigeria were best described by the equation $\hat{d}_t = 8.16 + 1.07d_{t-1}$; at other times when P_{33} exceeded 0.5, we assumed that the system was in the medium-risk state and was best modeled as $\hat{d}_t = 3.33 + 0.26d_{t-1}$. Finally when P_{22} exceeded 0.5, the corresponding state space was the low-risk state and in such cases, $\hat{d}_t = 0.16 + 0.01d_{t-1}$ provided the best fit for the system. Table A2 showed prediction error analysis by model. We observed that the 3-state case again provided the best fit for the death rates.

Table A2: Model comparison by prediction errors

Model	Max. abs. error	Mean square error
1-state	66.738	36.662
2-states	61.436	28.240
3-states	18.544	11.780

In addition, we plotted the observed against the predictions from the 3 candidate models in Figure A1 below. In agreement with the results presented in Tables A1 and A2, we observed that the 3-state model provided the best fit to the observed data. In all the daily predictions, the 1-state model could not reproduce the minimum death rate as its minimum estimated value was 2.38.

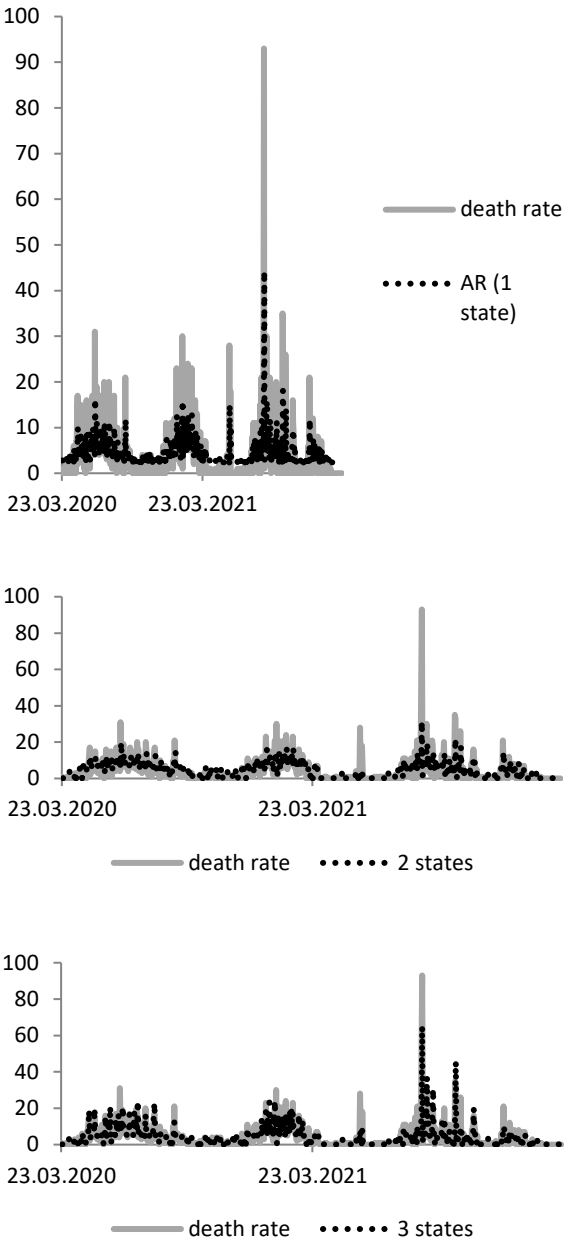


Figure A1: Predicted versus observed by models

STATISTICS IN TRANSITION new series, September 2024
Vol. 24, No. 3, pp. 141–154, https://doi.org/10.59170/stattrans-2024–032
Received – 28.07.2023; accepted – 25.04.2024

Nonparametric Bayesian optimal designs for Unit Exponential regression model with respect to prior processes(with the truncated normal as the base measure)

Anita Abdollahi Nanvapisheh[1], **Soleiman Khazaei**[2], **Habib Jafari**[3]

Abstract

Nonlinear regression models are extensively applied across various scientific disciplines. It is vital to accurately fit the optimal nonlinear model while considering the biases of the Bayesian optimal design. We present a Bayesian optimal design by utilising the Dirichlet process as a prior. The Dirichlet process serves as a fundamental tool in the exploration of Nonparametric Bayesian inference, offering multiple representations that are well-suited for application. This research paper introduces a novel one-parameter model, referred to as the 'Unit-Exponential distribution', specifically designed for the unit interval. Additionally, we employ a stick-breaking representation to approximate the D-optimality criterion considering the Dirichlet process as a functional tool. Through this approach, we aim to identify a Nonparametric Bayesian optimal design.

Key words: D-optimal design, Bayesian optimal design, Unit Exponential model (UE), Dirichlet process, stick-breaking prior, nonparametric Bayesian.

1. Introduction

Within the realm of experimental design, the concept of optimal design refers to a specific category of designs that are classified based on certain statistical criteria. It is widely acknowledged that a well-designed experiment can significantly enhance the accuracy of statistical analyses. Consequently, numerous researchers have dedicated their efforts to address the challenge of constructing optimal designs for nonlinear regression models. Experimental design plays a pivotal role in scientific research domains, including but not limited to biomedicine and pharmacokinetics. Its application in these fields enables researchers to conduct rigorous investigations and yield valuable insights.

Optimal designs are sought using optimality criteria, typically based on the information matrix. Until 1959, research primarily focused on linear models, where the models were linear with respect to the parameters. However, in nonlinear models, the presence of unknown parameters introduced complexities in the design problem, as the optimality criteria depended on these unknown parameters [3, 5]. To address this challenge, researchers

[1]Department of Statistics, Razi University, Kermanshah, Iran. E-mail: anita.abdollahi@yahoo.com.
ORCID: https://orcid.org/0000-0003-3248-0347.
[2]Department of Statistics, Razi University, Kermanshah, Iran. E-mail: s.khazaei@razi.ac.ir.
ORCID: https://orcid.org/0000-0003-2537-9232.
[3]Department of Statistics, Razi University, Kermanshah, Iran. E-mail: h.jafari@razi.ac.ir.
ORCID: https://orcid.org/0000-0001-5191-2796.

proposed various solutions, including local optimal designs [2, 7, 11, 19, 30], sequential optimal designs, minimax optimal designs, Bayesian optimal designs [28, 21-24], and pseudo-Bayesian designs [26]. Chernoff (1953) introduced the concept of local optimality, which involves specifying fixed values for the unknown parameters and optimizing a function of the information matrix to determine the design for these specified parameter values. This approach aimed to overcome the difficulties associated with the dependence of the design problem on unknown parameters in nonlinear models.

The selection of unknown parameters in local designs is typically obtained from previous studies or experiments specifically conducted for this purpose. The effectiveness of local designs heavily relies on the appropriate selection of these parameters. However, a significant challenge arises when the investigated problem lacks robustness in relation to weak parameter estimation. To address this, an alternative approach for local optimal designs involves utilizing a prior distribution for the unknown parameters instead of relying solely on initial guess. In the Bayesian method, the first step is to represent the available information in the form of a probability distribution for the model parameter, known as the prior distribution. A Bayesian optimal design aims to maximize the relevant optimality criterion over this prior distribution. Nevertheless, it is crucial to acknowledge that the selection of the prior distribution within the Bayesian framework can be problematic and may potentially lead to erroneous results. The choice of the prior distribution is subjective, relying on the researcher's beliefs, and it significantly influences the final outcome. Unfortunately, the Bayesian approach lacks a definitive method for selecting the prior distribution. Numerous researchers have investigated the effect of the prior distribution on determining design points in various types of optimal designs. For instance, Chaloner and Lorentz [10], Chaloner and Duncan [8], Burghaus and Dette [4], Chaloner and Vardinelli [9], Pronzato and Walter [29], Mukhopadhyay and Haines [26], Dette and Ngobauer [12, 13], Fedorov [14, 15], and Firth and Hinde [17] have contributed extensively to this field. Chapter 18 of Atkinson et al.'s book [3] provides further reading on this topic. Moreover, in situations where there is insufficient evidence from previous studies on the topic of interest, specifying an appropriate prior distribution becomes challenging. In such cases, subjective or noninformative prior distributions are used, incorporating all available information regarding the uncertainty of the parameter values.For more information, refer to Burghaus and Dette [4]. This research paper presents the introduction of a novel one-parameter model, referred to as the UE distribution, specifically designed for the unit interval in Section 2. As we know, in applied statistic, a common issue is to deal with the uncertainty phenomena observed in the interval (0, 1). For example, in real life we often encounter measures like proportion or fraction of a certain characteristic, scores of some ability tests, different index, rates, etc., which lie in the interval (0, 1). In such cases continuous distributions with domain (0, 1) are indispensable to probabilistic modeling of the phenomena. So, in regression models where the response variable is in the form of ratio, rate or percentage, we use the unit exponential regression model to model the data that are concentrated in a certain sub-interval of the range of their domains. In Section 3, the optimal design for nonlinear models is derived. Finally, Section 4 concludes the paper with some closing remarks.

2. The Unit-Exponential distribution

The exponential distribution is continuous distribution in statistics and probability theory. If $Y \sim \text{Exp}(\theta)$, then using the transformation $X=\frac{Y}{1+Y}$ we have a new distribution with support on the unit interval such that the CDF and the PDF of the resulting distribution are respectively [1]:

$$F(x \mid \theta) = 1 - \exp(\frac{-\theta x}{1-x});\ 0 \leq x < 1,\ \theta > 0, \tag{1}$$

$$f(x \mid \theta) = \frac{\theta}{(1-x)^2} \exp(\frac{-\theta x}{1-x});\ 0 \leq x < 1,\ \theta > 0. \tag{2}$$

The Hazard Rate Function (HRF) of this distribution is as follows:

$$h(x \mid \theta) = \frac{f(x \mid \theta)}{1 - F(x \mid \theta)} = \frac{\theta}{(1-x)^2};\ 0 \leq x < 1,\ \theta > 0. \tag{3}$$

In the following figure, the PDF and the HRF of this distribution are plotted for different values of the parameter θ. According to this figures, it can be seen that the HRF is increasing in $0 \leq x < 1$.

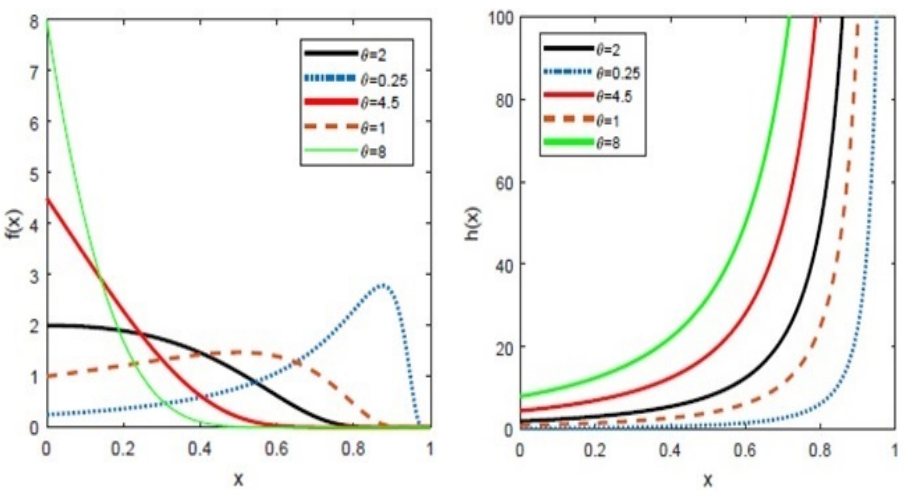


Figure 1: Plot of density function (left) and hrf (right)

3. Optimal Design for Nonlinear Models

In the context of nonlinear experimental design, a common issue arises where the relationship between the response variable y and the independent variable x is given by the equation $y = \eta(x, \theta) + \varepsilon$ where $x \in \chi \subseteq \text{R}$ and y is a response variable and $\theta \in \Theta$ is the

unknown parameter vector and ε is a normally distributed residual value with mean 0 and known variance $\sigma^2 > 0$. For simplicity, we assume $\sigma^2 = 1$ in this problem. If $\eta(x,\theta)$ is differentiable with respect to θ, then the information matrix at a given point x can be represented as follows:

$$I(\xi,\theta) = \frac{\partial}{\partial\theta}\eta(x,\theta)\frac{\partial}{\partial\theta^T}\eta(x,\theta). \tag{4}$$

There exist several optimality criteria used to obtain the optimal design, including D-optimality and A-optimality. These criteria are functions of the information matrix and can be expressed as follows:

$$\Psi_D(\xi,\theta) = -\log(det(M(\xi,\theta)))\ ,\ \Psi_A(\xi,\theta) = tr(M^{-1}(\xi;\theta)),$$

where ξ denotes a design with two components; the first component represents specific values from the design space χ and the second component corresponds to the weights assigned to these values, so that design ξ can be defined as follows:

$$\xi = \left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_l \\ w_1 & w_2 & \dots & w_l \end{array} \right\} \in \Xi, \tag{5}$$

where Ξ=$\{\xi \mid 0 \le w_j \le 1\ ;\sum_{j=1}^{l} w_j = 1\ , x \in \chi\}$, [25].

When considering a discrete probability measure ξ with finite support, the information function of ξ can be expressed as follows [3]:

$$M(\xi,\theta) = \sum_{j=1}^{l} w_j I(x_j,\theta). \tag{6}$$

Because of the dependence of the information matrix $M(\xi,\theta)$ to the unknown parameter θ, one approach to address this issue is to employ the Bayesian method and incorporate a prior distribution for the parameter vector. The Bayesian D-optimality criterion can be formulated as follows:

$$\Psi_\Pi(\xi) = E(\psi(\xi;\theta)) = \int_\Theta \psi(\xi;\theta)d\Pi(\theta) = \int_\Theta -log(det(M(\xi,\theta))d\Pi(\theta), \tag{7}$$

where Π represents the prior distribution for θ and the Bayesian D-optimal design is attained by minimizing (7). According to Dette and Neugebauer [11], in the general case of optimal designs which can include designs with two and more points, if the support of the prior distribution has n points, then the maximum number of Bayesian optimal design points is given by $n\ \frac{p(p+1)}{2}$. Hence, in the specific scenario of nonlinear models with one parameter ($p = 1$), this implies that the support of the Bayesian optimal design does not contain more points than the support of the prior distribution.

In certain situations, specifying a prior distribution on the parameter space Θ can be challenging for the experimenter. In such cases, an alternative approach is to consider an

unknown prior distribution Π for the parameter θ. In this condition, Π is treated as a parameter itself. Consequently, equation (7) becomes a random functional, and it becomes necessary to determine its distribution or approximation. From a Bayesian perspective, we construct a prior distribution on the space of all distribution functions to address this issue. Ferguson (1973) introduced the concept of the Dirichlet process in this context, an overview of which will be provided in Section 3.1.1.

3.1. Nonparametric Bayesian D-optimal design

In this section, we introduce the nonparametric Bayesian optimal designl. In the nonparametric Bayesian framework, it is assumed that $\theta \mid P \sim P$, where P is a random probability distribution and $P \sim \Pi$. A general method of construction of a random measure is to start with the stochastic processes. Ferguson (1973) formulated the requirements which must be imposed on a prior distribution and proposed a class of prior distributions, named the Dirichlet processes. One of the main argument in using the Dirichlet distribution in practical applications is based on the fact that this distribution is a good approximation of many parametric probability distributions. Below we give the definition of the Dirichlet process.

3.1.1 Dirichlet Process (DP)

To have a random distribution G distributed according to a Dirichlet process (DP), its marginal distributions must follow a Dirichlet distribution. Specifically, let H be a distribution over Θ and α be a positive real number. For any finite measurable partition $A_1, A_2, ..., A_r$ of Θ the vector $(G(A_1), G(A_2), ..., G(A_r))$ is random since G is random. We say G is the Dirichlet process distributed with base distribution H and concentration parameter α, written $G \sim \mathrm{DP}(\alpha, H)$, if the following conditions hold:

$$(G(A_1), G(A_2), ..., G(A_r)) \sim Dir(\alpha H(A_1), ..., \alpha H(A_r)), \tag{8}$$

for every finite measurable partition $A_1, A_2, ..., A_r$ of Θ.

The parameters of H and α play intuitive roles in the definition of the DP. The base distribution H represents the mean of the Dirichlet process, such that for any measurable set $A \subset \Theta$ we have $E[\mathrm{G(A)}] = H(A)$. On the other hand, the concentration parameter α can be viewed as an inverse variance: $V[G(A)] = H(A)(1 - H(A))/(\alpha + 1)$. The larger α is, the smaller the variance, and the DP will concentrate more of its mass around the mean. The concentration parameter is also referred to as the strength parameter, referring to the strength of the prior when using the DP as a nonparametric prior in Bayesian nonparametric modelsl, It can be interpreted as the amount of mass or sample size associated with the observations. It is worth noting that α and H only appear as their product in the definition of the Dirichlet process (equation 8). Consequently, some authors treat $\tilde{\mathrm{H}} = \alpha H$, as the single (positive measure) parameter of the DP, writing $\mathrm{DP}(\tilde{\mathrm{H}})$ instead of $\mathrm{DP}(\alpha, H)$. This parametrization can be notationally convenient, but loses the distinct roles α and H play in describing the DP.

As the concentration parameter α increases, the mass of the DP becomes more concentrated around its mean. Consequently, when α approaches infinity ($\alpha \to \infty$), $G(A)$ approaches $H(A)$ for any measurable set A, indicating weak or pointwise convergence of G to H. However, it is important to note that this does not imply a direct convergence of G to H as a whole. In fact, as we will explore later, samples drawn from the DP will typically be discrete distributions with probability one, even if the base distribution H is smooth. Therefore, G and H may not be absolutely continuous with respect to each other. Despite this, some authors still utilize the DP as a nonparametric extension of a parametric model represented by H. However, if the desire is to maintain smoothness, it is possible to extend the DP by convolving G with kernels, resulting in a random distribution with a density function.

