**Volume 25, Number 1, March 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**

Scientific Secretary
Marek Cierpiał-Wolan, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: m.cierpial-wolan@stat.gov.pl*
Managing Editor
Adriana Nowakowska, *Statistics Poland, Warsaw, e-mail: a.nowakowska3@stat.gov.pl*
Secretary
Patryk Barszcz, *Statistics Poland, Warsaw, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 – 608 33 66*
Technical Assistant
Rajmund Litkowiec, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: r.litkowiec@stat.gov.pl*



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

## CONTENTS

*STATISTICS IN TRANSITION* new series, *March 2024*
*Vol. 25, No. 1, pp. III*

# Submission information for Authors

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

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

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

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

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

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

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

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

*STATISTICS IN TRANSITION* new series, *March 2024*
*Vol. 25, No. 1, pp. V–VI*

# Policy Statement

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

***

## Abstracting and Indexing Databases

*Statistics in Transition new series* is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

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

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Electronic Journals Library

Elsevier – Scopus

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

Genamics JournalSeek

Google Scholar

Index Copernicus

J-Gate

JournalGuide

JournalTOCs

Keepers Registry

MIAR

Microsoft Academic

OpenAIRE

ProQuest – Summon

Publons

QOAM (Quality Open Access Market)

ReadCube

RePec

SCImago Journal & Country Rank

TDNet

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

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. VII–X*

# From the Editor

The March issue of *Statistics in Transition new series* presents readers with a set of eleven articles written by twenty-four authors from nine countries (in order of appearance): USA, Poland, Nepal, Indie, Nigeria, Indonesia, Iran, Algeria, and Saudi Arabia. This volume opens with an *invited article* by Piotr Kokoszka, Mengting Lin, HaonanWang, and Stephen Hayne.

## Invited paper

In the first paper, ***Statistical risk quantification of two-directional internet traffic flows,*** **Piotr Kokoszka, Mengting Lin, HaonanWang,** and **Stephen Hayne** discuss recent developments in statistical methodology for the quantification of risk of source-destination pairs in an internet network within the framework of functional data analysis and copula modeling. It was summarized in the form of computational algorithms that use bidirectional source-destination packet counts as input. The usefulness of the proposed approach was evaluated by an application to real internet traffic flows and via a simulation study. The performance, and relative performance, of the two algorithms using simulated data that has certain features of real data sets, but also certain characteristics that are known targets, were assessed.

## Research articles

**Henryk Gurgul's** and **Robert Syrek's** article ***Mutual information between Polish subindexes – the use of copula entropy around the time of the COVID-19 pandemic*** demonstrates application of the copula theory to describe the dependence structure between variables, while the information theory provides the tools necessary to measure the uncertainty associated with these variables. What both theories have in common is copula entropy, which is strictly related to mutual information. The findings of this study of the dependence of the (sub)indexes of the Polish stock market during the pandemic period seem to be useful not only to investors in Poland, but in other countries as well, especially in Central Europe, in making investment decisions. The results of calculating the interdependencies between WIG, sectoral indexes and among sectoral indexes of the Polish economy using copula entropy and Pearson's correlation are quite different.

The next article, by **Arjun Kumar Gaire** and **Yogendra Bahadur Gurung,** entitled ***Skew Log-logistic distribution: properties and application***, presents a novel three-parameter skew-log-logistic distribution. It starts from the development of a new random variable based on Azzalini and Capitanio's (2013) proposition, including various statistical properties of this distribution. A maximum likelihood method for estimating the distribution's parameters is employed. The density function exhibits unimodality with heavy right tails, while the hazard function exhibits rapid increase, unimodality, and slow decrease, resulting in a right-skewed curve. Furthermore, four real datasets are utilized to assess the applicability of this new distribution. The AIC and BIC criteria are employed to assess the goodness of fit, revealing that the new distribution offers greater flexibility compared to the baseline distribution.

In the paper ***A chain ratio-type exponential estimator for population mean in double sampling,*** by **Nirupama Sahoo** and **Sananda Kumar Jhankar,** an efficient ratio-type exponential estimator for estimating the population mean by incorporating two auxiliary variables in two-phase (double) sampling is proposed. The bias and the mean square error of the presented estimator have been obtained up to the first order of approximation. The new estimator offers more precision in comparison to other competing estimators, theoretically as well as empirically, by considering a known value of some population parameter.

**Abimibola V. Oladugba's** and **Oluwagbenga T. Babatunde's** paper, ***Improved calibration estimation of population mean in stratified sampling using two auxiliary variables***, discusses possibilities to improve the standard estimator of the population mean in a stratified sampling through calibration estimation approach using two auxiliary variables. A simulation study was carried out to evaluate the performance and efficiency of an estimator with respect to three estimators proposed in the literature for estimating the population mean in a stratified sampling (using two auxiliary variables). The proposed estimator has the least absolute relative bias and mean square error for all the cases under consideration. The results showed that the new estimator proved to be more efficient than the three existing estimators considered.

**Agnieszka Palma** and **Dorota Kałuża-Kopias** in the paper ***Inter-voivodship migration in Poland in the 2000–2020 period based on Markov chain analysis*** deal with the scale and directions of inter-voivodship migration in Poland in selected years of the 2000–2020 period. The study focused on permanent residence migration and aimed to identify areas of migration attractiveness and migration catchment voivodships. The application of the Markov Chain allowed for evaluation of the population flow between voivodships. The results of the study indicate that the most favourable situation remains in the Mazowieckie voivodship, which is considered the most attractive area for people from other regions of the country – mainly from the

Lubelskie, Podlaskie, and Łódzkie voivodships, and to a lesser extent from the Świętokrzyskie and Warmińsko-Mazurskie voivodships. The approach used allowed for determining the properties of the transition probability matrix as well as stationary probability in order to characterise the mechanism of inter-voivodship migrations in the years 2000, 2010 and 2020.

In the next article, ***Implementation of K-Nearest Neighbor using the oversampling technique on mixed data for the classification of household welfare status,*** **Nur Mutmainnah Djafar** and **Achmad Fauzan** took up the task to classify the household welfare status in Kulon Progo using the K-Nearest Neighbor (KNN) method. Since imbalance was found between the poor and non-poor categories, an oversampling technique was employed. Imbalanced data affect classification, especially when it comes to predicting the results of the classification. The following oversampling techniques were employed: Random Oversampling (RO), the Adaptive Synthetic (ADASYN) and the Synthetic Minority Oversampling Technique (SMOTE). It was found that, of the three techniques, RO was the most efficient with $k = 5$, which yielded the best performance in terms of sensitivity, specificity, the G-mean. Therefore, it can be concluded that the classification model performed well enough to classify household welfare status, especially among the poor (minority group).

The paper ***On Bayesian inference of reliability parameter in Burr-type XII model based on imprecise data: a survey on fuzzy modelling*** by **Iman Makhdoom** and **Abbas Pak** examines the classical and Bayesian inference procedures for the BT XII distribution parameters, including the corresponding reliability parameter when the available data are described regarding fuzzy numbers. In this context, the authors considered three priors as noninformative prior, i.e. $a_1 = b_1 = a_2 = b_2 = 0$, less informative prior, i.e. $a_1 = b_1 = a_2 = b_2 = 0.01$, and informative prior, i.e. $a_1 = b_1 = a_2 = b_2 = 4$. Considering the criterion MSE for all methods, with increasing $n$, the estimates are improved. The performance of the Bayes estimates with assumptions of noninformative prior and less informative prior regarding AVs and MSEs is almost identical. The simulation study for all methods shows that the estimate of $R$ is satisfactory, even for samples with sizes small and moderate. Using the NR or EM algorithms for the computation of MLEs gives similar estimation results.

**Joanna Kisielińska's** article presents ***Estimation of quantiles with the exact bootstrap method.*** The estimation uses the bootstrap method in the so-called exact. Three bootstrap estimators were used: two of them based on one order statistic, and the third on a linear combination of two order statistics (for an integer). It has been shown that there is no general form of the distribution of the exact bootstrap estimator based on two order statistics. However, it is possible to calculate such a distribution – the algorithm that performs such a task is presented. The bootstrap confidence intervals

were constructed using the exact percentile method. It has been shown that if the estimator is based on a single order statistic, it is known in advance which elements of the primary sample are the limits of the confidence intervals (so there is no need to resample). The intervals determined by the exact percentile method were compared with those constructed using other methods.

**Research Communicates and Letters**

**Thara Belhamra, Halim Zeghdoudi,** and **Vinoth Raman** analyse ***Reliability for Zeghdoudi distribution with an outlier, fuzzy reliability and application.*** This study focuses on estimating reliability P[Y<X], where $Y$ has a Zeghdoudi distribution with parameter $a$, $X$ has a Zeghdoudi distribution with one outlier present and parameter $c$, and the remaining $(n - 1)$ random variables are from a Zeghdoudi distribution with parameter $b$, in order for $X$ and $Y$ to be independent. Several findings of a simulation study and the maximum likelihood estimate of $R$ are provided. Some results related to fuzzy dependability were also presented. In order to demonstrate the adaptability of the Zeghdoudi distribution, the authors use real data on the survival times (in days) of 72 Algerian people who were infected with coronaviruses, and then compared the outcomes with those of other distributions. Studies have been done on the maximum likelihood estimator for $R$ and fuzzy dependability.

**Piyush Kant Rai's** and **Sweta Singh's** paper, ***Composite estimators for domain estimation and sensitivity performance interval of their weights*** presents some composite estimators based on various combinations of two different existing estimators. To account for the absence of optimum weights, the sensitivity performance intervals for weights with respect to the proposed composite estimators were obtained and the sensible values of the involved weights have been determined. The aim of this procedure was to confine the superiority for different composite combinations – i.e. simple direct vs. direct ratio, simple direct vs. synthetic ratio and direct ratio vs. synthetic ratio composite estimators – as compared to the existing estimators. It was concluded that the composite estimators for the weights lie in the sensitivity performance intervals that are less varying in terms of MSE. The outcomes of the study will be useful to develop efficient composite estimators for the estimation domain in general, and for small area estimation in particular.

**Włodzimierz Okrasa**
Editor

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 1–22,* https://doi.org/10.59170/stattrans-2024-001
Received – 11.09.2023; accepted – 05.12.2023

# Statistical risk quantification of two-directional internet traffic flows

**Piotr Kokoszka**[1], **Mengting Lin**[2], **Haonan Wang**[3], **Stephen Hayne**[4]

## Abstract

We develop statistical methodology for the quantification of risk of source-destination pairs in an internet network. The methodology is developed within the framework of functional data analysis and copula modeling. It is summarized in the form of computational algorithms that use bidirectional source-destination packet counts as input. The usefulness of our approach is evaluated by an application to real internet traffic flows and via a simulation study.

**Key words:** Copula, Functional data, Internet traffic, Principal components, Risk quantification.

## 1. Introduction

Malicious cyberattacks have emerged as a growing threat to economic performance and national security. They can be launched by criminal organizations or autocratic governments. A significant challenge facing the internet security community is to develop algorithms that can automatically detect abnormal network access patterns. Attackers use many different techniques, such as distributed denial of service attacks (DDoS), intrusions that lead to the installation of malware for exfiltration or ransomware intrusion, misconfigured servers for reflection and amplification attacks. By sending a misconfigured server request using a spoofed IP address, the server will unknowingly bombard the target with a frequency 50 or more times higher than that of the response. Attacks of various types have been subjects of extensive research, with thousands papers on the above topics. Some representative recent contributions are Dong and Sarem (2019), Nishanth and Mujeeb (2020), Sambangi and Gondi (2020) and Awan *et al.* (2021).

In this paper, we propose statistical methodology aimed at detecting attacks manifested as unusual traffic between a source and a destination IP addresses. Our focus is on identifying such pairs and ranking them according to the threat they may pose. Related papers, focusing on outlier detection in multivariate functional data, are Dai and Genton (2018) and

[1]Department of Statistics, Colorado State University, Fort Collins CO 80523, USA.
E-mail: Piotr.Kokoszka@colostate.edu. ORCID: https://orcid.org/0000-0001-9979-6536.
[2]Department of Statistics, Colorado State University, Fort Collins CO 80523, USA.
ORCID: https://orcid.org/0009-0002-7712-9585.
[3]Department of Statistics, Colorado State University, Fort Collins CO 80523, USA.
ORCID: https://orcid.org/0000-0002-8892-6232.
[4]Department of Computer Information Systems, Colorado State University, Fort Collins CO 80523, USA.
ORCID: https://orcid.org/0000-0002-9578-3364.

Amovin-Assagba *et al.* (2022). Dai and Genton (2018) propose graphical tools for identifying the set of potentially outlying curves by taking into account unusually large magnitudes and/or shapes. They do not rank the pairs, even though this might be possible by elaborating on their approach. Amovin-Assagba *et al.* (2022) also focus on identifying the set of outlying pairs, but do not rank them in any way. They postulate a specific model motivated by the industrial application they consider. Such a model, and the clustering technique they use, need not be suitable for the data we consider. Basically, related existing approaches focus on identifying the set of outliers rather than assigning numerical measures of separation from most curves.

Our method is based on multivariate functional principal components and copula modeling. Internet streaming data are recorded at densely spaced time points, so they can be modeled as densely observed functions. This suggests that functional data analysis (FDA) approaches might be suitable. Following the monographs of Bosq (2000), Ramsay and Silverman (2005) and Ferraty and Vieu (2006), FDA has grown into a mature field of statistics. Its advantage over competing approaches is that all information in the time series of traffic traces, e.g. shape, variation, and timing, can be taken into account. Functional principal component analysis (FPCA) is a statistical method used to uncover main patterns in functional data, see e.g. Chapter 11 of Kokoszka and Reimherr (2017). FPCA is a powerful dimension reduction, or feature extraction, tool when a sample of functions from a single population is observed. In our setting, we are dealing with bidirectional traffic flows, so we need an analog of FPCA for samples whose elements are pairs of functions. A suitable tool is therefore Multivariate (bivariate in our case) FPCA. Such methods have recently been studied by Happ and Greven (2018), Górecki *et al.* (2018), Krzyśko and Smaga (2020, 2021), even though earlier related work exists, e.g. Berrendero *et al.* (2011), Jacques and Preda (2014), Chiou *et al.* (2014).

A copula describes the joint distribution of random vectors with standard uniform marginal distributions. Many excellent monographs are available, e.g. Nelsen (2006), Joe (2015), Hofert *et al.* (2018) and Czado (2019). A copula model decomposes a multivariate distribution function into two elements: the marginal distributions and the copula which captures the dependence relationship of the marginals. In recent years, copulas have been used to handle multivariate cybersecurity risks, e.g. Peng *et al.* (2018), and for predicting the effectiveness of cyber defense early-warning, e.g. Xu *et al.* (2017). Both FPCA and copula modeling show flexibility and efficiency that we also demonstrate for our methodology that combines and suitably refines them for our task.

To summarize our contribution, this paper develops statistical methodology to identify IP addresses of source-destination pairs that exhibit unusual and suspicious behavior and quantify their cybersecurity risks. We use the term risk to refer to the level of extreme behavior relative to the bulk of the data. We treat the bi-directional internet flows as bivariate functional data and compute scores using a multivariate FPCA (MFPCA) algorithm. The scores provide low dimensional representations of the traffic between the node IP addresses of each pair. Then, we propose a multivariate copula to compute the cybersecurity risk. The copula model is estimated after outlying scores have been removed because it is used to compute probabilities of extreme observations under the assumption of normal traffic. Even though we deal with a specific application, we propose a general paradigm that can be used

to develop effective screening tools to detect unusual multiple functional data objects.

It is informative to put the approach we propose into the context of previous research. Methods for detecting internet anomalies can be divided into signature-based methods and profile-based methods, Liao *et al.* (2013). Several requirements are necessary for signature-based methods to identify suspects, including the need for labeled data, prior results from anomalies, and an external supervisor. However, using this method, it is not possible to detect new intrusions that are unknown, Modi *et al.* (2013). A number of approaches have been proposed for the detection and prevention of DDoS attacks by using classification algorithms. The majority of such techniques require pre-training on a set of labeled data before they are applied. There are several popular approaches to data analysis, including Support Vector Machines, Bayesian Networks, and Neural Networks, Ahmed *et al.* (2016). Although these algorithms have performed well in certain situations in which "known" anomaly data exist, they can be difficult to incorporate into a larger set of algorithms due to the reliance on labeled data. It is likely that there will be no real knowledge for the classification of network traffic, which means supervised techniques can only be applied when approximated labels are available. It is inevitable that the results of training will be skewed by incorrectly labeled data, Soysal and Schmidt (2010).