An alternative definition of the Dirichlet process is proposed by Ferguson [16], who, defined a random probability measure, which is a Dirichlet process on (Θ, B(Θ)), as:

$$P(.) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(.), \tag{9}$$

where θ_i $(i > 1)$ is a sequence of $i.i.d.$ random variables with common distribution Q, δ_{θ_i} represents a probability measure that is degenerate at θ where δ_{θ_i}=1 if $\theta_i \in A$ and 0 otherwise, and p_i's are the random weights satisfying p_i>0 and $\sum_{i=1}^{\infty} p_i$=1. The random distribution P is discrete with probability one. Several authors have proposed alternative series representations of the Dirichlet process. Sethuraman [31], and Zarepour and Al Labadi [32] are among those who have contributed to this area. In the upcoming section, we will discuss the nonparametric Bayesian D-optimal design for the UE model.

3.1.2 Nonparametric Bayesian D-optimal design for UE model

Now, let us consider the following regression model:

$$E(y|x) = \eta(x,\theta) = \frac{\theta}{(1-x)^2} \exp(\frac{-\theta x}{1-x}), \theta > 0. \tag{10}$$

Therefore, the Bayesian D-optimality criterion, denoted as $\Psi_\Pi(\xi)$, can be expressed as follows:

$$\Psi_\Pi(\xi) = E(\psi(\xi;\theta)) = \int_\Theta \psi(\xi;\theta) d\Pi(l\theta) \tag{11}$$

$$= \int_\Theta -log(\sum_{j=1}^{l} w_j [\exp(\frac{-\theta x_j}{1-x_j})(\frac{1}{(1-x_j)^2} - \frac{\theta x_j}{(1-x_j)^3})]^2) d\Pi(\theta) \tag{12}$$

where Π is the prior distribution for θ. The Bayesian D-optimal design is attained by minimizing equation (11). In the nonparametric Bayesian framework, we consider $P \sim \mathrm{DP}(\alpha, P_0)$ and its collective representation as $P(.) = \sum_{i=1}^{\infty} p_i \, \delta_{\theta_i}(.)$. In this context, the optimality criterion can be expressed as follows:

$$\Psi_{\Pi}(\xi) = \sum_{i=1}^{\infty} p_i(-log(\sum_{j=1}^{l} w_j[\exp(\frac{-\theta_i x_j}{1-x_j})(\frac{1}{(1-x_j)^2} - \frac{\theta_i x_j}{(1-x_j)^3})]^2)). \tag{13}$$

Chernoff [7] demonstrated that when searching for a local optimal design, there exists an optimal design where all the mass is concentrated at a single point within the design's support. Caratheodory's theorem also confirms the existence of a one-point optimal design. However, when employing the Bayesian optimality criterion, a more complex situation arises. Braess and Dette showed that with a uniform prior distribution, as the support of the prior distribution increases, the number of optimal design points for the single-parameter model also increases. Challoner suggested that if the researcher aims to obtain a one-point optimal design, it is advisable to consider a small support for the uniform prior distribution. The same principle applies to nonparametric Bayesian designs. In this case, assuming a uniform distribution over the interval [1, B] as the basic distribution, the one-point optimal design can be achieved.

Equation (11) represents a stochastic function of the Dirichlet process. According to Ferguson's definition of the Dirichlet process, the direct calculation of (12) is not straightforward. To address this challenge and obtain an approximation of the optimal nonparametric Bayesian criterion, methods such as the stick-breaking process is employed [31]. Saturaman (1994) introduced this method as a significant approach for generating realizations of the Dirichlet process, which we will explain below. Additionally, we will highlight the discreteness of the Dirichlet process within the framework of the stick-breaking process. To generate a realization of the Dirichlet process P with a concentration parameter α and base distribution H we can follow the stick-breaking process.

First, we generate a sequence of random samples $\theta_1, \theta_2, ...$ from the base distribution H. Additionally, we generate a sequence of random samples $V_1, V_2, ...$ from the $Beta(1, \alpha)$ distribution. We define a sequence of probabilities $p_1, p_2, ..., p_k, ...$ as follows. We start by choosing a point called V_1 on a unit-length piece of wood and set p_1 equal to V_1. In other words, $p_1 = V_1$. Then, we divide the remaining part of the wood into two parts, $V_2(1-V_1)$ and $(1-V_1)(1-V_2)$. We consider the first part as p_2. To calculate p_3, we divide the remaining part of the wood into two parts in the same manner as in step 2. We continue this process, dividing each remaining part into two parts and assigning the first part as the next weight in the sequence. By following these steps, we obtain a sequence of weights $p_1, p_2, ..., p_k, ...$ that represents the probabilities associated with the generated samples $\theta_1, \theta_2, ...$ This sequence of weights reflects the stick-breaking process used to approximate the Dirichlet process. So:

$$\begin{aligned} p_1 &= V_1, \\ p_i &= V_i \prod_{j=1}^{i-1} (1-V_j), \quad i \geq 2 \end{aligned}$$

According to the structure described, it can be proven that $\sum_{i=1}^{\infty} p_i$=1. For this purpose we have:

$$1-\sum_{i=1}^{\infty} p_i = 1-V_1-V_2(1-V_1)-V_3(1-V_2)(1-V_1)-...$$

$$=(1-V_1)(1-V_2-V_3(1-V_2)-...)$$

$$\vdots$$

$$=\prod_{i=1}^{\infty}(1-V_i) \tag{14}$$

By problem 32 in Chapter 1 of Folland (1999), we have [18]:

$$\prod_{i=1}^{\infty}(1-V_i)=0 \Leftrightarrow \sum_{i=1}^{\infty} V_i=\infty$$

So, for every $\varepsilon \in (0,1)$ we can write the following relation:

$$\sum_{i=1}^{\infty} Pr(V_i>\varepsilon)=\infty$$

And using Borel-Cantelli's Lemma, we will have:

$$Pr(V_i>\varepsilon, i.o)=1 \Rightarrow \sum_{i=1}^{\infty} V_i=\infty \quad a.s \tag{15}$$

Therefore, by setting the relation (3.10) equal to zero, we will have $\sum_{i=1}^{\infty} p_i$=1.

In this section, we focus on the use of a truncated normal distribution as the base measure in the DP. To obtain the results, we employ nonlinear optimization programming using the R package Rsolnp. The nonparametric Bayesian optimal designs are examined using the stick-breaking method, and tables presenting the results are provided. To better understand the influence of the α parameter, we present the results for four different values of α=1, 5, 10, 50. It is important to note that we consider a bounded design space χ=[0,1] without any loss of generality. Tables 4-7 display the results obtained when the concentration parameter (α) and uncertainty in the base measure increase. Based on these results, we can observe in the class of two-point design, that largest weight corresponds to the smallest point. This pattern is consistent across the investigated range of α values. According to the results, when the value of α increases, the support points in two-point design do not significantly change. The smallest point will have the most weight that this weight almost increases or remains fixed by increasing the concentration parameter. Also, for three-point design, minimum support point has the greatest weight. In addition, in the range under investigation, the results show that we do not have a three-point design for $\mu=5, \sigma=2$, and in fact, it converts to the design by less points. This observation is more clear for larger concentration parameter. But, by increasing the parameter space, optimal two and three-point design are obtained.

Table 1: Nonparametric Bayesian D-optimal designs with truncated normal base distribution and concentration parameter when α=1. First row: support points; second row: weights.

Parameters	*Design*	*Two − point*		*Three − point*		
$\mu = 5, \sigma = 2$	x	0.0000000	0.3419671	–	–	–
	w	0.9999995	0.0000005	–	–	–
$\mu = 50, \sigma = 30$	x	0.0245942	0.2728781	0.0341488	0.2776520	0.5086153
	w	0.9696928	0.0303072	0.9789477	0.0210526	0.0000007
$\mu = 150, \sigma = 90$	x	0.0065494	0.2995342	0.01578005	0.2997138	0.5013164
	w	0.9999903	0.0000097	0.9999994	0.0000003	0.0000003

Table 2: Nonparametric Bayesian D-optimal designs with truncated normal base distribution and concentration parameter when α=5. First row: support points; second row: weights.

Parameters	*Design*	*Two − point*		*Three − point*		
$\mu = 5, \sigma = 2$	x	0.0000000	0.3233669	–	–	–
	w	0.9999995	0.0000005	–	–	–
$\mu = 50, \sigma = 30$	x	0.0204877	0.2772758	0.03387816	0.2638516	0.5001318
	w	0.9799968	0.0200032	0.9494947	0.0505048	0.0000005
$\mu = 150, \sigma = 90$	x	0.0009694	0.2993877	0.01462319	0.3000076	0.4991483
	w	0.9999854	0.0000146	0.9999999	0.0000004	0.0000004

Table 3: Nonparametric Bayesian D-optimal designs with truncated normal base distribution and concentration parameter when α=10. First row: support points; second row: weights.

Parameters	*Design*	*Two − point*		*Three − point*		
$\mu = 5, \sigma = 2$	x	0.0000000	0.3021963	–	–	–
	w	0.9999995	0.0000005	–	–	–
$\mu = 50, \sigma = 30$	x	0.0156330	0.2706337	0.0257019	0.2071970	0.5050722
	w	0.9898957	0.0101043	0.9265122	0.0734868	0.0000010
$\mu = 150, \sigma = 90$	x	0.0006769	0.2990424	0.0126487	0.2992510	0.5007835
	w	0.9863551	0.0136449	0.9999868	0.0000135	0.0000007

Table 4: Nonparametric Bayesian D-optimal designs with truncated normal base distribution and concentration parameter when α=50. First row: support points; second row: weights.

Parameters	*Design*	*Two − point*		*Three − point*		
$\mu = 5, \sigma = 2$	x	0.0000000	0.3030561	–	–	–
	w	0.9999995	0.0000005	–	–	–
$\mu = 50, \sigma = 30$	x	0.0132530	0.2859840	0.0236265	0.2357064	0.5016003
	w	0.9999973	0.0000027	0.9361683	0.06383056	0.0000001
$\mu = 150, \sigma = 90$	x	0.0000608	0.2990344	0.0107339	0.2991683	0.5012125
	w	0.9999865	0.0000135	0.9999959	0.0000020	0.0000021

Table 5 presents the results when assuming a constant mean of the base distribution and increasing the variance. Specifically, in the two-point designs, it can be observed that the smallest point has the highest weight. This table provides insights into the distribution of weights in this scenario.

Table 5: Nonparametric Bayesian D-optimal designs with truncated normal base distribution and concentration parameter when α=1. First row: support points; second row: weights.

Parameters	*Design*	*Two points*		*Three − point*		
$\mu = 50, \sigma = 2$	x	0.0000000	0.3000000	–	–	–
	w	0.9999942	0.0000059	–	–	–
$\mu = 50, \sigma = 30$	x	0.0237781	0.2842176	0.0384189	0.2794133	0.5005586
	w	0.9795880	0.0204120	0.9587626	0.04123712	0.0000002
$\mu = 50, \sigma = 90$	x	0.0108601	0.2875706	0.0257537	0.2810997	0.4938304
	w	0.9899937	0.0100063	0.9791666	0.02083332	0.0000002

4. Concluding Remarks and Future Works

Nonlinear regression models are widely used in various scientific fields, and the Bayesian method is commonly employed to obtain optimal designs in such models. However, one of the challenges in the Bayesian framework is the subjective selection of the prior distribution, which can potentially lead to incorrect results. The choice of the prior distribution is often based on the researcher's beliefs, and it strongly influences the final outcome. Unfortunately, the Bayesian approach lacks a systematic method for selecting the prior distribution. To overcome these limitations and reduce reliance on restrictive parametric assumptions, nonparametric Bayesian methods are pursued. In this study, we consider the prior distribution as an unknown parameter and utilize the Dirichlet process to derive nonparametric

Bayesian D-optimal designs. Specifically, we focus on a nonlinear model with one parameter, namely the Unit-Exponential distribution. We investigate the Bayesian D-optimal design for the unit exponential regression model (equation 10) using a truncated normal prior distribution, examining various parameter values. By adopting a nonparametric Bayesian approach and utilizing the Dirichlet process, we aim to address the challenges associated with selecting the prior distribution in Bayesian optimal design construction. This allows us to account for uncertainty and mitigate the impact of restrictive parametric assumptions, providing more flexible and robust designs for nonlinear regression models.

In this study, we focus on utilizing the Polya Urn Scheme as the base distribution in the Dirichlet process. To better understand the influence of the concentration parameter α, we present the results in tables for four different values of α=1, 5, 10, 50. These tables provide valuable insights into the nonparametric Bayesian optimal designs, showcasing the distribution of weights and support points. By analyzing the results for different values of α, we can better understand the impact of this parameter on the design outcomes. This approach allows us to explore and evaluate the performance of the nonparametric Bayesian optimal designs under varying levels of concentration parameter α.

In the investigated range, the results reveal interesting findings. For small parameter values, there are no two-point designs observed. However, by increasing uncertainty in the base measure, another optimal point is obtained with a very small weight, resulting in a design where the smallest point has the highest weight. These designs can be considered as one-point designs, as the weight of the additional point becomes negligible.

In three-point designs, similar observations can be made. In some cases, two of the obtained optimal points are very similar, leading to a design with fewer points. This indicates that the additional point does not significantly contribute to the design in such cases.

Moreover, as the uncertainty in the base measure and the concentration parameter in the Dirichlet process increase, the support points in the two-point designs do not undergo significant changes. The weight of the smallest point increases rapidly, and it becomes the point with the highest weight. This weight tends to either increase or remain relatively stable with an increase in the concentration parameter.

It is important to note that this approach can be applied to other optimality criteria and various models with two or more parameters. For example, nonparametric Bayesian optimal designs using the A- or E-optimality criterion for the nonlinear model discussed in this paper, along with a Dirichlet process prior, hold potential for further research. We hope to report new results in this area in the near future.

Acknowledgements

Thanks to anyone for support, funding and such may be included in the non-numbered Acknowledgements section.

References

Abdollahi, A., H. Jafari and S.Khazaei, (2024). Locally, Bayesian and Nonparametric Bayesian optimal designs for Unit Exponential regression model. *Journal of Communications in Statistics – Theory and Methods*, doi.org/10.1080/03610926.2024.2328182.

Aminnejad, M., Jafari H., (2017). Bayesian A and D-optimal designs for gamma regression model with inverse link function. *Communications in Statistics-Simulation and Computation*, 46 (10), pp. 8166–89, doi:10.1080/03610918.2016.1271888.

Atkinson, A. C., A. N. Donev and R. D. Tobias, (2007). Optimum experimental design, with SAS. *Oxford*, UK: Oxford University Press.

Burghaus, I., Dette H., (2014). Optimal designs for nonlinear regression models with respect to non-informative priors. *Journal of Statistical Planning and Inference*, 154, pp. 12–25, doi:10.1016/ j.jspi.2014.05.009.

Burkner P. C., Schwabe R., Holling H., (2019). Optimal designs for the generalized partial credit model. *British Journal of Mathematical and Statistical Psychology*, doi:10.1111/ bmsp. 12148.

Braess, D., Dette, H., (2007). On the number of support points of maximin and Bayesian optimal designs. *The Annals of Statistics*, 35(2), pp. 772–792.

Chernoff, H., (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 24 (4), pp. 586–602, doi:10.1214/aoms/1177728915.

Chaloner, K. M., Duncan, G. T., (1983). Assessment of a beta prior distribution:PM elicitation. *The Statistician*, pp. 174–180.

Chaloner, K., Verdinelli, I., (1995). Bayesian experimental design: a review. *Statistical Science*, 10, pp. 273–304.

Chaloner, K., Larntz, K., (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21, pp. 191–208.

Dette, H., Melas, V. B. and Wong, W. K., (2006). *Locally D-optimal designs for exponential regression Statistica Sinica*, 16, pp. 789–803.

Dette, H., Neugebauer, H. M., (1996). Bayesian optimal one point designs for one parameter nonlinear models. *Journal of Statistical Planning and Inference*, 52(1), pp. 17–31.

Dette, H., Neugebauer, H. M., (1997). Bayesian D-optimal designs for exponential regression models. *Journal of Statistical Planning and Inference*, 60(2), pp. 331–349.

Fedorov V, Hackl P., (2012). Model-Oriented Design of Experiments. *Springer Science and Business Media*, Vol. 125, doi:10.1007/978-1-4612-0703-0.

Fedorov V. V., Leonov S. L., (2013). Optimal design for nonlinear response models. *CRC Press. bability Letters*, 82 (5), pp. 916–24, doi:10.1016/j.spl.2012.01.020.

Ferguson, T. S., (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1 (2), pp. 209–30, doi:10.1214/aos/1176342360.

Firth D, Hinde J., (1997). On Bayesian D-optimum Design Criteria and the Equivalence Theorem in Non-linear Models. *Journal of the Royal Statistical Society B*, 59(4), pp. 793–797, doi:10.1111/1467-9868.00096.

Folland, G. B., (1999). Real analysis: modern techniques and their applications. John Wiley and Sons.

Ford, I., Torsney, B. and Wu, C. F. J., (1992). The use of a canonical form in theconstruction of locally optimal designs for non-linear problems. *J. Roy. Statist. Soc. (Ser.) B*, 54, pp. 569–583.

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

Goudarzi, M., H. Jafari and Khazaei S., (2019). Nonparametric Bayesian optimal designs for exponential regression model. *Communications in Statistics-Simulation and Computation*, doi:10.1080/03610918.2019.1593454.

Grashoff U., Holling H., Schwabe R., (2012). Optimal designs for the Rasch model. *Psychometrika*, 77(4), pp. 710–723, doi:10.1007/s11336-012-9276-2.

Kiefer, J., (1959). Optimum experimental designs. *J. R. Statist. Soc. B.*, 21, pp. 272–319.

Kiefer, J., Wolfowitz, J., (1959). Optimum designs in regression problems. *The Annals of Mathematical Statistics*, 30(2), pp. 271–294.

Kiefer, J., (1974). General equivalence theory for optimum designs (approximate theory). *The Annals of Statistics*, 2 (5), pp. 849–79, doi:10.1214/aos/1176342810.

Mukhopadhyay, S., Haines, L. M., (1995). Bayesian D-optimal designs for the exponential growth model. *Journal of Statistical Planning and Inference*, 44(3), 385–397.