Furthermore, an analysis of frequency domains has proven to be effective in detecting DDoS attacks, Fouladi *et al.* (2016). Compared to normal traffic in which energy is distributed among different frequencies, most DDoS attack energy is found at lower frequencies. Such methods have been used to discover abnormalities and analyze traffic patterns, Fouladi *et al.* (2013). Low rate DoS attacks (LDoS) are distinguished from normal traffic using spectrum energy and thresholding methods. Spectrum energy and thresholding are used to separate them, Wu *et al.* (2015). Spectral analysis is one of the methods used by the authors in order to detect DoS attacks, Hussain *et al.* (2003). It should be noted that most studies of frequency domain analysis in identifying DoS and DDoS attacks are carried out in simulation environments.

The remainder of this paper is structured as follows. Section 2 begins with an introduction of the MFPCA followed by algorithms for identification of outliers and copula based risk quantification. In Section 3, we apply our methods to a DDoS data set. The analysis is supplemented by a simulation study in Section 4.

## 2. Statistical methodology

In Section 2.1, we review the MFPCA and interpret it in context of source-destination traffic flows. Section 2.2 describes strategies used to remove outlying pairs so that a model for normal traffic (for the whole source-destination network) can be constructed. Finally, Section 2.3 explain the estimation of this model.

### 2.1. Multivariate functional principal components

To make the exposition more relevant, we introduce multivariate functional principal component analysis (MFPCA) in the context of time series of packet counts.

Suppose there are $N$ SIP-DIP (source-destination) pairs. Sources are outside, and destinations are inside an organization or a protected network. Let $(X_i(t), Y_i(t))$ be a bivariate

time series associated with the $i$th SIP-DIP pair. Here, $X_i(t)$ denotes the count of packets in hour $t$ in the SIP $\to$ DIP direction (i.e. inbound), and $Y_i(t)$ denotes the count of packets in hour $t$ in the DIP $\to$ SIP direction (i.e. outbound). These are pairs of noisy functions over the time interval $[0,T]$. We create their smooth versions and set

$$\mathbf{h}_i(t) = [h_i^{(1)}(t), h_i^{(2)}(t)]^\top. \tag{2.1}$$

The smoothing serves two purposes: 1) it is the first step in dimension reduction because it eliminates noise, 2) within an FDA software it converts discrete data to functional objects. The latter can be done in such a way that the functional objects look almost exactly as raw data, but then no noise reduction is achieved. We performed the smoothing using 100 B-spline basis functions. In the context of the data studied in Section 3, it corresponds to approximately using averages over 2.5 h, thus focusing only on persistent anomalies or attacks. Using 250 basis functions would practically correspond to working with raw data and would thus include anomalies lasting an hour or less, which we want to exclude, unless they are so large that their influence spreads over a few hours. Using 50 basis functions would focus on anomalies impacting at least five hours. The latter choice produces basically the same risk rankings as the 100 basis functions we use in the remainder of the paper. The details of smoothing are not essential to understand the remainder of the paper, we refer e.g. to Chapter 1 of Kokoszka and Reimherr (2017). Examples of the raw count data and smooth series are shown in Figure 1.

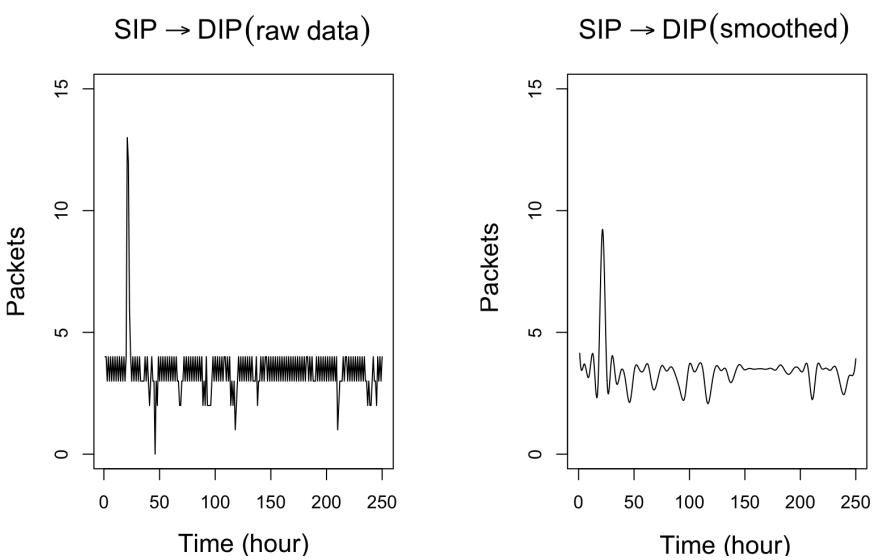


Figure 1: Example of DIP→SIP traffic and its smooth version.

We begin by describing the MFPCA algorithm of Happ and Greven (2018). We initially assume that each pair $i$ comes from the same population, in particular, the functions $h_1^{(k)}, \ldots, h_N^{(k)}$, have the same distributions as a population function $h^{(k)}$. This corresponds to

the absence of any outliers. We set

$$\boldsymbol{\mu}(t) = [\mu^{(1)}(t), \mu^{(2)}(t)]^\top = [E[h_i^{(1)}(t)], E[h_i^{(2)}(t)]]^\top, \qquad 1 \le i \le N, \tag{2.2}$$

and consider the Karhunen–Loève expansions

$$h_i^{(k)}(t) - \mu^{(k)}(t) = \sum_{m=1}^{\infty} \xi_{i,m}^{(k)} \phi_m^{(k)}(t) \approx \sum_{m=1}^{M} \xi_{i,m}^{(k)} \phi_m^{(k)}(t), \quad k = 1,2. \tag{2.3}$$

The functions $\phi_m^{(k)}$ are the functional principal components of the functions $h_i^{(k)}$. Their scores are $\xi_{i,m}^{(k)} = \langle h_i^{(k)} - \mu^{(k)}, \phi_m^{(k)} \rangle$. At this stage, decomposition (2.3) is performed for each $k$ separately. For each $k = 1,2$, the functions $\phi_m^{(k)}$ are orthonormal in the Hilbert space $L^2([0,T])$ and provide optimal data-driven basis systems in the sense that a specified accuracy of approximation that can be achieved with the smallest possible truncation level $M$. We refer e.g. to Chapter 11 of Kokoszka and Reimherr (2017) for an introductory account of FPCA and to Ramsay and Silverman (2005) and Horváth and Kokoszka (2012) for many examples of applications of FPCA.

Based on the sample $h_i^{(k)}$, $1 \le i \le N$, we can estimate the FPCs $\phi_m^{(k)}$ and the scores $\xi_{i,m}^{(k)}$. We denote the corresponding estimators by $\hat{\phi}_m^{(k)}$ and $\hat{\xi}_{i,m}^{(k)}$. We set

$$\Xi_i = (\hat{\xi}_{i,1}^{(1)}, ..., \hat{\xi}_{i,M}^{(1)}, \hat{\xi}_{i,1}^{(2)}, \ldots, \hat{\xi}_{i,M}^{(2)}) \tag{2.4}$$

and denote by $\boldsymbol{\Xi}$ the $N \times 2M$ matrix whose $i$th row is $\Xi_i$. Next, we set

$$\widehat{\mathbf{Z}} = (N-1)^{-1} \boldsymbol{\Xi}^\top \boldsymbol{\Xi} \quad (\dim[\widehat{\mathbf{Z}}] = 2M \times 2M). \tag{2.5}$$

The entries of the matrix $\widehat{\mathbf{Z}}$ are estimators of the covariances $E[\xi_m^{(k)} \xi_{m'}^{(l)}]$, $k,l = 1,2$, $m,m' = 1,\ldots M$.