Nadar M., Papadopoulos A., Kızılaslan F., (2013). Statistical analysis for Kumaraswamy's distribution based on record data. *Stat Pap*, 54, pp. 355–369.

Parsamaram, P., Jafari H., (2016). Bayesian D-optimal Design for the logistic regression model with exponential distribution for random intercept. textitJournal of Statistical Computation and Simulation, 86, pp. 1856–68, doi:10.1080/00949655.2015.1087525.

Pronzato, L, E. walter, (1985). Robust experiment design via stochastic approximation. *Mathematical BIOSCIENCES*, doi.org/10.1016/0025-5564(85)90068-9.

Rodriguez-Torreblanca C., Rodríguez-Díaz J. M., (2007). Locally D- and c-optimal designs for Poisson and negative binomial regression models. *Metrika*, Vol. 66, pp 161–172.

Sethuraman, J., (1994). A Constructive De nition of Dirichlet Priors. *Statistica Sinica 4*, p. 639, p. 650.

Zarepour, Mahmoud and Al Labadi, Luai, (2012). On a rapid simulation of the Dirichlet process. *Statistics and Probability Letters 82.5*, p. 916, p. 924.

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 155–173, https://doi.org/10.59170/stattrans-2024–033
Received – 14.07.2023; accepted – 15.04.2024

The use of the Bennet indicators and their transitive versions for scanner data analysis

Jacek Białek[1]

Abstract

Although modern price index theory is based on comparisons of ratios of prices, quantities and expenditures, we may be more interested in the magnitude of differences in these characteristics in many business applications. The benefit of using these differences is that there is no problem associated with the occurrence of zero prices and quantities, a problem that arises when we work with ratios. In practice, we most often care about decomposing value difference into indicators of contributions from price and quantity differences. The best-known price and quantity indicators are the Bennet indicators, which are not transitive. Although there have been papers in the literature that propose a transitive version of the Bennet indicators, they deal with comparisons across firms in cross-section or panel contexts.

This paper revises the price and quantity Bennet indicators and their multilateral versions for the analysis of scanner data. Specifically, instead of considering comparisons across firms, countries or regions, the transitive versions of the Bennet indicators are adapted to work on scanner data sets observed over a fixed time window. Since the scanner data sets have a high turnover of products, which can make it difficult to interpret the difference in sales values over the compared time periods, the paper also considers a matched sample approach. One of the objectives of the study is to compare bilateral and multilateral Bennet indicator results across all available products or strictly matched products over time. It also examines the impact of data filters used and the level of data aggregation on the price and quantity Bennet indicators. According to the best author's knowledge, this study is a pioneer in the field of implementing the multilateral Bennet indicators in scanner data analysis.

Key words: scanner data, the Bennet indicator, transitivity, multilateral indicators.

1. Introduction

Modern price index theory is based on comparisons of ratios of prices, quantities and expenditures (von der Lippe, 2007; International Labour Office, 2004; International Monetary Fund, 2020). These index numbers are used to construct various economic measures, such as the Gross Domestic Product (GDP) or the Consumer Price Index (CPI). Nevertheless, in many business applications we may be more interested in the magnitude of differences in prices, quantities and sale values. The approach based on differences in prices may concern many economic areas, e.g.: revenue change decompositions, profit and cost change decompositions, or the analysis of changes in consumer surplus (Diewert, 2005). An important benefit of using such differences is that there is no problem associated with the occurrence

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

of zero prices and quantities, a problem that arises when we work with ratios. It may be especially useful in many business contexts where not all goods are produced and purchased in every period.

Zero prices and quantities are also common in scanner data sets due to the high turnover of products in supermarkets (known as *product churn*). Scanner data, which support CPI calculations in many countries, mean transaction data that specify turnover and numbers of items sold by barcodes, e.g.: Global Trade Article Number (GTIN), formerly known as the European Article Number (EAN), or Stock Keeping Unit (SKU) codes. Scanner data have numerous advantages compared to traditional survey data collection because such data sets are much bigger than traditional ones and they contain complete transaction information, i.e. information about prices and quantities at the barcode level. Nevertheless, at the barcode level, in some product groups, significant shares of new and disappearing products are observed (e.g. in the case of seasonal goods, or goods that are highly sensitive to trends and fashion, such as cosmetics). Consequently, the occurrence of zero-scanner prices is even common, which causes analytical problems for statistical offices that use this kind of data (imputation of missing data) as well as for supermarket owners wishing to compare the sales performance of different product segments in two time periods. Thus, an approach based on differences in price values, quantities and expenditures can also be very useful in analyzing scanner data.

A differential pricing approach that decomposes at the overall level and at the individual product level the change in sales value into price and volume effects can also be valuable for any NSI (National Statistical Institute) that implements scanner data in inflation measurement. Of course, the approach presented in the paper is not an alternative to the CPI or HICP, but in the case of the aforementioned decomposition, it would enable the selection of the most significant scanner products in terms of sales value and thus reduce large scanner data sets to the minimum necessary. It does not make sense to take niche sales into account when determining price indices and that is why the inflation basket should include the most popular products. The discussed approach could therefore be useful when filtering products (sales) in order to limit them to the most important ones.

The difference approach to index numbers is well established in the economic literature, where it was introduced in the early 20^{th} century (Bennet, 1920). Please note that index numbers, expressed in terms of differences, are referred to as *indicators* (Diewert, 2005). Recently, one can see a return of interest in this approach on the part of statisticians and economists (Balk et al., 2004; Diewert, 2005; Fox, 2006; Cross and Färe, 2009; de Boer and Rodrigues, 2020). However, according to the best of the author's knowledge, there is a lack of papers in the literature that apply the Bennet indicator to the analysis of scanner data, which is the main objective of this paper. The contribution of this article to the indicator considerations is as follows: (1) the Bennet indicators and their transitive versions defined for comparisons across firms (or regions) in cross-section or panel context are adopted to work on scanner data sets observed over a fixed time window; (2) the axiomatic properties of the Bennet multilateral indicator are verified; (3) variants of the Bennet indicators based on matched samples are considered, which may be more accurate due to a high turnover of scanner products. In particular, a comparison of bilateral and multilateral Bennet indicator results across all available products or strictly matched products is made; (4) the impact of

data filters used and the level of data aggregation on the price and quantity Bennet indicators is also examined.

The structure of the paper is as follows: Section 2 presents the bilateral Bennet indicator in the terminology of scanner data, Section 3 adopts the transitive Bennet indicator from the field of inter-firm comparisons and presents its axiomatic properties, Section 4 presents the results of an empirical study in which the bilateral and multilateral Bennet indicators are applied to the analysis of scanner data and are compared, and Section 5 lists the most important conclusions of the research carried out.

2. The Bennet indicators

Fisher (1922) provided an axiomatic approach to the index theory (the so called *test approach*) and Konüs (1939) provided an economic framework for index numbers. Although back in the late 20^{th} century Bennet's indicators (or any other indicators) did not yet have the kind of fundamental basis that indexes do, recent theoretical results are fundamentally changing that. Many recent papers provide a promising background for the construction of the Bennet indicators. For instance, Chambers (2001) proposed a new economic framework for indicators by using Diewert's (1976) quadratic lemma. Balk et al. (2004) developed the theory of economic price and quantity indicators by deriving an exact relationship between indicators and directional distance functions. In the cited paper, observable bounds for the indicators are also derived. Finally, in the paper by Diewert (2005), an additive test approach is developed.

As a rule, the Bennet indicators are calculated using firm-level price and quantity data. Authors of most theoretical papers on the Bennet indicators use the context of production theory and/or concentrate on the input side of firms or regions (Balk et al., 2004; Cross and Färe, 2009; Fox, 2006). According to the best of the author's knowledge, this study is pioneering on the ground of implementing the multilateral Bennet indicators in scanner data analysis. In this section, the Bennet indicator formula will be expressed with the additional distinction between available and matched products that is made in the analysis of scanner data.

Let us denote sets of homogeneous products belonging to the same product group in the months 0 and t by G_0 and G_t respectively, and let $G_{0,t}$ denote a set of matched products in both moments 0 and t. Let G_0^t denote the set of available products in the months 0 and t, i.e. $G_0^t = G_0 \cup G_t$. Let p_i^τ and q_i^τ denote the price (more precisely: *unit value*) and quantity of the i-th product at the time $\tau \in \{0,t\}$, where we assume that $p_i^\tau = q_i^\tau = 0$ if the i-th product is not available at the time τ. Under the above significations, the Laspeyres and Paasche price and quantity indicators, which are additive counterparts of the Laspeyres and Paasche price and quantity indices (International Labour Office, 2004), can be written as follows:

$$IP_{0,t}^L = \sum_{i \in G_0^t} q_i^0 (p_i^t - p_i^0), IQ_{0,t}^L = \sum_{i \in G_0^t} p_i^0 (q_i^t - q_i^0), \tag{1}$$

$$IP_{0,t}^P = \sum_{i \in G_0^t} q_i^t (p_i^t - p_i^0), IQ_{0,t}^P = \sum_{i \in G_0^t} p_i^t (q_i^t - q_i^0). \tag{2}$$

The additive counterpart of the Fisher (1922) indices are the indicators of Bennet (1920) defined as the arithmetic mean of the Laspeyres and Paasche price indicators:

$$IP^B_{0,t} = \frac{1}{2}(IP^L_{0,t} + IP^P_{0,t}) = \sum_{i \in G^t_0} \frac{q^0_i + q^t_i}{2}(p^t_i - p^0_i), \tag{3}$$

$$IQ^B_{0,t} = \frac{1}{2}(IQ^L_{0,t} + IQ^P_{0,t}) = \sum_{i \in G^t_0} \frac{p^0_i + p^t_i}{2}(q^t_i - q^0_i). \tag{4}$$

Let $V^0_{G_0}$ and $V^t_{G_t}$ denote total expenditures on all available products in the periods 0 and t respectively, i.e. $V^0_{G_0} = \sum_{i \in G_0} p^0_i q^0_i$ and $V^t_{G_t} = \sum_{i \in G_t} p^t_i q^t_i$. The Bennet indicators allow us to decompose the absolute change in the total value additively into a price effect and a quantity effect (Bennet, 1920; Diewert, 2005):

$$V^t_{G_t} - V^0_{G_0} = IP^B_{0,t} + IQ^B_{0,t}. \tag{5}$$

Equation 5 means that the Bennet indicators satisfy the *sum test* known from the axiomatic approach. A list of basic axiomatic properties of the Bennet indicators is quite long, i.e. it is easy to show that these indicators fulfill the following tests for indicators: *identity*, *monotonicity in prices (quantities)*, *homogeneity of degree 1 in prices (quantities)*, *time reversibility*, as well as *dimensional invariance* (for more mathematical details see Diewert (2005); Balk (2008) or Balk (2016)).

For companies or regions, the set of commodities is relatively constant over time and zero prices or quantities are relatively rare. In the case of scanner data, due to the high turnover of products, deficiencies on the price and quantity side are rampant, especially over longer time periods. For the statistical office or the owner of a retail chain, information about the difference in total sales of scanner products in the compared periods may not be meaningful if the set of products simultaneously available in these periods is small. It seems that complementary, if not superior, information would be the knowledge of the difference in total sales in the compared periods but limited only to matched products. Following this thought, let us consider a price and quantity Bennet indicator defined only for matched products:

$${}_mIP^B_{0,t} = \sum_{i \in G_{0,t}} \frac{q^0_i + q^t_i}{2}(p^t_i - p^0_i), \tag{6}$$

$${}_mIQ^B_{0,t} = \sum_{i \in G_{0,t}} \frac{p^0_i + p^t_i}{2}(q^t_i - q^0_i). \tag{7}$$

Now, let $V^0_{G_{0,t}}$ and $V^t_{G_{0,t}}$ denote total expenditures on all matched products in the periods 0 and t respectively, i.e. $V^0_{G_{0,t}} = \sum_{i \in G_{0,t}} p^0_i q^0_i$ and $V^t_{G_{0,t}} = \sum_{i \in G_{0,t}} p^t_i q^t_i$. By analogy with the original approach (Bennet, 1920), it could be shown that the the Bennet indicators based on matched samples allow for the following decomposition:

$$V^t_{G_{0,t}} - V^0_{G_{0,t}} = {}_mIP^B_{0,t} + {}_mIQ^B_{0,t}. \tag{8}$$

3. The multilateral Bennet indicators

A lack of index *transitivity* is a well-known problem in the literature on international comparisons or scanner data (Gini, 1931; Eltetö and Köves, 1964; Szulc, 1964; Ivancic et al., 2011; Chessa, 2016). For international or inter-regional comparisons, *transitivity* means that estimates of price dynamics and quantities of selected attributes do not depend on the choice of underlying country or region. Similarly, for comparisons across firms, computing transitive price and quantity indices (indicators) do not depend on the choice of the firm-benchmark.

For some time, multilateral price indices, which were originally used for cross-country or cross-regional comparisons, have been adopted for calculating inflation based on scanner (also web-scraped) data. Commonly known methods include the GEKS method Gini, 1931; Eltetö and Köves, 1964), the Geary-Khamis method (Geary, 1958; Khamis, 1972), the CCDI method (Caves et al., 1982), or the Time Product Dummy Methods (de Haan and Krsinich, 2018). A multilateral index is compiled over a given time window composed of $T+1$ successive months (typically $T=12$), i.e. the time window consists of periods: $0,1,2,...,T$. Multilateral price indices take as input all prices and quantities of the previously defined individual products, which are available in a given time window. Multilateral price indices are *transitive*, which here means that the calculation of the price dynamics for any two moments in the time window does not depend on the choice of the base period. By definition, *transitivity* eliminates the chain drift problem. The chain drift can be formalized in terms of the violation of the *multi period identity test*. According to this test, one can expect that when all prices and quantities in a current period revert back to their values from the base period, then the index should indicate no price change and it equals one.

In the case of any price indicator IP and quantity indicator IQ, *transitivity*, in mathematical notation, means that the following relationships occur for any $0<s<t$:

$$IP_{0,s}+IP_{s,t}=IP_{0,t}, \tag{9}$$

$$IQ_{0,s}+IQ_{s,t}=IQ_{0,t}. \tag{10}$$

It can be shown that the bilateral Bennet indicators are not transitive (Fox, 2006). Chambers (1998) showed a method by which any price or quantity indicator can be made transitive. His method is applied to the Bennet indicator to explicitly derive for the first time a transitive Bennet indicator. In this paper, however, we adopt a transformation method by Fox (2008) to derive transitive Bennet price and quantity indicators (see Section 3.1).

3.1. Bennet's transitive indicators design

The design of the Bennet multilateral indicators is an adaptation of Fox's (2006) ideas for the scanner data case, so let us first introduce the significations for the considered time interval $[0,T]$. Let $G_{[0,T]}$ denote the set of available products in the whole interval $[0,T]$, i.e. $G_{[0,T]}=\cup_{\tau=0}^{T}G_{\tau}$, and let $G^{m}_{[0,T]}$ denote the set of matched products in the whole interval

$[0,T]$, i.e. $G^m_{[0,T]} = \cap^T_{\tau=0} G_\tau$. Let us introduce the additional notations:

$$ {}_TIP^B_{\tau,t} = \sum_{i \in G_{[0,T]}} \frac{q_i^\tau + q_i^t}{2}(p_i^t - p_i^\tau), \tag{11} $$

$$ {}_{mT}IP^B_{\tau,t} = \sum_{i \in G^m_{[0,T]}} \frac{q_i^\tau + q_i^t}{2}(p_i^t - p_i^\tau), \tag{12} $$

$$ {}_TIQ^B_{\tau,t} = \sum_{i \in G_{[0,T]}} \frac{p_i^\tau + p_i^t}{2}(q_i^t - q_i^\tau), \tag{13} $$

$$ {}_{mT}IQ^B_{\tau,t} = \sum_{i \in G^m_{[0,T]}} \frac{p_i^\tau + p_i^t}{2}(q_i^t - q_i^\tau). \tag{14} $$

Following and adopting Fox's (2006) transformation of the bilateral Bennet indicators, let us first construct the multilateral Bennet indicators for all available products in a fixed time interval. To do this, we need to average the bilateral Bennet indicators over a given time interval as follows:

$$ \overline{IP}^B_{t_0} = \frac{1}{T+1} \sum_{\tau=0}^{T} {}_TIP^B_{\tau,t_0}, \tag{15} $$

$$ \overline{IQ}^B_{t_0} = \frac{1}{T+1} \sum_{\tau=0}^{T} {}_TIQ^B_{\tau,t_0}. \tag{16} $$

Now, the price and quantity multilateral Bennet indicator can be defined respectively:

$$ MIP^B_{0,t} = \overline{IP}^B_t - \overline{IP}^B_0, \tag{17} $$

$$ MIQ^B_{0,t} = \overline{IQ}^B_t - \overline{IQ}^B_0. \tag{18} $$

Note that the multilateral indicators (17) and (18) are *transitive*. In fact, we have

$$ MIP^B_{0,s} + MIP^B_{s,t} = \overline{IP}^B_s - \overline{IP}^B_0 + \overline{IP}^B_t - \overline{IP}^B_s = \overline{IP}^B_t - \overline{IP}^B_0 = MIP^B_{0,t}, \tag{19} $$

and

$$ MIQ^B_{0,s} + MIQ^B_{s,t} = \overline{IQ}^B_s - \overline{IQ}^B_0 + \overline{IQ}^B_t - \overline{IQ}^B_s = \overline{IQ}^B_t - \overline{IQ}^B_0 = MIQ^B_{0,t}. \tag{20} $$

It is easy to show that multilateral indicators (17) and (18) fulfill the following tests for indicators: *homogeneity of degree 1 in prices (quantities)*, *time reversibility*, as well as *dimensional invariance* (the proof is straightforward and thus it is omitted). Note that the

sum test actually holds:

$$
\begin{aligned}
&MIP^B_{0,t} + MIQ^B_{0,t} = \frac{1}{T+1}\sum_{\tau=0}^{T}({}_TIP^B_{\tau,t} - {}_TIP^B_{\tau,0} + {}_TIQ^B_{\tau,t} - {}_TIQ^B_{\tau,0}) = \\
&= \frac{0.5}{T+1}\sum_{\tau=0}^{T}\sum_{i\in G_{[0,T]}}(q_i^\tau p_i^t - q_i^\tau p_i^\tau + q_i^t p_i^t - q_i^t p_i^\tau - q_i^\tau p_i^0 + q_i^\tau p_i^\tau - q_i^0 p_i^0 + q_i^0 p_i^\tau + \\
&+ p_i^\tau q_i^t - p_i^\tau q_i^\tau + p_i^t q_i^t - p_i^t q_i^\tau - p_i^\tau q_i^0 + p_i^\tau q_i^\tau - p_i^0 q_i^0 + p_i^0 q_i^\tau) = \\
&= \frac{0.5}{T+1}\sum_{\tau=0}^{T}\sum_{i\in G_{[0,T]}}(2p_i^t q_i^t - 2p_i^0 q_i^0) = \sum_{i\in G_{[0,T]}}(p_i^t q_i^t - p_i^0 q_i^0) = \\
&= \sum_{i\in G_{[0,T]}} p_i^t q_i^t - \sum_{i\in G_{[0,T]}} p_i^0 q_i^0 = V^t_{G_{[0,T]}} - V^0_{G_{[0,T]}}.
\end{aligned}
\tag{21}
$$

Note that from (15) and (17) we obtain

$$
MIP^B_{0,t} = \frac{1}{T+1}\Big(\sum_{\tau\notin\{0,t\}}({}_TIP^B_{\tau,t} - {}_TIP^B_{\tau,0}) + {}_TIP^B_{0,t} - {}_TIP^B_{0,0} + {}_TIP^B_{t,t} - {}_TIP^B_{t,0}\Big). \tag{22}
$$

Since ${}_TIP^B_{0,0} = {}_TIP^B_{t,t} = 0$ and ${}_TIP^B_{0,t} = -{}_TIP^B_{t,0}$ (*time reversability*), from (22), we have that

$$
MIP^B_{0,t} = \frac{1}{T+1}\Big(\sum_{\tau\notin\{0,t\}}({}_TIP^B_{\tau,t} + {}_TIP^B_{0,\tau}) + 2{}_TIP^B_{0,t}\Big). \tag{23}
$$

Since the bilateral Bennet price indicator satisfies the *monotonicity in prices* test, we conclude from (23) that the multilateral Bennet indicator also satisfies this test. In fact, if prices in the current period t rise, then the values of each of the indicators ${}_TIP^B_{\tau,t}$ and ${}_TIP^B_{0,t}$ rise too. On the other hand, if we increased prices in the base period, we would get smaller values of each of the indexes ${}_TIP^B_{0,\tau}$ and ${}_TIP^B_{0,t}$. As a consequence, the multilateral Bennet price indicator behaves identically, and thus it satisfies the *monotonicity in prices* test. In an analogous way, it can be shown that the Bennet multilateral quantity indicator satisfies the *monotonicity in quantities* test.

Now, let us assume that there is the following relationship between prices and quantities of the current and base periods: $p_i^0 = p_i^t$ and $q_i^0 = q_i^t$ for each $i \in G_{[0,T]}$. Since the multilateral

price Bennet indicator satisfies *transitivity*, we have

$$\begin{aligned} &MIP^B_{0,1} + MIP^B_{1,2} + MIP^B_{2,3} + ...MIP^B_{t-1,t} = MIP^B_{0,t} = \\ &= \frac{0.5}{T+1} \sum_{\tau=0}^{T} \sum_{i \in G_{[0,T]}} ((q_i^\tau + q_i^t)(p_i^t - p_i^\tau) - (q_i^\tau + q_i^0)(p_i^0 - p_i^\tau)) = \\ &= \frac{0.5}{T+1} \sum_{\tau=0}^{T} \sum_{i \in G_{[0,T]}} ((q_i^\tau + q_i^0)(p_i^0 - p_i^\tau) - (q_i^\tau + q_i^0)(p_i^0 - p_i^\tau)) = 0. \end{aligned} \tag{24}$$

Since the equality of prices and quantities from the current and base periods entails a relationship (24), we conclude that the Bennet price indicator satisfies the *multi-period identity test* (in the additive version for indicators). In an analogous way, it can be shown that the multilateral quantity Bennet indicator is *transitive*. According to the best of the author's knowledge, this is the first suggestion in the literature that the *multi-period identity* test should be included in the construction of transitive indicators. In our opinion, however, this is a very natural requirement, which in the construction of multilateral price indices is a key property that eliminates the problem of chain drift observed in the calculation of inflation based on scanner data (Chessa, 2015; Diewert, 2020).

In the case of well-known and widely recognized multilateral indices (e.g. GEKS, CCDI, TPD, or Geary-Khamis), we observe a certain regularity: these indices satisfy the *multi-period identity test* (they are free of chain drift) but do not satisfy the *identity test* (Białek, 2022). An analogous regularity applies to the Bennet multilateral price (and quantity) indicator. Breaking the *identity test* will be demonstrated with a simple example with regard to the multilateral price Bennet indicator.

Example. Let us consider a data set included in the publication by Eurostat (2022), i.e. a data set concerning four individual products observed in four periods $\{0,1,2,3\}$. Let us make a change in prices in the last period, assuming $p_i^3 = p_i^0$ for each $i-$th product (Tab. 1).

Table 1: Example data set with four individual products

Individual product	p^0	q^0	p^1	q^1	p^2	q^2	p^3	q^3
1	2.97	15	2.96	25	2.93	32	2.97	33
2	3.64	44	3.50	79	3.36	65	3.64	90
3	6.75	49	6.71	41	6.67	35	6.75	53
4	3.37	35	3.29	59	3.37	30	3.37	31

After calculations (*), we obtain the difference in total sales value in the last and first period, which equals 234.42. In the case of the bilateral Bennet indicators we obtain: $IP^B_{0,3} = 0$ and $IQ^B_{0,3} = 234.42$, which confirms that the bilateral Bennet indicators meet the *identity test*. However, let us note that at the same time we obtain: $MIP^B_{0,3} = 2.55 \neq 0$ and $MIQ^B_{0,3} = 231.87$, which means that the multilateral price Bennet indicator does not satisfy the *identity test*.

(*) the corresponding R script is available at:
https://github.com/JacekBialek/important_documents/blob/main/IdentityBennet.R

Remark. Let us conclude this session by pointing out that all the axiomatic properties valid for the multilateral Bennet indicators estimated on the set of all available products $G_{[0,T]}$ also carry over to the analogous multilateral Bennet indicators estimated on the set of matched products $G^m_{[0,T]}$. The transitive Bennet indicators based on matched products can be written as

$$ {}_mMIP^B_{0,t} = \frac{1}{T+1}\sum_{\tau=0}^{T}({}_mIP^B_{\tau,t} - {}_mIP^B_{\tau,0}), \tag{25} $$

$$ {}_mMIQ^B_{0,t} = \frac{1}{T+1}\sum_{\tau=0}^{T}({}_mIQ^B_{\tau,t} - {}_mIQ^B_{\tau,0}). \tag{26} $$

In an analogous way to (21), it can be shown that holds:

$$ {}_mMIP^B_{0,t} + {}_mMIQ^B_{0,t} = V^t_{G^m_{[0,T]}} - V^0_{G^m_{[0,T]}}. \tag{27} $$

3.2. Multilateral Laspeyres and Paasche indicators vs multilateral Bennet indicators

Let us introduce additional notations for the Laspeyres (superscript "L") and Paasche (superscript "P") indicators defined on the basis of the whole set of available products:

$$ {}_TIP^L_{\tau,t} = \sum_{i\in G_{[0,T]}} q^\tau_i(p^t_i - p^\tau_i), \tag{28} $$

$$ {}_TIQ^L_{\tau,t} = \sum_{i\in G_{[0,T]}} p^\tau_i(q^t_i - q^\tau_i), \tag{29} $$

$$ {}_TIP^P_{\tau,t} = \sum_{i\in G_{[0,T]}} q^t_i(p^t_i - p^\tau_i), \tag{30} $$

$$ {}_TIQ^P_{\tau,t} = \sum_{i\in G_{[0,T]}} p^t_i(q^t_i - q^\tau_i). \tag{31} $$

Let us define, analogously to (17) and (18), the average bilateral Laspeyres and Paasche indicators over a given time interval as follows:

$$ \overline{IP}^L_{t_0} = \frac{1}{T+1}\sum_{\tau=0}^{T} {}_TIP^L_{\tau,t_0}, \tag{32} $$

$$ \overline{IQ}^L_{t_0} = \frac{1}{T+1}\sum_{\tau=0}^{T} {}_TIQ^L_{\tau,t_0}, \tag{33} $$

$$ \overline{IP}^P_{t_0} = \frac{1}{T+1}\sum_{\tau=0}^{T} {}_TIP^P_{\tau,t_0}, \tag{34} $$

$$ \overline{IQ}^P_{t_0} = \frac{1}{T+1}\sum_{\tau=0}^{T} {}_TIQ^P_{\tau,t_0}. \tag{35} $$

Now, the price and quantity multilateral Laspeyres and Paasche indicators can be defined respectively:

$$MIP^L_{0,t} = \overline{IP}^L_t - \overline{IP}^L_0, \tag{36}$$

$$MIQ^L_{0,t} = \overline{IQ}^L_t - \overline{IQ}^L_0, \tag{37}$$

$$MIP^P_{0,t} = \overline{IP}^P_t - \overline{IP}^P_0, \tag{38}$$

$$MIQ^P_{0,t} = \overline{IQ}^P_t - \overline{IQ}^P_0. \tag{39}$$

It is easy to verify that the multilateral Laspeyres and Paasche indicators are *transitive*. Please note that the Laspeyres indicators satisfy *identity*. In fact, if prices in the current period are the same as prices in the base period, i.e. $p^0_i = p^t_i$ for each i, then we obtain

$$\begin{aligned} MIP^L_{0,t} &= \frac{0.5}{T+1}\sum_{\tau=0}^{T}\sum_{i\in G_{[0,T]}}(q^\tau_i(p^t_i - p^\tau_i) - q^\tau_i(p^0_i - p^\tau_i)) = \\ &= \frac{0.5}{T+1}\sum_{\tau=0}^{T}\sum_{i\in G_{[0,T]}}(q^\tau_i(p^0_i - p^\tau_i) - q^\tau_i(p^0_i - p^\tau_i)) = 0. \end{aligned} \tag{40}$$

It can be shown in a similar way that if quantities in the current period are the same as quantities in the base period, i.e. $q^0_i = q^t_i$ for each i, then we obtain $MIQ^L_{0,t} = 0$. Note that this conclusion is analogous to results from the paper of Białek (2022). The cited paper shows that although the multilateral GEKS index, which is *transitive*, does not satisfy the *identity test* for indices, its version based on the Laspeyres index (GEKS-L) does. Unfortunately, in general, we observe that $V^t_{G_{[0,T]}} - V^0_{G_{[0,T]}} \neq MIP^L_{0,t} + MIQ^L_{0,t}$ and $V^t_{G_{[0,T]}} - V^0_{G_{[0,T]}} \neq MIP^P_{0,t} + MIQ^P_{0,t}$. Nevertheless, as can easily be shown (the proof is omitted), there is a relationship analogous to the one we observe for the Laspeyres and Paasche bilateral indicators:

$$V^t_{G_{[0,T]}} - V^0_{G_{[0,T]}} = MIP^L_{0,t} + MIQ^P_{0,t}, \tag{41}$$

$$V^t_{G_{[0,T]}} - V^0_{G_{[0,T]}} = MIP^P_{0,t} + MIQ^L_{0,t}. \tag{42}$$

Since, from (41) and (42), we conclude that the Laspeyres and Paasche multilateral indicators are equally good, we may use their arithmetic mean to define a proper multilateral indicator. Nevertheless, as expected, we then get the following:

$$\begin{aligned} &\frac{MIP^L_{0,t} + MIP^P_{0,t}}{2} = \\ &= \frac{0.5}{T+1}\sum_{\tau=0}^{T}\sum_{i\in G_{[0,T]}}(q^\tau_i(p^t_i - p^\tau_i) - q^\tau_i(p^0_i - p^\tau_i) + q^t_i(p^t_i - p^\tau_i) - q^0_i(p^0_i - p^\tau_i)) = \\ &= \frac{0.5}{T+1}\sum_{\tau=0}^{T}\sum_{i\in G_{[0,T]}}((q^\tau_i + q^t_i)(p^t_i - p^\tau_i) - (q^\tau_i + q^0_i)(p^0_i - p^\tau_i)) = \overline{IP}^B_t - \overline{IP}^B_0 = MIP^B_{0,t}. \end{aligned} \tag{43}$$

In an analogous way, it can be shown that

$$\frac{MIQ^{L}_{0,t} + MIQ^{P}_{0,t}}{2} = MIQ^{B}_{0,t}. \tag{44}$$

4. Empirical illustration

In the following empirical study, we use scanner data from one retail chain in Poland, i.e. monthly data on *stationery and hygiene products* (COICOP 5: 121322) sold in over 500 outlets during the period from December 2021 to December 2022 (334268 records, which means 62 MB of data in *csv* format). The COICOP 5 product group consists of the following local COICOP 6 product subgroups: tissues (60 products: IDs), wet wipes (88 IDs), toilet paper (117 IDs), baby diapers (193 IDs), sanitary pads (20 IDs), sanitary napkins (67 IDs), and tampons (22 IDs).

Before calculating the price indices, the data sets were carefully prepared. Product classification was performed using the `data_selecting()` and `data_classification()` functions from the `PriceIndices` R package (Białek, 2021). The first function required manual preparation of dictionaries of keywords and phrases that identified individual product groups. The second function was used for problematic, previously unclassified products and required manual preparation of learning samples based on historical data. The classification itself was based on machine learning using random trees and the XGBoost algorithm (Tianqi and Carlo, 2016). Next, the product matching was carried out based on the available GTIN (Global Trade Item Number) bar codes, internal retail chain codes and product labels. To match products, we used the `data_matching()` function from the `PriceIndices` package. To be more precise: products with two identical codes or one of the codes identical and an identical description were automatically matched. Products were also matched if they had identical one of the codes and the Jaro-Winkler distance (Jaro, 1989) of their descriptions was smaller than the fixed precision value: 0.02.

The sales value difference, the Bennet price indicator ("price effect" in the figure) and the Bennet quantity indicator ("quantity effect") were compared, with calculations made for different variants. First, the bilateral versions of the Bennet indicators (Fig. 1 and 2) was compared separately with the multilateral versions (Fig. 3 and 4). The bilateral indicators, like the multilateral ones, were also considered in two cases: without filtering the original data (Fig. 1 and 3) and with the application of a low sales filter with its parameter $\lambda = 1.25$ (van Loon and Roels, 2018) (Fig. 2 and 4). The low sales filter was used to eliminate products with relatively low sales from the sample (almost 29% of products were removed).

Clarification is needed on how to interpret Fig. 1-4. Let us first emphasize that this part of the analysis considers the most disaggregated level of data, i.e. the GTIN barcode level. For the characteristics under study (value difference, price or quantity effect), their sorted in ascending order values calculated for all months from the analyzed time window and determined for all available products in this time window are marked on the OX axis. On the OY axis, analogous values for the corresponding months are marked, with only matched products included this time. In this way, easy-to-interpret figures are obtained. Namely: the theoretical red line (the curve $y = x$, which is named "identity") would indicate a situation in which the inclusion of product matching does not change the values of the indicators

estimated on the set of all available products. If the empirical curve (the green one, which is named "observed") is under the theoretical red line, it means that product matching caused a decrease in the values of the characteristics under study. The more the empirical line diverges from the theoretical one, the greater the above-mentioned effect becomes. On the other hand, if the empirical curve is above the theoretical line, we conclude that considering only matched products in determining the difference in the value of sales and the Bennet indicators led to an increase in the value of these characteristics compared to estimates based on all available products.

As can be observed in Fig. 1-2, for the bilateral approach, data filtering clearly increases the product matching effect especially in the context of the price Bennet indicator. Here, after applying data filtering, we observe a partial transition of the empirical curve above the theoretical line, which was not observed before filtering. Interestingly, in the multilateral approach, we observe an analogous effect but on the side of the Bennet quantity indicator (see Fig. 3 and 4). Consequently, after filtering, here we observe greater differences in sales values in the compared months for matched products than for all available products (the empirical curve is generally above the theoretical line, see Fig. 4).