The eigenvalues of the positive definite matrix $\widehat{\mathbf{Z}}$ are denoted by $\lambda_s$ and the orthonormal vectors belonging to them by $\hat{\mathbf{c}}_s$, i.e.

$$\widehat{\mathbf{Z}} \hat{\mathbf{c}}_s = \lambda_s \hat{\mathbf{c}}_s, \quad s = 1, \ldots, 2M, \tag{2.6}$$

with the convention that the eigenvalues $\lambda_s$ are ordered from the largest to the smallest. Each $\hat{\mathbf{c}}_s$ is a column vector of length $2M$. The multivariate eigenfunctions are estimated by $\hat{\psi}_m^{(k)}$ where

$$\hat{\psi}_m^{(k)}(t) = \sum_{j=1}^{M} \hat{c}_{(k-1)M+j,m} \hat{\phi}_j^{(k)}(t), \quad m = 1,2,\ldots,M, \;\; k = 1,2. \tag{2.7}$$

The multivariate scores are calculated as

$$\hat{\rho}_{i,m} = \sum_{k=1}^{2} \sum_{j=1}^{M} \hat{c}_{(k-1)M+j,m} \hat{\xi}_{i,j}^{(k)}, \quad m = 1,2,\ldots,M, \;\; i = 1,...,n. \tag{2.8}$$

There is a correlation between the two sets of scores since the number of packets sent from SIP to DIP is correlated with the number of packets sent from DIP to SIP. The MFPCA algorithm has the advantage of revealing a joint variation in the number of packets sent in both directions that cannot be captured by separate FPCA.

We emphasize that the $\phi_m^{(k)}$ and $\xi_{i,m}^{(k)}$ are the functional principal components and scores from univariate FPCA, while the $\hat{\psi}_m^{(k)}$ are the multivariate functional principal components of the $k$th variable and $\hat{\rho}_{i,m}$ are the corresponding scores of the $i$th multivariate functional observation. Thus, in the MFPCA, the functional principal components of both variables share the same score. These scores reflect the variability of pairs rather than their individual components. While the objects at the population level are defined under the assumption of identical distributions, the estimators discussed above can be computed for any sample of SIP-DIP pairs.

We conclude this section by introducing the concept of the copula, see Genest and Nešlehová (2012) for a recent review. Consider a random vector $(Z_1,\ldots,Z_d)$ with univariate continuous marginal distribution $F_1,\ldots,F_d$, respectively. Then the random vector $(U_1,\ldots,U_d) = (F_1(Z_1),\ldots,F_d(Z_d))$, where $F_k(z) = P(Z_k \leq z)$ has marginals that are uniformly distributed on the interval $[0,1]$. The copula of $(Z_1,\ldots,Z_d)$ is defined as the joint cumulative distribution functions of $(U_1,\ldots,U_d)$, i.e.

$$C(u_1,\ldots,u_d) = P\left(Z_1 \leq F_1^{-1}(u_1),\ldots,Z_d \leq F_d^{-1}(u_d)\right). \tag{2.9}$$

Equivalently, for any random vector $(Z_1,\ldots,Z_d)$ with distribution function $F(z_1,\ldots,z_d)$ and marginal distributions $F_1,\ldots,F_d$, there is a copula $C$ such that

$$F(z_1,\ldots,z_d) = C(F_1(z_1),\ldots,F_d(z_d)).$$

Therefore, assuming that the margins $F_1,\ldots,F_d$ are continuous and that the unique underlying copula is absolutely continuous, the joint density function can be represented as

$$f(z_1,...,z_d) = c(F_1(z_1),\ldots,F_d(z_d))\prod_{i=1}^{d} f_i(z_i),$$

where $f_i(z_i)$ is the corresponding marginal density function of $Z_i$ and $c(u_1,...,u_d)$ is the $d-$dimensional copula density function. We refer to $C$ or $c$ as a copula model.

In Sections 2.2 and 2.3, we use the letter $d$ in place of $M$. Our recommendation is to perform the MFPCA for some larger $M$, and then depending on the variance explained, use $d < M$ initial components.

### 2.2. Identification of risky source-destination pairs

We will use bivariate FPCA and a probabilistic copula-based method as our anomaly detection and risk quantification techniques. There are three stages. First, we consider the bi-directional streams $[(h_i^{(1)}(t), h_i^{(2)}(t))]$, $i = 1,...,N$, as bivariate functional data and

compute the scores $\hat{\rho}_{i,m}$ defined by (2.8). Then, a copula model is estimated based on the score vectors $\boldsymbol{\rho}_i = (\rho_{i1}, ..., \rho_{id})$, $i = 1, ..., N$, obtained from the bivariate FPCA after outlying scores or outlying functions have been removed. Finally, a copula model is used to compute the risk of each SIP-DIP pair. We propose two strategies to remove outliers. In the first algorithm, we remove extremely large scores before fitting a copula. In the second algorithm, we remove pairs of functions associated with extreme scores, recompute the scores, and then fit a copula. The justification for removing outlying pairs of curves is to ensure that a copula is estimated on data that can be reasonably assumed to come from the same distribution, so a single copula model is appropriate. Outliers come from different distributions than the bulk of the data. These two strategies are summarized in Algorithms 1 and 2 below. In Algorithm 3, we explain how extremely large scores are identified.

ALGORITHM 1

1. For the smooth versions $(h_i^{(1)}(t), h_i^{(2)}(t))$, $i = 1, ..., N$, estimate the multivariate functional principal components $\psi_m^{(k)}$, $k = 1, 2$, and the scores $\hat{\rho}_{i,m}$, $m = 1, ..., d$.
2. If pair $i$ has extremely large $\hat{\boldsymbol{\rho}}_i$, then it is considered as an outlier. Remove $\hat{\boldsymbol{\rho}}_i$ from the estimated scores.
3. Estimate a copula model based on the remaining scores $\hat{\boldsymbol{\rho}}_i = (\hat{\rho}_{i1}, ..., \hat{\rho}_{id})$.

ALGORITHM 2

1. Step 1 is the same as in Algorithm 1.
2. If pair $i$ has extremely large $\hat{\boldsymbol{\rho}}_i$, remove $(h_i^{(1)}(t), h_i^{(2)}(t))$.
3. Estimate the multivariate functional principal components $\psi_m^{(k)}$ and the scores $\hat{\rho}_{i,m}$ again.
4. Iterate Step 2 and Step 3 until there is no more $\hat{\boldsymbol{\rho}}_i$ identified as outlying.
5. Estimate a copula model based on estimated scores $\hat{\boldsymbol{\rho}}_i = (\hat{\rho}_{i1}, ..., \hat{\rho}_{id})$.

Step 2 of both algorithms identifies outlying pairs $i$ using the following Algorithm 3 due to Billor *et al.* (2000). We note that any effective way of identifying pairs of outlying curves could be used; the approaches of Hubert *et al.* (2005), Dai and Genton (2018) or Amovin-Assagba *et al.* (2022) could be effective. As we will see in Section 3, a more significant difference arises depending on whether Algorithms 1 or 2 are used.

ALGORITHM 3

1. Compute the Mahalanobis distance for each $\hat{\boldsymbol{\rho}}_i$:

$$\text{Mahalanobis distance} = (\hat{\boldsymbol{\rho}}_i - \bar{\hat{\boldsymbol{\rho}}}_i)^\top \mathbf{S}^{-1} (\hat{\boldsymbol{\rho}}_i - \bar{\hat{\boldsymbol{\rho}}}_i),\ i = 1, \ldots, N,$$

   where $\bar{\hat{\boldsymbol{\rho}}}_i$ and $\mathbf{S}$ are the mean and the sample covariance matrix of the $\hat{\boldsymbol{\rho}}_1, \ldots, \hat{\boldsymbol{\rho}}_N$. Select a potential basic subset of size $k$ ($k > M$) of smallest Mahalanobis distances that can safely be assumed free of outliers.

2. Compute the discrepancies:

$$d_i = \sqrt{(\hat{\boldsymbol{\rho}}_i - \bar{\boldsymbol{\rho}}_b)^\top \mathbf{S}_b^{-1} (\hat{\boldsymbol{\rho}}_i - \bar{\boldsymbol{\rho}}_b)},\ i = 1, \ldots, N,$$

where $\bar{\boldsymbol{\rho}}_b$ and $\mathbf{S}_b$ are the sample mean and the sample covariance matrix of the observations in the basic subset.

3. Denote by $\chi^2_{d,\alpha/N}$ the $(1-\alpha/N)$th quantile of the chi-square distribution with $d$ degrees of freedom. The level $\alpha$ depends on how many risky pairs we want to identify; see the discussion following (2.13).

Set the new basic subset to all points with discrepancies less than $c$, where

$$c = \sqrt{\chi^2_{d,\alpha/N}} \left( \max \left\{ 0, \frac{h-k}{h+k} \right\} + 1 + \frac{d+1}{N-d} + \frac{1}{N-h-d} \right)$$

with $h = (N+d+1)/2$.

4. The stopping rule: Iterate Step 2 and 3 until the size of the basic subset no longer changes.

5. Nominate the observations excluded by the final basic subset as outliers.

### 2.3. Risk quantification using a copula model

Among several copula candidates, we settled on the $t$-copula that is widely used in finance and risk analysis, see Demarta and McNeil (2005). We also considered the popular normal copula, but it did not lead to a good separation of risks for the most extreme pairs. The R package `copula` contains many other copula models that could be used in various settings, and could be better than the $t$-copula in different applications.

The $d$-dimensional $t$-copula with $\nu$ degrees of freedom and association matrix $\Sigma$ is the probability distribution on $[0,1]^d$ whose distribution function is given by

$$C_{\nu,\Sigma}(\boldsymbol{u}) = \int_{-\infty}^{t_\nu^{-1}(u_1)} \ldots \int_{-\infty}^{t_\nu^{-1}(u_d)} \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^d |\Sigma|}} \left( 1 + \frac{\boldsymbol{x}' \Sigma^{-1} \boldsymbol{x}}{\nu} \right)^{-\frac{\nu+d}{2}} d\boldsymbol{x} \tag{2.10}$$

where $t_\nu(\cdot)$ is the distribution function of a univariate $t$-distribution with $\nu$ degrees of freedom. The probability density function corresponding to (2.10) equals to

$$c_{\nu,\Sigma}(\boldsymbol{u}) = \frac{dt_{\nu,\Sigma}(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d))}{\prod_{i=1}^d dt(t_\nu^{-1}(u_i), \nu)}, \tag{2.11}$$

where $dt_{\nu,\Sigma}(\cdot)$ and $dt(\cdot,\cdot)$ are the densities of multivariate and univariate $t$-distribution, respectively. We used the R package **copula** to fit copula (2.10). While the $t$-copula provides a useful separation of risks for the data we study in Section 3, different copulas could be more appropriate for different data sets. Our criterion is that the highest risks should be clearly separated from each other and the bulk of the data.

Risk usually refers to the uncertainty of an outcome given a situation. Cybersecurity risk is the potential for a cybersecurity threat to occur. Following an established practice, we use tail probabilities to quantify risk. To explain the idea, we consider the first two scores, i.e., $\boldsymbol{\rho}_i = (\rho_{i1}, \rho_{i2})$. This corresponds to $d = 2$ used in Section 3. In general, the four cases in (2.12) would be replaced by $2^d$ cases. Define the probability of scores more extreme than those of the observed pair $(\hat{\rho}_{i1}, \hat{\rho}_{i2})$ as

$$p_i = \begin{cases} P(\rho_{i1} \geq \hat{\rho}_1, \rho_{i2} \geq \hat{\rho}_2), & \text{if } \hat{\rho}_{i1} \geq 0 \text{ and } \hat{\rho}_{i2} \geq 0 \\ P(\rho_{i1} \leq \hat{\rho}_1, \rho_{i2} \geq \hat{\rho}_2), & \text{if } \hat{\rho}_{i1} < 0 \text{ and } \hat{\rho}_{i2} \geq 0 \\ P(\rho_{i1} \leq \hat{\rho}_1, \rho_{i2} \leq \hat{\rho}_2), & \text{if } \hat{\rho}_{i1} < 0 \text{ and } \hat{\rho}_{i2} < 0 \\ P(\rho_{i1} \geq \hat{\rho}_1, \rho_{i2} \leq \hat{\rho}_2), & \text{if } \hat{\rho}_{i1} \geq 0 \text{ and } \hat{\rho}_{i2} < 0. \end{cases} \tag{2.12}$$

The extreme (risky) regions may have a different form, and will look differently in higher dimensions, but (2.12) is a commonly used definition on the plane. We require that in every quadrant, both scores are extreme, rather than just one of them. If the $i$th pair of traffic flows is anomalous, then it should occur infrequently, i.e., the probability of obtaining $\rho_i$ at least as extreme should be small. To associate high risk with large positive values, we work with negative log probabilities. Thus, the cybersecurity risk of pair $i$ is defined as

$$R_i = -\log(\varepsilon + p_i), \tag{2.13}$$

where $\varepsilon > 0$ is a small value, the same in all calculations. (In Section 3, we use $\varepsilon = 0.001$.) The risks $R_i$ can be used to rank the pairs from most risky to least risky. One can also set a probability threshold $\alpha$, and consider the pairs satisfying $R_i > -\log(\varepsilon + \alpha)$ as exceptionally risky. We emphasize that $\alpha$ has an interpretation as a probability only within the copula model. Alternatively, one can report $\alpha$ corresponding to 10 or 20, or any other number of most risky pairs. In most applications, we are dealing with thousands of pairs.

## 3. Application to bi-directional packet flows

### 3.1. Data description and preliminary analysis

The data set we study consists of a collection of time series of bi-directional packet flows, aggregated hourly, between source Internet protocol (SIP) addresses and destination IP (DIP) addresses captured at a large university from October 20th to 30th, 2013. These data are collected 3 months before a major DDoS attack occurred around January 10th, 2014. The data, transformed with Crypto-PAn, as well as the source code, accompany this paper at the journal's website. During the 250-hour time window over which the data were collected, there are 869 unique SIPs connected with 1869 unique DIPs, and a total of approximately 1.2 million data packets were sent. We consider $N = 3049$ unique SIP-DIP pairs, where SIP is an IP outside the university network and DIP is inside. Each pair is associated with two observed time series, an inbound packet flow and an outbound packet flow. The pairs are labeled with integers $1, 2, \ldots, 3049$, the SIPs with $S1, S2, \ldots, S869$ and the DIPs with $D1, D2, \ldots, D1869$. This is needed to anonymize the IP addresses and ease the notation, the real addresses are long string of integers.

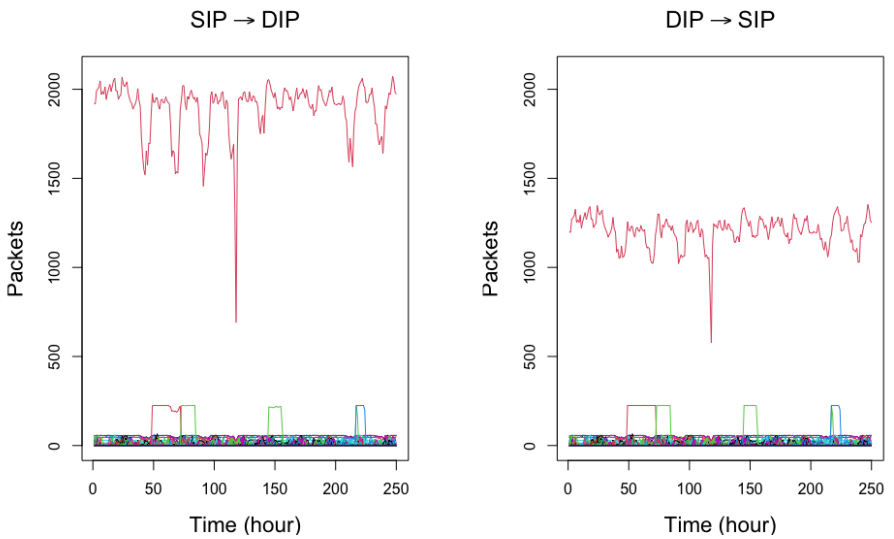


Figure 2: Time series plots of traffic traces. Left: inbound (SIP to DIP); Right: outbound (DIP to SIP). Each time series depicts the hourly count of packets between a SIP-DIP pair.