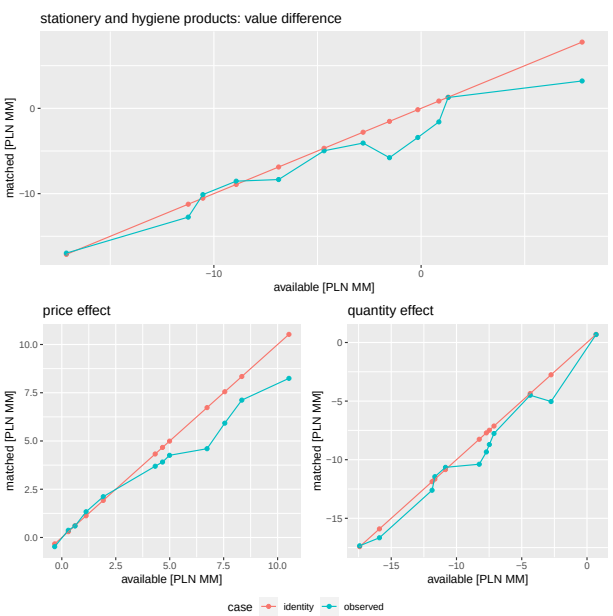


Figure 1: Comparison of the difference in the value of sales and the bilateral Bennet indicators calculated for all available products and for matched products (no data filters are applied)

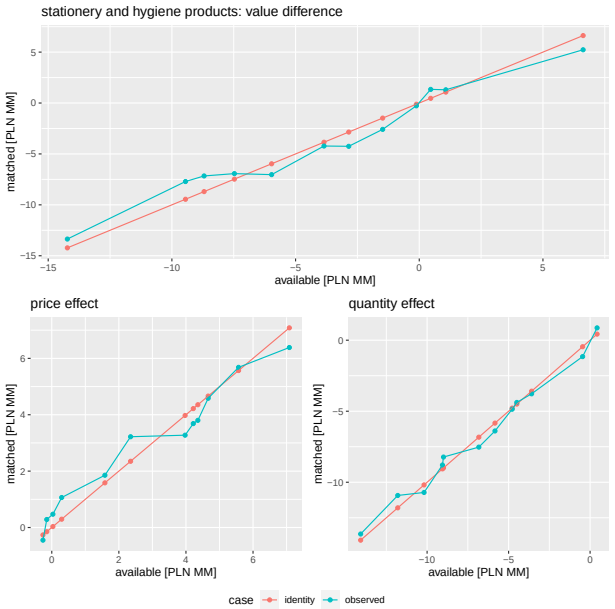


Figure 2: Comparison of the difference in the value of sales and the bilateral Bennet indicators calculated for all available products and for matched products (data filters are applied)

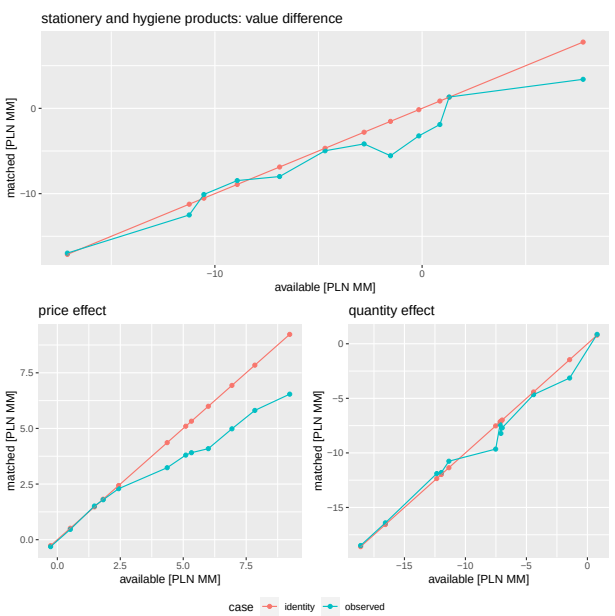


Figure 3: Comparison of the difference in the value of sales and the multilateral Bennet indicators calculated for all available products and for matched products (no data filters are applied)

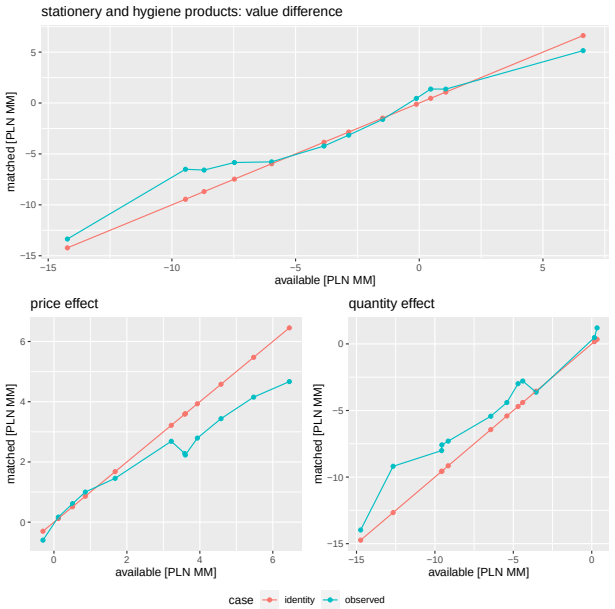


Figure 4: Comparison of the difference in the value of sales and the multilateral Bennet indicators calculated for all available products and for matched products (data filters are applied)

Undoubtedly, in the case of defining a homogeneous product at the level of the GTIN bar code, the differences between multilateral and bilateral indicators must simply result from the fact that some of these products are withdrawn from sale during the time window. Both in the case of bilateral and multilateral approaches, a significant impact of data filtering on the relationship between indicators determined for matched products and indicators determined based on all available products can be seen. In order to take a closer look not so much at the relationship as at the values of the quantity and price Bennet indicators, it was decided to additionally take into account the level of data aggregation. To be more precise, the values of the Bennet indicators (in all versions) were determined both for the product defined very narrowly (GTIN code level) and broadly (COICOP 6 level). In the case of the COICOP 6 level, as a rule, we do not observe a loss of products (subgroups), and thus the theoretical line will always coincide with the empirical line (matching the products will not change the values of the indicators). However, it can be expected that not only data filtering, but also a change in the level of data aggregation will substantially affect the price and quantity values of the Bennet indicators. An interesting direction of research also seems to be the verification of the hypothesis that the transition from bilateral to multilateral Bennet indicators will substantially change the shares of individual subgroups of products after decomposition of indicators even at a higher level of aggregation. Taking the above aspects into account, a number of comparisons were made for the current period set at the end of the time window (December, 2022).

Table 2: Comparison of the Bennet indicators across data aggregation level and data filtering (all available products are considered, the normalized (∗) values are in brackets)

GTIN level: bilateral approach		
characteristics	no data filtering	with data filtering
sales value difference	7769806.28 (100)	6624556.59 (100)
price Bennet indicator	10525438.59 (135.47)	7082331.13 (106.91)
quantity Bennet indicator	-2755632.31 (-35.47)	-457774.54 (-6.91)
GTIN level: multilateral approach		
characteristics	no data filtering	with data filtering
sales value difference	7769806.28 (100)	6624556.59 (100)
price Bennet indicator	9228411.35 (118.77)	6452990.29 (97.41)
quantity Bennet indicator	-1458605.07 (-18.77)	171566.3 (2.59)
COICOP 6 level: bilateral approach		
characteristics	no data filtering	with data filtering
sales value difference	7769806.28 (100)	6624556.59 (100)
price Bennet indicator	9952719.32 (128.09)	8079099.68 (121.96)
quantity Bennet indicator	-2182913.04 (-28.09)	-1454543.09 (-21.96)
COICOP 6 level: multilateral approach		
characteristics	no data filtering	with data filtering
sales value difference	7769806.28 (100)	6624556.59 (100)
price Bennet indicator	8456853.19 (108.84)	6875623.85 (103.79)
quantity Bennet indicator	-687046.91 (-8.84)	-251067.26 (-3.79)

∗ *By normalized values we mean those obtained by taking the difference in sales value set at 100 as a reference point.*

Table 3: Price and quantity contributions across bilateral and multilateral approach (all available products are considered, the normalized (∗) values are in brackets: total sales value difference = 100)

product contributions: bilateral approach			
COICOP 6 subgroup	sales value difference	price contributions	quantity contributions
tissues	2423488.86 (36.58)	1738333.88 (26.24)	685154.98 (10.34)
toilet paper	-149921.79 (-2.26)	92165.13 (1.39)	-242086.92 (-3.65)
baby diapers	3200407.98 (48.31)	5963636.80 (90.02)	-2763228.82 (-41.71)
sanitary pads	1142085.37 (17.24)	208307.56 (3.14)	933777.81 (14.10)
sanitary napkins	-103572.86 (-1.56)	31282.13 (0.47)	-134854.99 (-2.04)
tampons	112069.03 (1.69)	45374.18 (0.68)	66694.85 (1.01)
product contributions: multilateral approach			
COICOP 6 subgroup	sales value difference	price contributions	quantity contributions
tissues	2423488.86 (36.58)	1446161.01 (21.83)	977327.85 (14.75)
toilet paper	-149921.79 (-2.26)	72092.56 (1.09)	-222014.35 (-3.35)
baby diapers	3200407.98 (48.31)	5054000.86 (76.29)	-1853592.88 (-27.98)
sanitary pads	1142085.37 (17.24)	226985.90 (3.43)	915099.47 (13.81)
sanitary napkins	-103572.86 (-1.56)	25785.94 (0.39)	-129358.80 (-1.95)
tampons	112069.03 (1.69)	50597.58 (0.76)	61471.45 (0.93)

∗ *By normalized values we mean those obtained by taking the difference in sales value set at 100 as a reference point.*

It can be noted that in our study, for unfiltered data, larger absolute values of the Bennet indicators were obtained in the bilateral case than in the multilateral one (see Tab. 2). Thus, the lack of data filtering generated greater separation of price and quantity effects. In the case of filtered data, the opposite relation is true, i.e. greater separation of price and quantity effects can be observed on the side of the multilateral Bennet indicators. In both approaches (bilateral and multilateral), regardless of the level of data aggregation, the application of the low sales filter led to a reduction in the absolute value of the Bennet indicators, and thus to a flattening of the difference between the price and quantity effects.

When analyzing the shares of individual product subgroups in the total volume, price and quantity effect, it can be seen that in the case of aggregation at the COICOP 6 level, the change in approach from bilateral to multilateral does not cause such large changes in the values of the indicators (Tab. 3). The only exception to this is the *baby diapers* subgroup, for which the largest increase in sales value was recorded during the time interval considered (3200408 PLN). For this subgroup of products, the normalized Bennet price indicator changed from 90.02 to 76.29, and the normalized Bennet quantity indicator from -41.71 to -27.98 when switching from a bilateral to a multilateral approach. However, we do not make general conclusions here and the relationships presented require in-depth research to be able to call them regularities.

5. Conclusions

The adaptation of the Bennet transitive indicators to multilateral versions, operating on a fixed time window, appears to be a valuable addition to the analysis of scanner data due to the product churn that occurs here. Such additional analysis, which separates volume, price and quantity effects at different levels of aggregation, can be a valuable addition to analyses conducted by statistical offices (e.g.: when determining the list of representatives of the CPI basket) but can also be a valuable source of information for the owner of a retail chain when determining product demand.

A valuable result from the work is the conclusion that the versions of the Bennet bilateral and multilateral indicators differ not only in the set of tests (axioms) they fulfill but also generate different values regardless of the level of data aggregation. The main theoretical conclusion is that multilateral Bennet indicators, while gaining *transitivity*, lose one of the leading axioms (*identity test*). However, as it was shown, it is possible to construct an indicator that satisfies both *transitivity* and *identity* (e.g.: the Laspeyres multilateral indicator). The paper also proposes that the construction of multilateral indices should take into account the appropriate version of the *multi-period identity test*. The main practical conclusion, on the other hand, is that the relationship between bilateral price and quantity indicators depends on the level of data aggregation, the choice between matched products and all available products, and the possible use of data filters. In particular, the application of a *low sales filter* led to a flattening of the difference between the price and quantity effects in the study, regardless of the level of data aggregation.

According to the best of the author's knowledge, this article is the first application of the Bennet multilateral indicators (in "matched" and "available" versions) for the analysis of scanner data. Nevertheless, it raises many questions for the future and opens up potentially

new research directions. In particular, from a theoretical point of view, the problem outlined in the work seems interesting, namely an attempt to construct multilateral indicators that satisfy both the *sum test* (*volume test*) and *identity*. By analogy with multilateral indices and the so-called *splicing methods*, techniques of combining data from a new month (and thus a new time window) with data from the previous window may also be important for practice. Various forms of extensions for multilateral Bennet indicators can be considered here, which would, for example, keep a fixed time window width and connect the new time window with the old one. From a practical point of view, it seems important to deepen research similar to the one presented in the *Empirical illustration* section. Taking into account a wide range of products and also intermediate levels of data aggregation could then allow for a certain generalization of practical conclusions.

References

Balk, B. M., (2008). *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference*. Cambridge University Press.

Balk, B. M., (2016). *A Review of Index Number Theory*, pp. 1–24. John Wiley and Sons, Ltd.

Balk, B. M., Färe, R., and Grosskopf, S., (2004). The theory of economic price and quantity indicators. *Economic Theory*, 23(1), pp. 149–164.

Bennet, T. L., (1920). The theory of measurement of changes in cost of living. *Journal of the Royal Statistical Society*, 83(3), pp. 455–462.

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

Białek, J., (2023). Improving quality of the scanner CPI: proposition of new multilateral methods. *Quality and Quantity*, 57, pp. 2893–2921. https://doi.org/10.1007/s11135-022-01506-6.

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

Chambers, R. G., (1998). *Input and Output Indicators*. Springer Netherlands, Dordrecht, pp. 241–271.

Chambers, R. G., (2001). Consumers' surplus as an exact and superlative cardinal welfare indicator. *International Economic Review*, 42(1), pp. 105–119.

Chessa, A., (2015). *Towards a generic price index method for scanner data in the dutch CPI*. In: 14th meeting of the Ottawa Group, Tokyo, pp. 20–22.

Chessa, A., (2016). *A new methodology for processing scanner data in the Dutch CPI*. Eurostat review of National Accounts and Macroeconomic Indicators, pp. 1:49–69.

Cross, R. M., Färe, R., (2009). *Value data and the Bennet price and quantity indicators*. Economics Letters, 102(1), pp. 9–21.

de Boer, P. and Rodrigues, J. F. D., (2020). Decomposition analysis: when to use which method? *Economic Systems Research*, 32(1), pp. 1–28.

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

Diewert, W., (2005). Index number theory using differences rather than ratios. *American Journal of Economics and Sociology*, 64, pp. 311–360.

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

Diewert, W. E., (2020). *The chain drift problem and multilateral indexes*. Technical report, Discussion Paper 20-07, Vancouver School of Economics.

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

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

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

Fox, K. J., (2006). A method for transitive and additive multilateral comparisons: A transitive bennet indicator. *Journal of Economics*, 87(1), pp. 73–87.

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

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

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

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

Ivancic, L., Diewert, W. E., and Fox, K. J., (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics*, 161(1), pp. 24–35.

Jaro, M., (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406), pp. 414–420.

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

Konüş, A., (1939). The problem of the true index of the cost of living. *Econometrica*, 7, pp. 10.

Szulc, B., (1964). Indices for multiregional comparisons. *Przeglad statystyczny*, 3, pp. 239–254.

Tianqi, C., Carlo, G., (2016). *Xgboost: A scalable tree boosting system*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM, pp. 785–794.

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

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

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 175–185, https://doi.org/10.59170/stattrans-2024-034
Received – 30.11.2023; accepted – 08.07.2024

Language independent algorithm for clustering text documents with respect to their sentiment

Jerzy Korzeniewski[1], **Adam Idczak**[2]

Abstract

Determining the sentiment of a written text is an important task in text research. This task can be performed either in the supervised or unsupervised version. In this paper, we propose a novel unsupervised algorithm for documents written in any language using documents written in Polish as an example. The clustering of Polish language texts with respect to their sentiment is poorly developed in the literature on the subject. The novelty of the proposed algorithm involves the abandonment of stoplists and lemmatisation. Instead, we propose translating all documents into English and performing a two-stage document grouping. In the first step of the algorithm, selected documents are assigned to a class of positive or negative documents based on a set of lexical and grammatical rules as well as a set of key-terms. Key-terms do not have to be entered by the user, the algorithm finds them. In the second step, the remaining documents are attached to one of the classes according to the rules based on the vocabulary found in the documents grouped in the first step. The algorithm was tested on three corpora of documents and achieved very good results.

Key words: text mining, document sentiment, document clustering.

1. Introduction

Text clustering is becoming more and more important every day with the rapid development of media massive output of news, posts, comments, articles, etc. Media and official government departments can effectively handle news and grasp the development trend of news popularity. Therefore, clustering of texts has become an important research topic in text clustering. Text clustering with respect to text sentiment is a very vital part of this research topic. By text sentiment we understand either positive or negative opinion expressed by the author about the subject of the document under study. Other variants of the meaning of the term 'sentiment' are possible like, e.g. 'sentence sentiment', but we refer it to the sentiment of the whole

[1] Department of Demography, Faculty of Economics and Sociology, University of Lodz, Lodz, Poland. E-mail: jerzy.korzeniewski@uni.lodz.pl. ORCID: https://orcid.org/0000-0001-6526-5921.

[2] Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz, Lodz, Poland. E-mail: adam.idczak@uni.lodz.pl. ORCID: https://orcid.org/0000-0001-9676-2410.

document. This approach is probably most difficult as it creates problems when there are several objects being described in the document. We propose a novel algorithm tested on Polish language texts, but the algorithm is independent of the language because its intrinsic feature is text initial translation into English. Another important feature of our proposal is the abandonment of any stoplists and lemmatisations. The reasons are quite obvious, in the case of the Polish language none of implementation of lemmatisation is free of errors, for example a word that has multiple meanings is converted into the base form of a word with a different meaning. Moreover, some words are not recognized and are not lemmatized at all. Even if there was any such procedure it would often kill the meaning of the text through harsh simplification of grammatical and lexical structures. Eder and Górski (2023) showed that there is no substantial difference between using lemmatised words and its original forms in classification problem.

The paper is organized as follows. In Part 2 we give a short summary of related work in this research topic. In Part 3 we describe the proposal of our algorithm. Part 4 contains empirical evaluation of the algorithm in the example of three Polish language text corpora and conclusions are given in Part 5.

2. Related work review

Although the task of unsupervised text clustering is closely related to the supervised version it is significantly less researched. It can be even considered to be a relatively new topic as its beginnings go back to hardly 2002 (Pang et al., Turney). For example, Turney used a specific unsupervised learning technique based on the mutual information between document phrases and the words "excellent" and "poor", where the mutual information is computed using numbers collected by a search engine. Li and Liu (2012) proposed an algorithm for document clustering whose idea was to cluster repeatedly the documents via the *k*-means with random choice of starting points with subsequent majority voting mechanism. The number of running times needed to be large enough to lessen the effect of outliers and instability. In the second stage the polarity of clusters was determined on the basis of external sources (WordNet). Souza (Souza et al., 2017) proposed a Particle Swarm Optimization (PSO) algorithm to cluster documents with respect to their sentiment. The cosine distance and silhouette index were used in assessing the clustering quality. The results were not very impressive. For example, the accuracy of clustering ranged from 40% to 60% and was sometimes losing to *k*-means. The algorithm mimicked the behavior of insects in their search for food and attracted much attention, particularly in biological sciences. Probably, the first to propose a probability-based model approach to sentiment analysis were Lin and He (Lin et al., 2009). Their model was based on the Latent Dirichlet Allocation (LDA) and

achieved accuracy equal to about 70%. Similar to this approach is the one based on SOM (Self Organizing Map) type of neural networks, e.g. Chifu et al. (2015). In their research the algorithm achieved accuracy of about 60%.