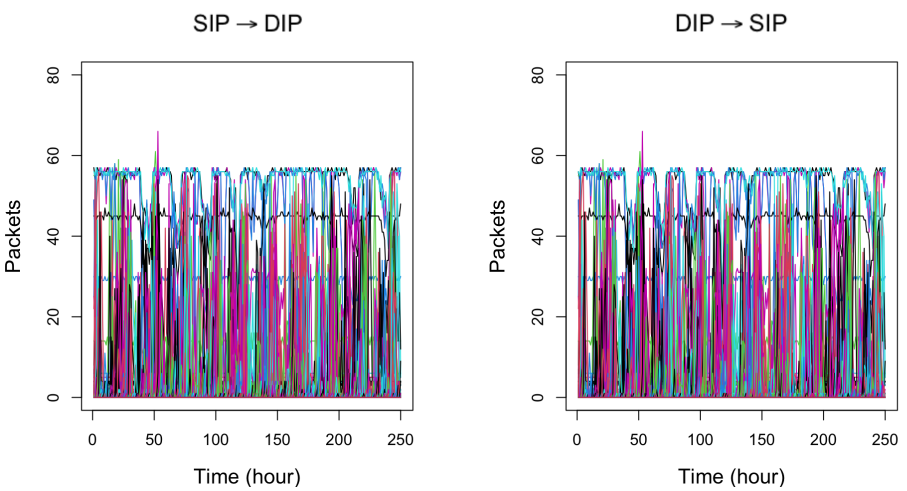


Figure 3: Zoom of Figure 2 with outlying traces removed.

Time series of inbound traffic traces (from SIP to DIP) and outbound traffic traces (from DIP to SIP) are depicted, respectively, in the left and right panels of Figure 2. The hourly count of packets is shown as $y$-axis, and the time (in hours) is shown as $x$-axis. It can be seen that there are some clearly, or potentially, outlying packet flows. These are the traces that need to be removed before the computation of the bivariate FPCA is performed. Detection and ordering of risky pairs in the remaining data set shown in Figure 3, cannot be done visually, or by an obvious algorithm. This is why we have developed copula-based algorithms.

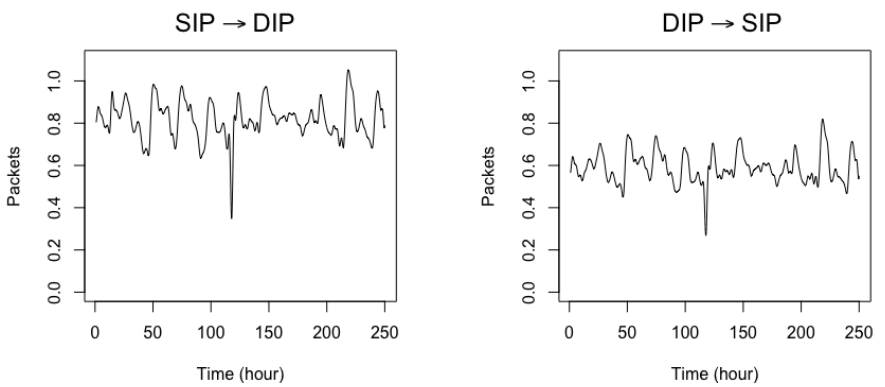


Figure 4: Mean functions of all functions in the sample.

Another justification of the need to develop an algorithm that uses only the pairs that are not obviously outlying comes from the examination of Figures 4 and 5. Figure 4 shows sample mean functions computed from all available data. It is seen that they strongly reflect the extremely outlying curves in Figure 2, one curve in each panel. Similarly, the initial FPCs, shown in Figure 5 reflect the deviations of the mean due to smaller outliers, except the first FPC that reflects the differences in level for most functions. These figures show that the FPCA based on all functions is not suitable for the quantification of risk because it reflects the most risky functions and mostly ignores the bulk of the data. For these reasons, in the following, we first apply the outlier removal algorithms proposed in Section 2.2.

We conclude this section with information about running times. On a 2.2 GHz Intel Core i7 processor, 16GB RAM, the average running times over three repetitions were 27.9 s for Algorithm 1 and 129.9 s for Algorithm 2, for the data set described at the beginning of this section.

### 3.2. Risk analysis using Algorithm 1

For reasons explained in Section 3.1, before fitting a copula model, we use Algorithm 3 in Section 2.2 to remove outlying curves. It identifies four pairs with abnormal scores (labeled 2, 794, 1077, and 1491). These pairs are excluded from the copula model estimation. Using only the remaining pairs, 95.6% of the variance is explained by the first two MFPCs, with 86.3% of the variance explained by the first MFPC and 9.3% by the second MFPC.

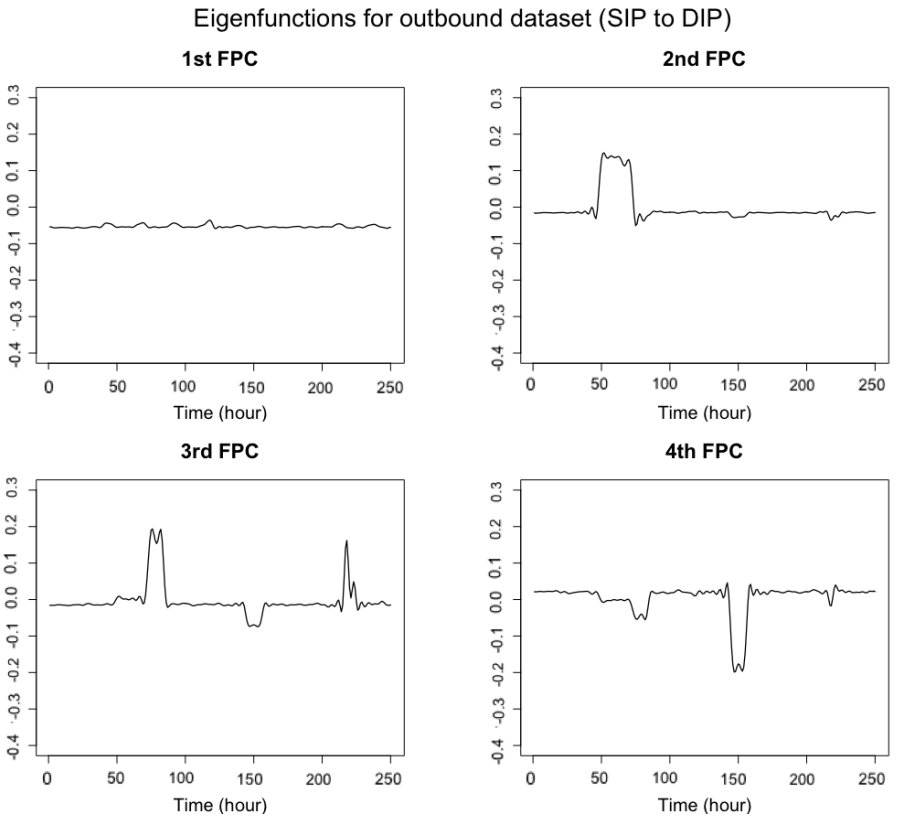


Figure 5: The first four sample FPCs for all functions in the SIP to DIP direction.

Using $d = 2$ is therefore sufficient to capture the main features of the data. After estimating the bivariate $t$ copula (2.10), we compute the probabilities $\hat{p}_i$ using (2.11) for *all* pairs $\hat{\boldsymbol{\rho}}_i = (\hat{\rho}_{i1}, \hat{\rho}_{i2})$, including those that were excluded in the copula estimation. Next, we compute the risks using equation (2.13) with $\varepsilon = 0.001$. The risks are in the range $[0.624, 2.336]$, i.e. $\hat{R}_i \in [0.624, 2.336]$. To give a better idea about the range of risk, we consider, say, 55 pairs with the highest risk. They have risks higher than 1.509. This corresponds to the cut-off level $\alpha = 0.22$, i.e for these 55 pairs, $\hat{R}_i > -\log(\varepsilon + 0.22) = 1.509$. Table 3.2 shows the risks for ten riskiest pairs.