As far as researching the Polish language text is concerned, the only methodological work known to the authors is Kocon et al. (2019), however, the methods developed in this work are dependent on external sources.

3. Description of the proposed algorithm

The main idea behind the construction of the algorithm is the lack of use any extensive usage of stoplists and lemmatizations. Instead, the algorithm uses the translation of documents written in Polish into English. Translated documents are grouped based on a two-stage algorithm. In the first stage, documents are grouped based on rules using lists of positive and negative terms and key-terms found. In the second stage, documents are grouped based on positive and negative bigrams. These positive and negative bigrams are found close to key-terms in positive and negative documents grouped in the first stage respectively. Below are the steps that make up the algorithm:

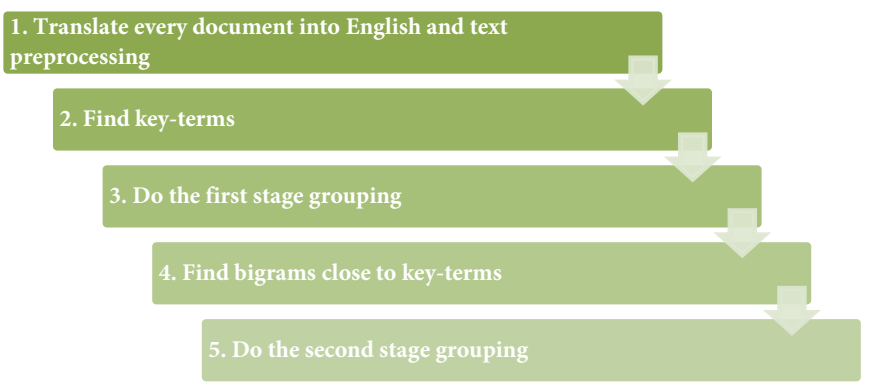


Figure 1: Algorithmic description of the experiment

Source: own work.

Translation of documents

The document corpora are translated fully automatically using the R language and the translate 2 function from the deeplr package[3]. This function uses the online DeepL translator[4].

[3] R documentation is available at https://cran.r-project.org/web/packages/deeplr/deeplr.pdf

[4] https://www.deepl.com/translator

Text preprocessing

Text preprocessing is an integral part of every text mining application. We reduce this step only to (1) remove all unwanted punctuation marks except sentence-ending marks, i.e. "?", "!", ".", (2) use our own shortened stoplist containing only stopwords such as: "a", "an", "the", "and", "or". These words are very common in the English language and rather do not convey any meaning.

In this step, lemmatization is usually carried out, which is to remove inflections and map a word to its root form. To perform lemmatization, dictionaries (e.g. *wordnet*) are needed to enable this type of conversion of an inflectional word to its basic form. We propose to skip of direct lemmatization Polish language at this stage by swapping highly inflected Polish for less inflected English. This action has a twofold effect on the text. First, implicitly, a 'soft' lammatization is performed. Hence, for example, Polish words *samochód, samochodu, samochodowi, samochodem, samochodzie* can be replaced by one English word *car*. Secondly, morphologically complex languages with relatively free word-order which is Polish are converted into more structured English, which is more suitable for creating grouping rules.

Finding key-terms

We define a key-term as any term which is directly preceded or followed by a word from both positive and negative lists at least once as shown in the diagram below:

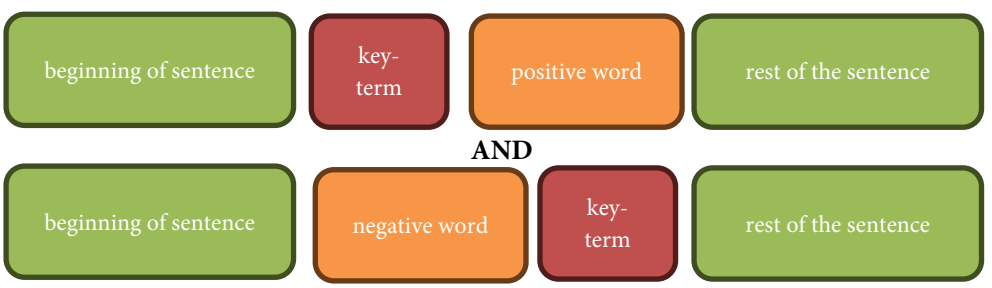


Figure 2: Key-term searching idea

Source: own work.

Example for searching *hotel* key-term:
Great hotel for business trips!
I am disappointed because of my stay in this ugly hotel.

Lists of positive and negative words were created in an approach based on data analysis with expert verification. The list of positive (negative) words was formed by terms next to (up to 3 words) to the most common nouns in the set of positive

(negative) documents. Expert verification consisted of removing words with no sentiment, in particular, these were nouns (acting as subjects) or verbs (most often the word *to be* in various forms). The positive (negative) list contains words that are commonly associated with a positive (negative) connotation. These are mainly adjectives, adverbs and nouns. The positive list is presented below:

adequate, amazing, awesome, beautiful, beautifully, best, better, classic, classics, clean, cleanliness, comfortable, competent, convincing, convenient, cool, cozy, decent, delicious, delighted, durable, ecstatic, effectively, efficient, efficiently, elegant, enjoying, excellent, exceptional, extra, fantastic, favourite, favourites, fine, firecracker, flawless, fresh, friendly, fun, good, great, happy, high-end, ideal, interesting, like, likes, long-lasting, love, lovely, magnificent, mega, neat, nice, nicely, ok, okay, outstanding, peaceful, perfect, pleasant, positive, positively, pretty, professional, professionally, reliable, revelation, right, satisfied, sensational, successful, successfully, super, superb, tasteful, tasty, timeless, top, true, truly, unbeatable, valuable, value, well, wonderful, worth, worthy.

The negative list is presented below:

bad, blatantly, break, broken, counterfeit, damage, damaged, defeat, defective, destroy, destroyed, dirty, disappointed, disappointing, disappointment, disillusion, disrupt, disturbance, disturbed, downside, dull, embarrassing, failure, faint, fake, fatal, hate, horrible, inadequate, lack, lacks, letdown, lousy, miserable, missing, monotone, mundane, plain, poor, poorest, poorly, problem, problems, scandal, scandalous, scandalously, shit, shitty, sorry, spoil, spoiled, tacky, terrible, terribly, ugly, unclean, unfortunately, uninteresting, unprofessional, unrealistic, unreliable, unsatisfied, unsuccessful, unsuccessfully, until, untrue, unvaluable, unworthy, valueless, waste, weak, worse, worst.

The first stage grouping

In the first stage grouping documents are grouped into two clusters (negatives or positives) according to the following 3 rules.

Rule 1:

- **positive label** is assigned to the document in which there is RECOMMEND (or its variation) followed by a full stop.
- **negative label** is assigned to the document in which there is RECOMMEND (or its variation) followed by a full stop and preceded with negation.

Rule 2:

- **positive label** is assigned to the short document (n<5) in which there is at least one term from the list of positive terms or at least one term from the list of

negative terms preceded with negation (and there are no negative terms in a document),

- **negative label** is assigned to the short document (n<5) in which there is at least one term from the list of negative terms or at least one term from the list of positive terms preceded with negation (and there are no positive terms in a document).

Rule 3:

- **positive label** is assigned to the document in which there is a key-term directly followed or preceded by a term from the list of positive terms and there are no terms from the list of negative terms,
- **negative label** is assigned to the document in which there is a key-term directly followed or preceded by a term from the list of negative terms and there are no terms from the list of positive terms.

Finding bigrams

Bigrams are created based on up to the nearest 4 words preceding or following the key-terms separately for positive and negative documents grouped in the first stage. The procedure searches among these 4 words for the first 2 "to the left" and the first 2 "to the right" of the key-term, skipping keywords and taking into account punctuation marks ending the sentence. We will further call these two words positive or negative bigrams (depending on which cluster the document belongs to). These bigrams are an essential part of the second stage grouping. The main idea for searching such bigrams is presented in Figure 3.

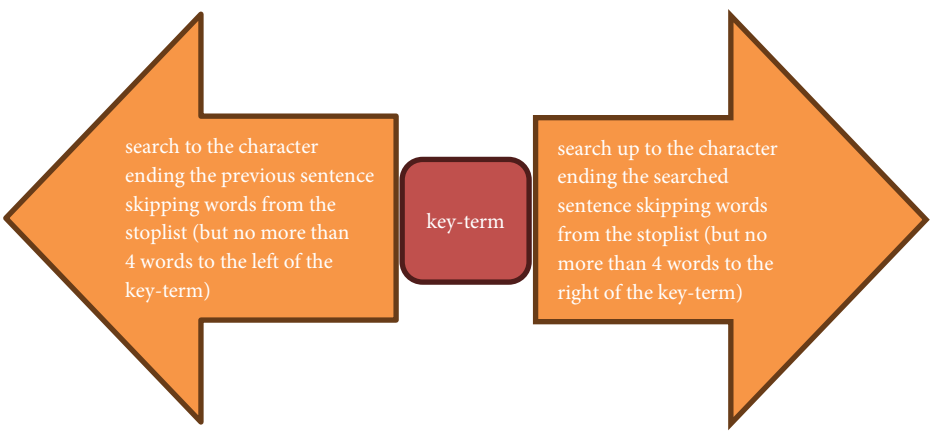


Figure 3: Bigram searching idea

Source: own work.

Real example (black words from the stoplist are skipped) – positive document with *reception* key-term:
For praise deserves the reception open 24 hours a day.
Bigrams created: *praise deserves*, *open 24*.

The second stage grouping

After the first grouping step, there are still documents that have not been assigned to one of the clusters. In the second stage, the positive and negative bigrams described above will be used to group these documents, i.e. **positive (negative) label** is assigned to the document in which there are more positive (negative) bigrams.

4. Empirical evaluation of algorithm

Data sets

Proposed method is evaluated on three datasets. These datasets were obtained from a company that collects online reviews. The first dataset called *Hotels* consists of reviews of hotels in Poland from 2020–2021. It consists of 4 385 terms. There were 221 negative and 1 667 positive documents as a total of 1 888 documents in the first dataset. The second corpus named *Perfumes* consists of reviews of perfumes from online shop from 2021. It consists of 2 675 terms. There were 271 negative and 2 333 positive documents as a total of 2 604 documents in the first dataset. The third collection of documents (*Courier*) contains reviews of courier companies from Poland written in 2021. This dataset includes 4 579 terms among 4 191 documents (541 negative, 3 650 positive).
Each document in these three datasets is labelled with positive or negative sentiment (positive or negative class). These labels were assigned manually by an opinion holder (by rating 1–5 stars where 1–2 stars documents are labelled as negative sentiment and 4–5 are marked as positive, 3 stars are excluded because of ambiguous connotation).

Results

To determine the optimal number of key-terms, we ran the algorithm many times setting a different number of most frequent terms in the range of 10–100, measuring the quality of classification and the percentage of unclassified documents[5]. Based on the results, we recommend and adopt in the study the use of 20 key-terms, which in the 3 datasets considered provided the smallest percentage of unclassified documents with high value of accuracy and F1 measures.
Hotels key-terms:

apartment, atmosphere, bathroom, beds, breakfast, conditions, decor, everything, food, helpful, hotel, large, location, place, price, restaurant, room, service, staff, tidy.

[5] Excessively many key-terms negatively affect clustering rules since the document (often short review) would mostly consist of keywords. Even some documents would remain unclassified because they would only consist of key-terms.

Perfumes key-terms:

as, delicate, fragrance, intense, it, lasting, long, masculine, more, my, original, packaging, perfume, price, product, quality, quiet, scent, sensual, smells.

Courier key-terms:

all, always, company, condition, contact, courier, delivery, everything, fast, it, order, package, pay, possible, price, quality, quick, service, shipment, time.

For each data set, positive and negative bigrams were determined according to the idea presented in Figure 3. The mechanical method of determining bigrams is not without drawbacks, as both lists repeat terms with no sentiment such as "have been", "has been", "can not", etc. It seems reasonable to correct such inaccuracies manually, but we wanted an algorithm that does not require human intervention. Note that this list is not fixed and will be specific to each dataset. Nevertheless, our results show that this effect does not adversely affect the quality of text clustering.

We evaluated the proposed algorithm on the abovementioned three datasets with 20 key-terms set-up. The results prove that the proposed algorithm yields high quality classification (see Table 1). The most commonly used measure of classification quality is accuracy, i.e. the percentage of correctly classified documents. The quality of classification was assessed by means of accuracy:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{1}$$

where:

TP (true positives) is the number of documents with positive sentiment classified as positive,

TN (true negatives) is the number of documents with negative sentiment classified as negative,

FP (false positives) is the number of documents with negative sentiment classified as positive,

FN (false negatives) is the number of documents with positive sentiment classified as negative.

Because of the unbalanced clusters within the datasets used, the F1 measure was used, which is the harmonic mean of precision and recall:

$$F1 = \frac{2*precision*recall}{precision+recall}, \tag{2}$$

where:

$$precision = \frac{TP}{TP+FP}, \tag{3}$$

$$recall = \frac{TP}{TP+FN}, \tag{4}$$

Table 1: Detailed clustering results on three datasets

Corpus	Number of documents	Number of terms	Cluster frequencies	Percent of grouped documents	Accuracy	F1	Computing time (in seconds)
Hotels	1 888	4 385	1 667 – positive 221 - negative	95.50	94.40	96.93	15.43
Perfumes	2 604	2 675	2 333 – positive 271 - negative	92.71	94.26	96.89	16.35
Courier	4 191	4 579	3 650 – positive 541 - negative	93.35	92.84	95.86	29.95

Source: own calculation based on three datasets.

According to Table 1 the proposed two-stage algorithm achieves very high accuracy and F1 for all investigated datasets (both measures are above 90% for all cases). The highest accuracy is obtained on *Hotels* (94.40%), the lowest accuracy is obtained on *Courier* (92.84%). Similarly, in the case of F1 measure the highest score is achieved on *Hotels* (96.93%) and the lowest F1 is reported on *Courier* (95.96%).

Table 1 also presents the percent of grouped documents, i.e. documents that are grouped into one of two clusters (positive or negative). The highest percentage is obtained on *Hotels* dataset (95.50%) and the lowest is observed on *Perfumes* (92.71%). It means that the proposed algorithm does still leave from 4.5% up to 7.29% of ungrouped documents depending on the dataset.

It is worth mentioning that all considered datasets are grouped in less than a minute. All calculations were performed in R using the tm[6], word2vec[7] and deeplr[8] packages.

5. Conclusions

In the article, a novel unsupervised algorithm for determining the sentiment of text documents written in any language was proposed. The proposal was tested using the example of three corpora of documents in the Polish language. The very important task

[6] R documentation is available at https://cran.r-project.org/web/packages/tm/tm.pdf

[7] R documentation is available at https://cran.r-project.org/web/packages/word2vec/word2vec.pdf

[8] R documentation is available at https://cran.r-project.org/web/packages/deeplr/deeplr.pdf

of establishing the sentiment of Polish texts is very poorly developed in the literature on the subject, therefore the algorithm fills this gap. The novelty of the proposed algorithm includes the abandonment of any extensive usage of stoplists and lemmatizations. Instead we first translated all documents into English and then carried out a two-stage documents grouping. In the first step the algorithm establishes key-terms characteristic for the subject and using these terms, as well as a set of lexical and grammatical rules, assigns some documents to a class of positive or negative documents. In the second step, the remaining documents are attached to one of the classes by means of the rules based on the vocabulary, especially bigrams, found in the documents grouped in the first step. The algorithm was tested on three corpora of documents and achieved very good results comparable with most popular supervised neural networks for English texts (see Chifu et al., 2015 or Sharma et al., 2013). The limitations of the algorithms include unique words appearing in one-time occasions. In such cases it is impossible to overcome this limitation without the use of external sources. Other kind of limitation comes from inadequate translations, usually concerning modern slang expressions not covered by online translators. In spite of all drawbacks we strongly believe that the algorithm proposed deserves attention and further research. For the first attempt we would suggest extending the list of special grammatical structures clearly defining the sentiment independently of the subject of the text.

References

Chifu, E., Chifu, V., Letia, T., (2015). *Unsupervised Aspect Level Sentiment Analysis Using Self-Organizing Maps*, IEEE, https://doi.org/10.1109/SYNASC.2015.75.

Eder, M., Górski, R. L. (2023). Stylistic Fingerprints, POS-tags, and Inflected Languages: A Case Study in Polish. *Journal of Quantitative Linguistics*, 30(1), pp. 86–103.

Kocon, J., Milkowski, P., Zasko-Zielinska, M., (2019). *Multi-Level Sentiment Analysis of PolEmo 2.0: Extended Corpus of Multi-Domain Consumer Reviews*, Proceedings of the 23rd Conference on Computational Natural Language Learning, pp. 980–991.

Manaa, M. E., Abdulameer, G., (2018). Web Documents Similarity using k-Shingle tokens and MinHash technique. *J. Eng. Appl. Sci.*, 13, pp. 1499–1505.

Lin, C., He, Y., (2009). *Joint sentiment/topic model for sentiment analysis*, 18th ACM conference on Information and knowledge management, pp. 375–384.

Li, G., Liu, F., (2012). Application of a clustering method on sentiment analysis. *Journal of Information Science*, 38(2) pp. 127–139.

Sharma, A., Dey, S., (2013). *Using self-organizing maps for sentiment analysis*, Cornell University Library.

Souza, E., Santos, D., Oliveira, G. et al., (2020). Swarm optimization clustering methods for opinion mining. *Nat Comput*, 19, pp. 547–575, https://doi.org/10.1007/s11047-018-9681-2.

Yuqiang Tong, Lize Gu, (2018). *A News Text Clustering Method Based on Similarity of Text Labels,* Advanced Hybrid Information Processing – Second EAI International Conference, ADHIP 2018.