**Table 1.** The 10 riskiest pairs according to Algorithm 1

| Pair | SIP | DIP | Risk |
|---|---|---|---|
| 2 | S2 | D2 | 2.336 |
| 1077 | S312 | D655 | 2.0679 |
| 1491 | S213 | D899 | 2.0404 |
| 2260 | S312 | D1296 | 1.999 |
| 10 | S10 | D1 | 1.896 |
| 51 | S46 | D1 | 1.870 |
| 40 | S36 | D1 | 1.861 |
| 33 | S30 | D1 | 1.858 |
| 1272 | S423 | D1 | 1.854 |
| 34 | S31 | D1 | 1.820 |

We note that the risky pairs are found, and their risks computed, using presmoothing with $B = 100$ splines, which is the value used for all analyses presented in this paper. This level of smoothing is suitable to capture the main and relevant features of the data we study. We refer again to Figure 1. We need a level of smoothing that preserves large spikes, but basically ignores typical variability that is not unusual in any way. Larger values of $B$ are not recommended because they would basically reproduce the raw data and distort the MFPCA that requires smooth functions as inputs. Using $B = 50$ produces basically the same risks and identifies almost the same sets of risky pairs. Using fewer that $B = 50$ basis functions is not recommended because the spikes are smoothed out too much.

We examined the patterns of high risk pairs in Table 3.2. The high-risk pairs can, roughly, be classified into three groups, which we denote (a), (b), and (c). Figure 6 shows examples of packet flows in each of the three groups. The curves in group (a) have high levels of packet flows with many rapid drops in the packet counts. Pair 2, the most outstanding outlier, is characterized by exceptionally large traffic. Pair 34 has a similar pattern as pair 2, but the traffic levels are much lower, so it is not displayed in Figure 6. The relatively high levels of activity in group (b) (pairs 1077, 1491, and 2260) last only for a short period of time, and at other times, no activity occurs. The curves in group (c) (pairs 10, 33, 40, 51, and 1272) have generally low levels with many spikes. Only two pairs in groups (b) and (c) are plotted, so as not to obscure the picture.

We also examined other pairs in the group of the 55 riskiest pairs, beyond those in Table 3.2. The general patterns are somewhat different. Basically, the patterns in panels (a) and (b) of Figure 3.2 are exceptional and correspond to outliers. For the majority of high-risk pairs three different groups can be identified. Figure 7 shows examples of packet flows in each of the three groups. The curves in group (a) have moderate levels of packet flows, but exhibit more variability than typical curves. The curves in group (b) have mostly high levels of packet flows with many rapid drops in the packet counts. Group (c) coincides with group (c) in Figure 6. It is basically a mirror image of group (b). The curves in that group have generally low levels with many upward spikes.

It is not possible to display risks for all 3049 pairs in our data set. To obtain some additional insights, we proceed as follows. In the 3049 pairs, there are SIPs that appear more often than others. We thus ranked the SIPs by the frequency with which they appear in the pairs. For example, the address S23 appears most frequently, in 241 out of 3049

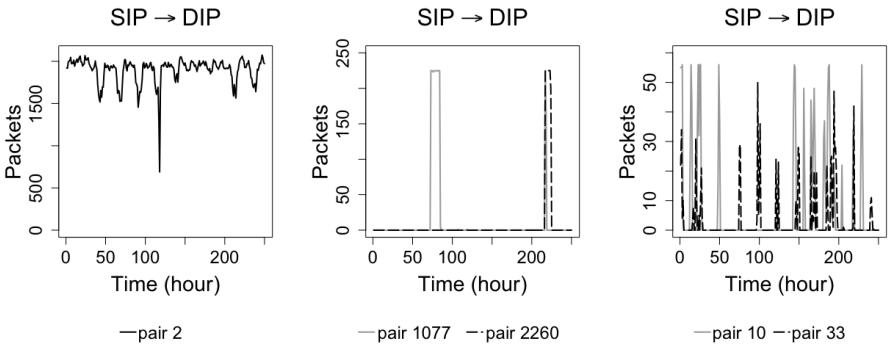


Figure 6: Examples of traffic traces corresponding to the pairs in Table 3.2.

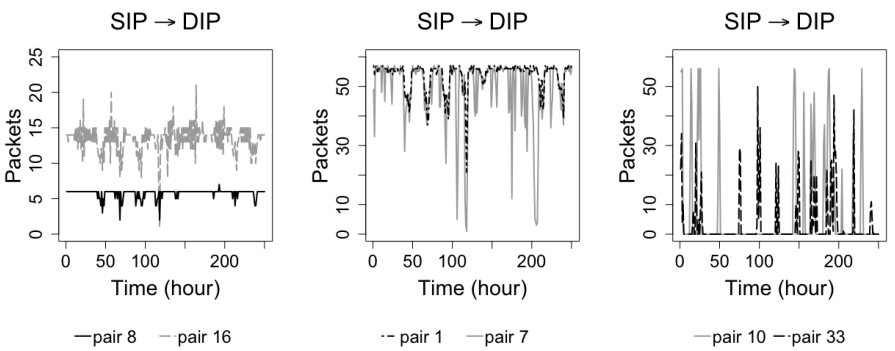


Figure 7: Examples of traffic traces corresponding to pairs identified as high risk under Algorithm 1: Left: pair 8 (S8 → D2), pair 16 (S15 → D2); Middle: pair 1 (S2 → D2), pair 7 (S7 → D2); Right pair 10 (S10 → D1), pair 33 (S30 → D1).

pairs. We performed the same ranking for the DIPs; the address D1 appears most often, in 285 out of 3049 pairs Figure 8 shows risks for the 10 most frequent DIP and SIP addresses. According to Figure 8, the pairs including the 10 most frequent SIPs tend to be less risky, whereas the pairs sent to the 10 most frequent DIPs require more attention, particularly D1 and D2. The DIP D2 was captured by the outlier detection Algorithm 3, but D1 was identified only after computing the risks. A finding of this type may indicate that D1 and D2 (that are within the university) may require special attention.

The results presented in this section illustrate the value of quantitative risk assessment. Certain SIP-DIP pairs are brought to attention by their high risk, even though they are difficult to identify visually due to the fact that we are dealing with thousands of pairs of curves with very complex shapes; in a cloud of thousands of curves it is difficult to see which are more unusual than others, and it is difficult to examine them visually one after another. Our method provides a tool for sorting the SIP-DIP pairs so that attention can be focused only on the riskiest ones.

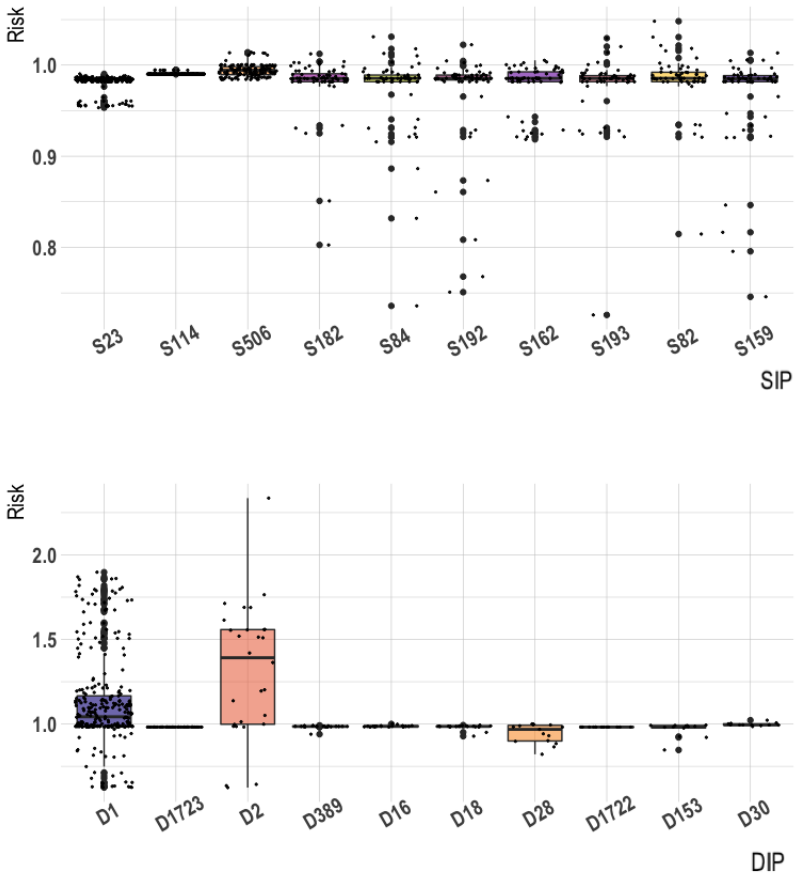


Figure 8: Top: Boxplots of risks for pairs with the 10 most frequent SIPs. Bottom: Boxplots of risks for pairs with the 10 most frequent DIPs. (Algorithm 1)