Zhang, W. M., Jiang, W. U., Yuan, X. J., (2010). K-means text clustering algorithm based on density and nearest neighbor, *J. Comput. Appl.*, 30(7), pp. 1933–1935.

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 187–196, https://doi.org/10.59170/stattrans-2024-035
Received – 22.10.2023; accepted – 15.05.2024

A finite state Markovian queue to let in impatient customers only during K-vacations

R. Sivasamy[1]

Abstract

We investigate a matrix analysis study for a single-server Markovian queue with finite capacity, i.e. an M/M/1/N queue, where the single server can go for a maximum, i.e. a *K* number of consecutive vacation periods. During these vacation periods of the server, every customer becomes impatient and leaves the queues. If the server detects that the system is idle during service startup, the server rests. If the vacation server finds a customer after the vacation ends, the server immediately returns to serve the customer. Otherwise, the server takes consecutive vacations until the server takes a maximum number of vacation periods, e.g. *K*, after which the server is idle and waits to serve the next arrival. During vacation, customers often lose patience and opt for scheduled deadlines independently. If the customer's service is not terminated before the customer's timer expires, the customer is removed from the queue and will not return. Matrix analysis provides a computational form for a balanced queue length distribution and several other performance metrics. We design a 'no-loss; no-profit cost model' to determine the appropriate value for the maximum value of *K* consecutive vacation periods and provide a solution with a numerical illustration.

Key words: impatient customers, vacation period, queue length, stationary distribution.

1. Introduction

The main objective of this research is to develop a matrix method to obtain the queue length distribution of a single-server service system M/M/1/N, where N represents the maximum capacity of customers waiting. This system allows the server to conditionally take up to K vacations during which the system remains inactive. However, if a server returning from vacation finds a queue of customers, the server starts service according to first-come first-served (FCFS) standards. In addition, we revisit the algorithmic approaches given by (Neuts, 1981) and (Latouche & Ramaswami, 1999) to find a solution for the class of finite two-dimensional continuous-time queue-length processes Z(t) = {L(t) , J(t); t≥0} defined in the space **E**:

$$\mathbf{E}= \mathbf{N}_0 \text{ x } \mathbf{K}_0, \mathbf{N}_0 = \{0, 1, \ldots, N, (< \infty),\} \text{ and } \mathbf{K}_0 = \{0, 1, \ldots, K\} \quad (1)$$

[1] Faculty of Social Sciences, Statistics Department, Gaborone, University of Botswana, P.bag 00705, Botswana. E-mail: sivasamyr@ub.ac.bw. ORCID: https://orcid.org/0000-0002-3158-928X.

The basic assumption is that the process **L**(t)= n (≥0) represents the observed queue length at time 't' and **J**(t) = j (j = 0, 1, . . ., (K-1)) indicates the $(j+1)^{th}$ vacation at time 't', where **J**(t) = K indicates the state of the server, whether idle or busy.

1.1. Finite state and finite capacity queue with impatient customers

Let us talk about a single server Markovian (a.k.a. Poisson) queue with a maximum capacity of N, i.e. M/M/1/N. The system services customers and the service time follows an exponential distribution with the rate μ. The customer arrival process follows the Poisson process with an average rate λ. If the server finds the system empty at a departure epoch, the server goes on vacation. If the server finds the customer at the end of the vacation, the server immediately returns to serve the customer. Otherwise, the server will make consecutive vacations until the server takes the maximum number of vacations, say K, then the server will be idle and wait to serve the next arrival.

Each vacation period is assumed to be distributed exponentially with the vacation rate ν. Suppose that during vacation periods of the server, each customer becomes impatient and activates an 'Impatient timer length' T which is exponentially distributed with parameter ψ. If the service period of the impatient client is not terminated before the client's timer expires, the client leaves the queue and does not return. Let us denote the above model by M/M/1/N/(K-vacations) queue with impatient customers.

Section 2 describes the process of limiting the length of the queue process, i.e. Z = {(L, J)} = $\lim_{t\to\infty}\{(L(t), J(t))\}$ of M/M/1/N/(K-vacations) service system with impatient customers. After formulating it as a positive recurrent process, using numerical algorithms, we obtain probability vectors for each level of process Z along with the stationary queue length distribution. Section 3 discusses the calculation of various performance metrics and probability distributions of server's vacation, busy and idle states. In addition, a "no loss; non-profit" cost function is designed to fix the maximum "K" of consecutive vacations, and a solution is also provided using numerical illustration. Section 4 provides a formal concluding report.

2. Performance of M/M/1/N/(K-vacations) queue with impatient customers in the long run

2.1. Citations

As the best members of this class, we select the M/M/1 queue, whose stationary measures were studied by (Zhang, Yue, & Yue, 2005) and impatient behavior with multiple vacations analyzed by (Ammar, 2015), (Sivasamy, 2020) and (Sudhesh & Azhagappan, 2019). It is assured from the contributions of (Kharoufeh, 2011) and (Sivasamy, Thillaigovindan, Paulraj, & Parnjothi, 2019) that the quasi birth-death (QBD) process framework could be used, in a way that is suitable for modelling all types

of M/M/1 queues. (Kharoufeh, 2011) discussed both discrete-time and continuous-time versions of the level-dependent quasi-birth-and-death (LDQBD) process that exhibits a block tri-diagonal structure.

Using standard probability arguments, we can state that the bivariate process Z = {(L, J)} forms an aperiodic, regular, and irreducible LDQBD over the state space E. We divide a two-dimensional space into a union. (N+1) levels, say L_i for i = 0, 1, …, N:

$$\boldsymbol{L_i} = ((\boldsymbol{i},\boldsymbol{0}),(\boldsymbol{i},\boldsymbol{1}),\dots(\boldsymbol{i},\boldsymbol{K}));\ \boldsymbol{i} \in \mathbf{N_0},\ \mathbf{E} = \bigcup_{\boldsymbol{i}=\boldsymbol{0}}^{\boldsymbol{N}} \boldsymbol{L_i} \tag{2}$$

where $L_i \in (2)$ is called the i^{th} level vector of size or order (K+1).

Using the properties of the QBD under FCFS rule, we organize the elements of the transition generator matrix G = (g(L_i, L_j)) described in (3):

$$\mathbf{G} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{A_0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \boldsymbol{A_2^{(1)}} & \boldsymbol{A_1^{(1)}} & \boldsymbol{A_0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A_2^{(2)}} & \boldsymbol{A_1^{(2)}} & \boldsymbol{A_0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{A_2^{(3)}} & \boldsymbol{A_1^{(3)}} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \boldsymbol{A_1^{(N-1)}} & \boldsymbol{A_0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{A_2^{(N)}} & (\boldsymbol{A_2^{(1)}} + \boldsymbol{A_0}) \end{pmatrix} \tag{3}$$

Each sub-matrix of G is a square matrix of order (K+ 1) indexed by j = 0, 1, . . . , K. The precise structures of B, A_0, A_1, and A_2 are described in (4) to (7), respectively:

$$\boldsymbol{B} = \begin{pmatrix} -(\boldsymbol{\lambda}+\boldsymbol{\gamma}) & \boldsymbol{\gamma} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -(\boldsymbol{\lambda}+\boldsymbol{\gamma}) & \boldsymbol{\gamma} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \ddots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \cdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \ddots & -(\boldsymbol{\lambda}+\boldsymbol{\gamma}) & \boldsymbol{\gamma} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & -\boldsymbol{\lambda} \end{pmatrix} \tag{4}$$

$$\boldsymbol{A_0} = \begin{pmatrix} \boldsymbol{\lambda} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\lambda} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \ddots & \vdots & \mathbf{0} \\ \vdots & \cdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \boldsymbol{\lambda} \end{pmatrix} \tag{5}$$

For n = 1,2,…, N,

$$\boldsymbol{A_1^{(n)}}= \begin{pmatrix} -(\boldsymbol{\gamma+\lambda+n\psi}) & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \boldsymbol{\gamma} \\ \mathbf{0} & -(\boldsymbol{\gamma+\lambda+n\psi}) & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \cdots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \ddots & -(\boldsymbol{\gamma+\lambda+n\psi}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & -(\boldsymbol{\lambda+\mu}) \end{pmatrix} \tag{6}$$

$$\boldsymbol{A_2^{(1)}}= \begin{pmatrix} \boldsymbol{\psi} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\psi} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \cdots & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \ddots & \boldsymbol{\psi} & \mathbf{0} \\ \mu & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{0} \end{pmatrix}, \boldsymbol{A_2^{(n)}}= \begin{pmatrix} \boldsymbol{\psi} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\psi} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\ \vdots & \cdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\psi} \end{pmatrix}, \text{ n=2, …, N} \tag{7}$$

Distribution: Let $\Pi = (\pi_0, \pi_1, \pi_2, \ldots \pi_N)$ denote the steady-state distribution of Z process and π_n the row vector $\pi_n = (\pi(n, 0), \ldots, \pi(n, K))$ associated to level L_n, $n \in N_0$.

2.2. Steady-state characteristics of the Z process

The level jumps of the Z process from sate (n, j) ∈ E are restricted only to its adjacent neighbors (n−1, j) and (n+1, j) but not to states of the form (n ± i, j) where i ≥ 2.

For each level, L_n, N ≥n≥0 in G, the diagonal elements of the matrix B and $\mathbf{A_1^{(n)}}$ are completely negative and off diagonal elements are non-negative. Matrices $\mathbf{A_2^{(n)}}$ and A_0 are not negative. In each row of G, the sum of the elements is zero (scalar). The structure of the generator matrix G reveals that the process possesses a limiting distribution of the system of equations Π G = 0, Π e =1 (scalar) where 0 denotes the zero vector. We now discuss a simple linear level reduction method in two phases for a positive recurrent case that leads to computation of the limiting distribution:

Phase 1: Iteratively reduce the state space from the level 'N' by removing one level at each step until we reach the level '0' and check if the generator of the level '**0**' corresponds to a positive recurrent Markov process. Compute the U(n) matrices and the rate matrices R_n in terms of the sub matrices of the generator G:

$$\mathbf{U_{(N)}} = \mathbf{A_1^{(N)}} + \mathbf{A_0} \tag{8}$$

$$\mathbf{U_{(n)}} = \mathbf{A_1^{(n)}} - \mathbf{A_0}[\mathbf{U_{(n+1)}^{-1}}][\mathbf{A_2^{(n+1)}}] \text{ for n = (N − 1), (N − 2), . . ., 1} \tag{9}$$

$$\mathbf{U}_{(0)} = \mathbf{B} - \mathbf{A_0}[\,\mathbf{U}_{(1)}^{-1}]\,[\,\mathbf{A}_2^{(1)}] \tag{10}$$

Phase 2: Construct the rate matrix R_n of order (Q+1), for n= 1, 2, …N:

$$\mathbf{R_n} = -\mathbf{A_0}[\,\mathbf{U}_{(n)}^{-1}] \tag{11}$$

Lemma 1: The system of matrix equations which govern all transitions of the of the Z process in terms of its steady-state probability vectors $\boldsymbol{\pi_n}$ and the known sub-matrices of its generator matrix G is derived by solving $\mathbf{\Pi}$ G = 0 and reported in (12):

$$\begin{aligned} \boldsymbol{\pi_0}\,\mathbf{B} + \boldsymbol{\pi_1}\,\mathbf{A}_2^{(1)} &= \mathbf{0} \\ \boldsymbol{\pi_{n-1}}\,\mathbf{A_0} + \boldsymbol{\pi_n}\,\mathbf{A}_1^{(n)} + \boldsymbol{\pi_{n+1}}\,\mathbf{A}_2^{(n+1)} &= \mathbf{0};\ \text{for n=1, 2, …, (N-1)} \\ \boldsymbol{\pi_{N-1}}\,\mathbf{A_0} + \boldsymbol{\pi_N}\left[\mathbf{A}_1^{(N)} + \mathbf{A_0}\right] &= \mathbf{0} \end{aligned} \tag{12}$$

Theorem 1: The unique stationary joint distribution vector $\mathbf{\Pi}$ of the queue length plus the inventory level of the Z = (L, J) process is given by $\boldsymbol{\pi_n} = \boldsymbol{\pi_{n-1}}\,\mathbf{R_n}$ for n = 1, 2, …, N and $\mathbf{\Pi}$e = 1.

Proof: The proof of this theorem is organized as an algorithm consisting of four steps:

Step 1: Re-organizing the last equation $\boldsymbol{\pi_{N-1}}\,\mathbf{A_0} + \boldsymbol{\pi_N}\left[\mathbf{A}_1^{(N)} + \mathbf{A_0}\right] =$ of (12), and using the definition $\mathbf{U}^{(N)} = \mathbf{A}_1^{(N)} + \mathbf{A_0}$ of (8), we conclude that $\boldsymbol{\pi_N} = \boldsymbol{\pi_{N-1}}\,\mathbf{R_N}$.

Step 2: Putting n = N–1 in (12), we have

$$\boldsymbol{\pi_{N-2}}\,\mathbf{A_0} + \boldsymbol{\pi_{N-1}}\,\mathbf{A}_1^{(N-1)} + \boldsymbol{\pi_N}\,\mathbf{A}_2^{(N)} = \mathbf{0} \tag{13}$$

Re-organizing the equation (13) with the substitution of $\boldsymbol{\pi_N} = \boldsymbol{\pi_{N-1}}\,\mathbf{R_N}$, we obtain that

$$\begin{aligned} &\boldsymbol{\pi_{N-2}}\,\mathbf{A_0} + \boldsymbol{\pi_{N-1}}\,\mathbf{A}_1^{(N-1)} + \boldsymbol{\pi_{N-1}}\,\mathbf{R_N}\,\mathbf{A}_2^{(N)} = \mathbf{0} \\ \Rightarrow\ &\boldsymbol{\pi_{N-1}}\,[\mathbf{A}_1^{(N-1)} + \{-\mathbf{A_0}\left[\mathbf{U}^{(N)}\right]^{-1}\}\,\mathbf{A}_2^{(N)}] = \boldsymbol{\pi_{N-1}}\,[\mathbf{U}^{(N-1)}] = -\boldsymbol{\pi_{N-2}}\,\mathbf{A_0} \\ \Rightarrow\ &\boldsymbol{\pi_{N-1}} = \boldsymbol{\pi_{N-2}}\,\mathbf{R_{N-1}}\ \textbf{on using the fact}\ \ \mathbf{R_{N-1}} = -A_0\ [\mathbf{U}^{(N-1)}]^{-1} \end{aligned} \tag{14}$$

Step 3: Continuing the similar iterative process for n = (N-2), (N-1), … 1 and using the results of step preceding step, we can establish that $\boldsymbol{\pi_n} = \boldsymbol{\pi_{n-1}}\,\mathbf{R_n} = \boldsymbol{\pi_0}\sum_{k=1}^{n}\mathbf{R_k}$ for n = 1, 2, …, N.

Step 4: To find the vector $\boldsymbol{\pi_0}$, the normalizing condition is $\boldsymbol{\Pi}\mathbf{e} = 1$, we use the following steps.

i) Solve $\boldsymbol{\pi_0}\,\mathbf{U^{(0)}} = \mathbf{0},\ \boldsymbol{\pi_0}\mathbf{e} = 1$.

ii) Compute $\boldsymbol{\pi_n} = \boldsymbol{\pi_{n-1}}\,\mathbf{R_n}$ for n = 1, 2, ..., N.

iii) Calculate $\boldsymbol{\Pi}\mathbf{e}$ and renormalize $\boldsymbol{\Pi}$ using $\boldsymbol{\Pi} = (\boldsymbol{\Pi} / \boldsymbol{\Pi}\mathbf{e})$.

Thus, now the steady-state probability vector $\boldsymbol{\Pi}$ of the Z process is completely determined. We can now discuss the steady-state probabilities of various events, such as the conditional probability that a server is on vacation, busy, or idle, etc., and measures of system performance. In addition, we can investigate the optimal number of vacations to minimize the average operating cost of the system.

3. Performance Measures: M/M/1/N/(K-vacations) queue with impatient customers

Let us compute the conditional mean system size $\overline{\mathbf{Q}}_\mathbf{j}$ given J = j of the $(j+1)^{th}$ vacation for j = 0,1, . . . ,(K-1), K:

$$\overline{\mathbf{Q}}_\mathbf{j} = \sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{n}\,\boldsymbol{\pi}(\mathbf{n},\mathbf{j}),\ \mathbf{j} = 0, 1, 2, \ldots, K \tag{15}$$

The conditional probability P_V that the server is on vacation:

$$\mathbf{P_V} = \sum_{\mathbf{n=0}}^{\mathbf{N}} \sum_{\mathbf{j=0}}^{\mathbf{K-1}} \boldsymbol{\pi}(\mathbf{n},\mathbf{j}) \tag{16}$$

The conditional probability P_B that the server is busy:

$$\mathbf{P_B} = \sum_{\mathbf{n=1}}^{\mathbf{N}} \boldsymbol{\pi}(\mathbf{n},\mathbf{K}) \tag{17}$$

The conditional probability P_I that the server is idle:

$$\mathbf{P_I} = \mathbf{1} - \mathbf{P_B} - \mathbf{P_V} = \sum_{\mathbf{n=1}}^{\mathbf{N}} \boldsymbol{\pi}(\mathbf{0},\mathbf{K}) \tag{18}$$

The expectation is that $P_V + P_B + P_I = 1$. The conditional probability P_{lost} of lost customers is given by:

$$\mathbf{P_{lost}} = \sum_{\boldsymbol{j=0}}^{\boldsymbol{K}} \boldsymbol{\pi}(\boldsymbol{N},\boldsymbol{j}) \tag{19}$$

The effective arrival rate λeff is given by:

$$\boldsymbol{\lambda_{eff}} = \frac{\boldsymbol{\lambda}}{\mathbf{1} - \mathbf{P_{lost}}} \tag{20}$$

The unconditional mean system size $\bar{\mathbf{Q}}$ is given by

$$\bar{\mathbf{Q}} = \sum_{\mathbf{j=0}}^{\mathbf{K}} \bar{\mathbf{Q}}_{\mathbf{j}} = \sum_{\mathbf{j=0}}^{\mathbf{K}} \sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{n}\, \boldsymbol{\pi}(\mathbf{n}, \mathbf{j}) \tag{21}$$

The mean waiting time $\bar{\mathbf{W}}$ can be calculated by Little's law. Thus,

$$\bar{\mathbf{W}} = \frac{\bar{\mathbf{Q}}}{\boldsymbol{\lambda_{eff}}} \tag{22}$$

3.1. Numerical illustrations

We now discuss the numerical calculations of some of the measurements discussed so far in the previous discussions. The crux of the calculation lies in the calculation of the stationary probability vectors $\{\pi_n;\ n = 0, 1, \ldots, N\}$ because there is no explicit expression for each probability function π_n. Calculations are based on the algorithm given from the matrix equations reported in Theorem 1.

Any external observer can find the server in one of the three mutually exclusive states i.e. "V: Vacation, B: Busy and I: Idle". Let the probabilities of these three states V, B and I be P_V, P_B, and P_I respectively. Then, P_V, P_B, and P_I ($P_V + P_B + P_I = 1$) values can be easily calculated and checked from the stationary probability vector Π satisfying the conditions ΠG = 0 and Π e = 1.

Using server state distribution $\{P_V, P_B, \text{and } P_I\ (P_V + P_B + P_I = 1)\}$, we now notice an issue related to fee collection and loss of rental owner. Assume the server is a leased machine. Let us say the server rental is USD (245.25, 80.8, 125.5) per unit time. Let the vacation state of the server cost be 425.25 USD per unit time to manage loss due to impatient customers and non-availability of service which each customer spends on vacation. Suppose that the management collects profit USD 80.8 per unit time from customers when the server is busy and USD 125.45 when server is idle. The objective of the management is to fix the maximum number 'K' of vacations for the server to take consecutively, which ensues 'no loss and no gain' status. For this experiment, random values of input quantities are selected as λ = 2.2, ψ = 1.1, ν = 2.25 and μ = 2.5. For j = V, B, and I, let us calculate the vector of probabilities $V(j) = (P_j^{(K=1)}, P_j^{(K=2)}, \ldots, P_j^{(K=9)}, P_j^{(K=10)})$, respectively for K = 1, 2,. . . , 10.

Let us compute the following costs:

- $T1(K) = 425.25\ V_{(1)}$, called vacation loss cost;
- $T2(K) = 80.6\ V_{(2)}$, called the busy server profit;
- $T3(K) = 125.45\ V_{(3)}$, called the idle server profit.

The objective is now finding a K* value such that the total cost T4(K*) = 0:

$$T4(K) = (T2(K) + T3(K)) - T1(K) > 0 \text{ for all } K < \boldsymbol{K}^*$$

$$T4(K) = (T2(K) + T3(K)) - T1(K) < 0 \text{ for all } K > \boldsymbol{K}^* \quad (23)$$

The general trend is that the function T1(K) increases as the values of K increase. On the other hand, both functions T2(K) and T3(K) decrease as the values of K increase. We call the total costs T4(K) = T2(K) T3(K) – T1(K). If any of the components of T4(K) >0, it gives a profit or a loss for that value of K.

To demonstrate these facts and to get an optimum on K to meet the no-loss no-profit condition, numerical values are computed for the given data set and the corresponding curve is plotted in Figure 1.

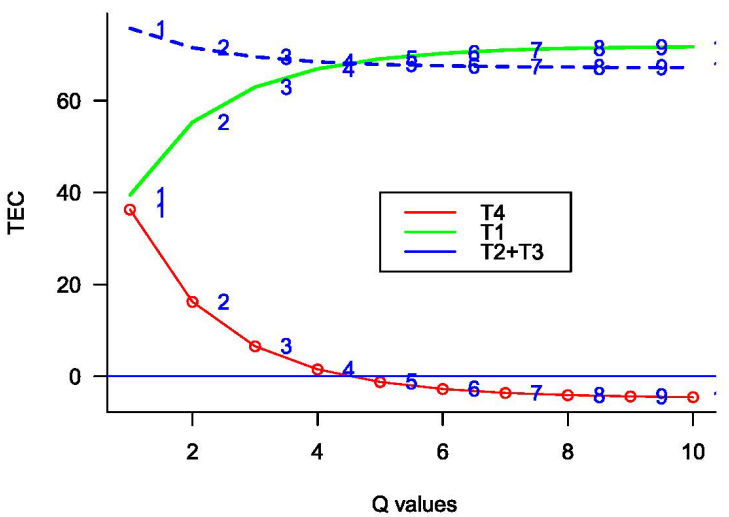


Figure 1: Relationships among Vacation, Busy and Idle states of the server's cost for "no loss and No gain level as K = 4

A simple look at Figure 1 tells us that an optimum of vacations is attained if K* = 4. This means that if management allows K = 1, 2 and 3 consecutive vacations, management earns a positive profit. There is no profit or loss from the maximum number of vacation periods K* = 4. But if K ≥ 5, the lead is only a negative profit or loss. This type of test can also be designed to monitor a typical K-value, which provides "no loss, no gain" efficiency while adding a larger number of cost effects.

4. Conclusions

A single-server M/M/1/N/(K-vacations) queue with impatient customers, which conditionally accepts impatient customers in every vacation period, is well studied by matrix analysis. The special thing is that if the customer's service does not end before the random deadline chosen by the customer, he can leave the queue. Additionally, the server can take multiple vacations in a row, but no more than K vacations.

The steady-state results of this study are supported by numerical algorithms to obtain the necessary probability vectors and the optimal number of $K^{\star}$ for consecutive vacations. We obtain the steady-state queue length distribution and the scalar value of the vector expression for multiple events and measurements. A "no Loss; no Profit" cost model is proposed to test the appropriate value for the maximum K-value of consecutive vacations and provide a solution using a numerical representation.

The proposed methodology is implemented by showing an exponential distribution of arrival times, service times, impatient customer deadlines and vacation periods. Our opinion is thatthe proposed theoretical and computational aspects of single-server Poisson queue M/M/1/N are useful not only for academics but also for all practitioners dealing with queues that have many vacations.

Future scope can be expanded by replacing exponential distribution with general distribution in single server or multi-server queues.

Acknowledgement

The author thanks his colleagues for their constructive suggestions and thanks the institutions of the University of Botswana for encouraging this research without financial support.

References

Ammar, S., (2015). Transient analysis of an M/M/1 queue with impatient behavior and multiple vacations. *Applied Mathematics and Computation, 260*, doi: 10.1016/j.amc.2015.03.066, pp. 97–105.

Kharoufeh, J. P., (2011). *Level-Dependent Quasi-Birth-and-Death Processes,* doi: 10.1002/9780470400531.eorms0460: Wiley Encyclopedia of Operations Research and Management Science.

Latouche, G. A., Ramaswami, V., (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling.* Pennsylvania: Society for Industrial and Applied Mathematics, Philadelphia.

Neuts, M., (1981). *Matrix geometric solutions in stochastic models—an algorithmic approach,* Baltimore: John Hopkins University Press.

Sivasamy, R., (2020). Two Server Poisson Queues with a Slow Service Provider for Impatient Customers. *International Journal of Mathematics and Statistics, 2,* ISSN 0974–7117 (Print); ISSN 0973-8347 (Online), https://www.ej-math.org, pp. 57–73.

Sivasamy, R., Thillaigovindan, N., Paulraj, G. A., and Parnjothi, N., (2019). Quasi birth-death processes of two-server queues with stalling. *OPSEARCH, 56,* https://doi.org/10.1007/s12597-019-00376-1, pp. 739–756.

Sudhesh, R., Azhagappan, A., (2019). Transient solution of an M/M/∞ queue with system's additional tasks and impatient customers. *Int. J. of Mathematics in Operational Research, 16, No. 1,* https://dblp.org/db/journals/ijmor/index, pp. 82–97.

Zhang, Y., Yue, D. A., and Yue, W., (2005). Analysis of an M/M/1/N queue with Balking, Reneging and Server Vacations. *International Symposium on OR and Its Applications, 1,* https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje)), pp. 37–67.

STATISTICS IN TRANSITION new series, September 2024
Vol. 25, No. 3, pp. 197–201

About the Authors

Abdollahi Nanvapisheh Anita is a PhD student at the Department of Statistics, Faculty of Science, Razi University. Her main areas of interest include distribution theory, regression, general linear model, non-linear model, optimal design, Bayesian and nonparametric Bayesian analysis. She has published more than 30 research papers in international/national journals and conferences. She has also published three scientific books in the statistical field. She is an active member of many scientific professional bodies including Iranian statistical society.

Ayodeji Idowu Oluwasayo is an Associate Professor of Statistics at the Department of Mathematics, Faculty of Science, Obafemi Awolowo University, Nigeria. Her main areas of interest include time series analysis and multivariate analysis. She is a 2017 alumnus of the Lindau Nobel laureate Meetings on Economic Sciences and a 2024 fellow of the Ife Institute of Advanced Studies. Idowu currently serves as an Associate Editor of the Ife Journal of Science.

Bayoud Husam A. is an Associate Professor of Statistics. His research interests include bioequivalence studies, information theory, reliability analysis and Bayesian analysis. Dr. Bayoud has published more than 25 research papers in international journals and conferences. He has also co-supervised three PhD students in Statistics. He is an active member of many scientific professional bodies.

Bharti has completed her PhD from the Department of Mathematics and Computing at Dr B R Ambedkar National Institute of Technology, Jalandhar, India in 2023 under the supervision of Dr R R Sinha. Dr Bharti has eight research papers published in national and international journals and has presented seven research papers at national and international conferences. She is currently an Assistant Professor at the Department of Mathematics at DAV University, Jalandhar, India.

Białek Jacek is an Associate Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. Simultaneously he holds the position of an expert in Statistics Poland at the Department of Trade and Services. His main areas of interest include price index theory, inflation measurement and implementation of scanner data in the CPI/HICP by using R, Python and SQL. He is an author of the PriceIndices R package.

Brzozowska-Rup Katarzyna is an Assistant Professor at the Department of Economics and Finance, Faculty of Management and Computer Modelling, Kielce University of Technology, Poland. She also serves as an expert consultant for the Statistical Office in Kielce, contributing to research on modelling the shadow economy. Her research activities encompass statistical and econometric methods applied to the unobserved economy, migration, and renewable energy.

Czapkiewicz Anna is an Associate Professor at the Department of Mathematical Applications in Economics at the AGH University in Krakow, Poland. She is also an expert at the Regional Statistical Office in Kielce, Poland. Her research interests are econometrics, financial markets, statistics, applications of mathematics in economics. Anna Czapkiewicz has published more than 60 research papers in international/ national journals and conferences.

Hasilová Kamila is an Associate Professor at the Department of Quantitative Methods of the University of Defence. She specializes in the statistical methods and mathematical modelling, with a focus on data analysis and non-parametric statistical methods applied in the area of reliability. She has published more than 50 research papers in peer-reviewed journals and international conferences. She is a member of the European Safety and Reliability Association, Technical Committee "Mathematical and Computational Methods in Reliability and Safety".

Hilow Hisham M. is Jordanian Professor of Statistics born in 1951. He got his PhD degree from Virginia Tech University in 1985. He started his educational career as an Assistant Professor at the Mathematics Department of the University of Jordan. The department has Bachelor, Master and PhD programs. Dr. Hilow was later promoted to the associate professorship rank and finally to full professorship. Dr. Hilow has supervised the thesis of around 20 Master students and the dissertations of three PhD students. Dr. Hilow was a member of dozens of Master and PhD defense committees. He has published around 40 research articles in well-renowned statistical journals. His research interest is mainly in design and analysis of experiments but some of his research is in mathematical statistics as well as in statistical applications. Dr. Hilow chaired the Mathematics Department of the University of Jordan from 2005 to 2009.

Horová Ivana is a Full Professor of Applied Mathematics. She is dealing with non-parametric methods for data analysis. Her main focus is on kernel smoothing techniques. Professor Horová has published 125 papers in international/national journals and presented her results at many international conferences. With her colleagues, she published the monograph “Kernel smoothing in MATLAB” (2012). She was the supervisor of 15 successful PhD students.

Idczak Adam is an Assistant at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. His current research interests are in the field of text mining, particularly methods for classifying and grouping text documents by sentiment. He has several publications to his credit in the field of statistical methods of text analysis. At the university, he teaches courses in statistics, survey sampling, small area estimation and credit scoring. He works as an expert on credit risk parameter modeling at the Polish bank Pekao S.A.

Jafari Habib is an Associate Professor at the Department of Statistics, Faculty of Science, Razi University. Simultaneously he holds the position of Secretary of the Scientific Committee of Iran Statistics Student Competition. His main areas of interest include regression, GLM, optimal design, conjoint optimal design and Bayesian optimal design. Dr. Jafari has published more than 90 research papers in international/ national journals and conferences. He is an active member of the Iranian statistical society.

Khazaei Soleiman is an Assistant Professor at the Department of Statistics, Faculty of Science, Razi University. His main areas of interest include Bayesian analysis, non-parametric analysis, non-parametric Bayesian analysis. Dr Khazaei has published many research papers in international/national journals and conferences. He is an active member of the Iranian statistical society.

Korzeniewski Jerzy is an Associate Professor at the Department of Demography, Faculty of Economics and Sociology, University of Lodz. His main field of scientific work is cluster analysis and its applications. The applications have quite a wide spectrum, for example, the analysis of the effectiveness of stock exchange investments or clustering Polish poviats. For many years he has worked on using cluster analysis to develop machine learning methods for studying linguistic texts. He has several publications on this subject including, e.g. a monograph devoted to the methods of analyzing the sentiment of linguistic texts.

Mpinda Berthine Nyunga is a Machine Learning (ML) research fellow at the University of Tübingen, Germany. Her research interests include medical image analysis, ML and Deep Learning models for medical images, and the explainability and interpretability of these models. With a background in mathematics and Machine Learning, she has served as a teaching assistant at the African Institute for Mathematical Sciences (AIMS) in Cameroon. Berthine is a co-author of several published papers and book chapters in international journals and conferences. She is also a member of the organizing committee for the Deep Learning Indaba conference and an organizer of the Women in Machine Learning and Data Science Kinshasa Chapter in the DRC. Adam Szulc is an Associate Professor at the Institute of Statistics and Demography of the Warsaw School of Economics. His fields of interests cover poverty and inequality

measurement, consumer demand systems, equivalence scales, social policy evaluations, and matching estimation. He is a member of the review boards of the Equilibrium - Quarterly Journal of Economics and Economic Policy and of the Argumenta Oeconomica.

Nitha K. U. works as a statistical assistant at the Department of Economics and Statistics. Her research interested areas are time series analysis and distribution theory.

Olawale Awe O. holds a PhD in Statistics from the University of Ibadan and an MBA from Obafemi Awolowo University, Nigeria. He is the Vice President of the International Association for Statistics Education (IASE) and an Elected Council Member of the International Statistics Institute (ISI). He also serves as Vice President of Global Statistical Engagements for the LISA 2020 Global Network, USA, and is a research professor and machine learning team leader at the Statistical Learning Laboratory (SaLLy), Federal University of Bahia, Brazil. Awe has published over 100 research papers and five books/monographs. As the pioneering LISA Fellow at the University of Colorado, Boulder, USA, he has made significant contributions to the global statistical community.

Qubbaj Huda H. is an Assistant Professor of Mathematics and Statistics, who has graduated recently from the University of Jordan. Currently, she holds the position of Head of logistics department at the Cell therapy center. Her areas of interest are biostatistics, statistical inference and data analysis in particular.

Sinha R. R. received his PhD in Sampling Techniques from the Department of Statistics at Banaras Hindu University, Varanasi, India, in 2001. He is currently an Associate Professor at the Department of Mathematics and Computing at Dr B R Ambedkar National Institute of Technology, Jalandhar, India. Along with overseeing the dissertations of PG candidates, he has guided two PhD and three MPhil candidates. Dr Sinha has published more than 35 research articles and six book chapters in national and international journals and proceedings on various topics related to sampling techniques. In addition to actively serving on editorial boards and as a reviewer for esteemed journals and academic societies, Dr Sinha has edited two books on "Statistical Modeling and Applications on Real-Time Problems".

Sivasamy R. is a Full Professor of Statistics at the University of Botswana. He is an internationally published researcher specializing in stochastic processes and their applications to queuing and optimization problems. He frequently partners with other researchers on research projects and supports researchers interested in Markov chain research. He has written several book chapters and monographs on queue inventory systems and pair trading strategies. Over one hundred publications in international journals are his own or co-authored works. He has lectured or presented at major conferences and edited a few papers.

Szulc Adam is an Associate Professor at the Institute of Statistics and Demography of the Warsaw School of Economics. His fields of interests cover poverty and inequality measurement, consumer demand systems, equivalence scales, social policy evaluations, and matching estimation. He is a member of the review boards of the Equilibrium – Quarterly Journal of Economics and Economic Policy and of the Argumenta Oeconomica.

Vališ David is a Full Professor of Mechanical Engineering. His research interests are oriented into the area of reliability, risk and safety – specifically to reliability tests and accelerated reliability tests, degradation modelling and condition assessment. Professor Vališ has published more than 100 research articles/papers in international/national journals and conferences. He has also published three books/monographs or chapters in them. Professor Vališ is an active member of scientific professional bodies such as European Safety and Reliability Association, International Electrotechnical Committee – Technical Committee 56 "Dependability" or International Standardisation Organisation – Technical Committee 262 "Risk Management".

Wanjohi Jane Wangui is a highly motivated data scientist who has recently graduated with a Structured Master's in mathematical science - Data science at the African Institute of Mathematical Sciences. During her academic journey, she enjoyed courses focused on inferential statistics, machine learning, and network science. Through various projects, she has cultivated programming skills in Python and gained practical experience in applying algorithms to solve complex problems.

Zámečník Stanislav is a PhD student at the Faculty of Science of Masaryk University. His research is focused on non-parametric methods, especially kernel smoothing, spherical data, and data analysis. Currently, he is completing his thesis on the topic of his research and has published several papers in this area.