### 3.3. Risk analysis using Algorithm 2

We now turn to the application of Algorithm 2 proposed in Section 2.2. The results of copula estimations are shown in Table 3.3. Due to the iterative procedure for the removal of outliers, Algorithm 2 is expected to identify more outliers than Algorithm 1. We emphasize that risks are computed for all pairs, including those identified as outliers. The difference relative to Algorithm 1 is that the copula is estimated on a smaller subset of "typical" pairs. For the data set we study, Algorithm 2 identified 54 pairs as outlying using $\alpha = 0.1$ in Step 3 of Algorithm 3. The first iteration identified, by definition, the same outliers as Algorithm 1: pairs 2, 794, 1077 and 1491. The second iteration identified 42 new outliers, third, 5 outliers and fourth 3. After the fourth iteration no more outliers were identified. The range of $\hat{R}_i$ computed by Algorithm 2 is $[0.870, 2.059]$. The risks are different than those obtained

from Algorithm 1, where the range was $[0.624, 2.336]$. We emphasize that the values of risks are used only for identifying and ranking risky pairs, they do not have an "absolute" interpretation. This point is further highlighted by comparing Figures 8 and 9. Table 3.3 displays the risks of the riskiest pairs identified by Algorithm 2. The pairs in Table 3.3 are different than those in Table 3.2, with some overlap (pairs 2260, 2, 1491, 1272, 40). This is to be expected because different copula models are used to compute them. A change in ranking can also occur if the method is applied to transformed data. We applied Algorithm 2 to $\log(1+\text{count})$ and obtained slightly different, but similar rankings using the level of smoothing similar to that used for the original data. This is understandable, because after any transformation, the curves take on different shapes.

**Table 2.** Results of the estimation of the $t$ copula based on the two algorithms

| | Algorithm 1 | Algorithm 2 |
|---|---|---|
| Degrees of freedom | 1.844 | 3.359 |
| Correlation matrix | $\begin{pmatrix} 1 & 0.344 \\ 0.344 & 1 \end{pmatrix}$ | $\begin{pmatrix} 1 & -0.0737 \\ -0.0737 & 1 \end{pmatrix}$ |
| Margin 1 | $t_{0.785}(\mu = -0.672, \sigma = 0.0289)$ | $t_{1.0769}(\mu = -0.0459, \sigma = 0.0273)$ |
| Margin 2 | $t_{0.965}(\mu = 0.0761, \sigma = 0.0295)$ | $t_{0.908}(\mu = 0.00433, \sigma = 0.0470)$ |

**Table 3.** The 10 riskiest pairs according to Algorithm 2

| Pair | SIP | DIP | Risk |
|---|---|---|---|
| 2260 | S312 | D1296 | 2.0594 |
| 794 | S312 | D13 | 2.0329 |
| 80 | S71 | D1 | 2.0294 |
| 2 | S2 | D2 | 1.995 |
| 1491 | S213 | D899 | 1.994 |
| 79 | S70 | D1 | 1.988 |
| 43 | S39 | D2 | 1.951 |
| 1272 | S423 | D1 | 1.945 |
| 57 | S49 | D1 | 1.938 |
| 40 | S36 | D1 | 1.929 |

## 4. Assessment of the methodology on simulated data

A question arises whether Algorithm 1 or Algorithm 2 provides a more useful risk ranking. To address this question, we need an informative simulation study, which is the focus of this section.

The chief difference between Algorithms 1 and 2 of Section 2.2 is as follows. In Algorithm 1, the MFPCs are computed using all available data, even the potential outliers. The largest outliers do not affect the MFPCs because they impact the mean functions that are subtracted before the computation of the MFPCs. In Algorithm 2, the MFPCs are computed after the outliers have been removed. For example, in Section 3.3 they were computed after 54 pairs had been removed. We assess the performance, and relative performance, of the two algorithms using simulated data that has certain features of our real data sets, but also certain characteristics that are known targets. In step 1 of the following data generation algorithm, we have two options, A and B. Option A might seem to, a priori, favor Algorithm

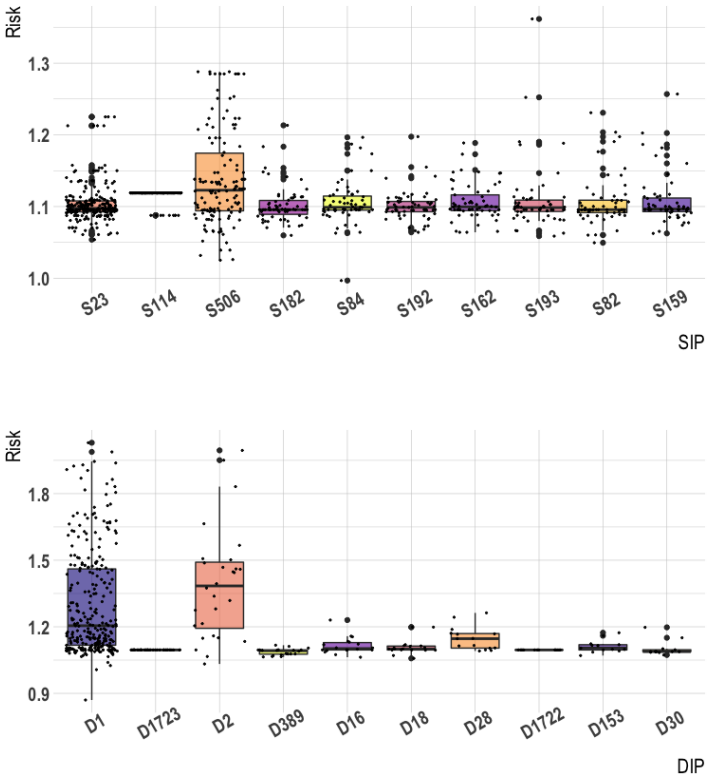


Figure 9: Top: Boxplot of risks for pairs with the 10 most frequent SIPs. Bottom: Boxplot of risks for pairs with the 10 most frequent DIPs. (Algorithm 2)

1 and Option B Algorithm 2.

1. A Estimate $\psi_1^{(1)}, \psi_2^{(1)}, \psi_3^{(1)}$ and $\psi_1^{(2)}, \psi_2^{(2)}, \psi_3^{(2)}$ using all data.

   B Remove 54 pairs identified by Algorithm 2 as outlying and estimate $\psi_1^{(1)}, \psi_2^{(1)}$, $\psi_3^{(1)}$ and $\psi_1^{(2)}, \psi_2^{(2)}, \psi_3^{(2)}$ based on the remaining 3049 - 54 = 2995 pairs.

2. For $1 \leq i \leq 2995$, generate

$$X_i = \sum_{j=1}^{3} \xi_{ij} \psi_j^{(1)}, \quad Y_i = \sum_{j=1}^{3} \eta_{ij} \psi_j^{(2)} \tag{4.14}$$

with iid scores $\xi_{ij}$ and $\eta_{ij}$ distributed according to

$$\xi_1 \sim t_{10}, \;\; \xi_2 \sim 0.5\, N(0,1), \quad \xi_3 \sim 0.1\, N(0,1),$$

$$\eta_1 \sim t_{11}, \;\; \xi_2 \sim 0.4\, N(0,1), \quad \xi_3 \sim 0.2\, N(0,1),$$

The first 2995 pairs are the typical low risk pairs.

3. For $2996 \le i \le 3041$, generate the pairs $(X_i, Y_i)$ according to (4.14), but $\xi_1$ and $\eta_1$ having different, "larger" distributions, as specified below. The remaining distributions are unchanged. These are the pairs with increasing risks. Pair 2996 has the smallest risk of them, pair 3041 the highest.

4. For $3042 \le i \le 3049$, generate the pairs $(X_i, Y_i)$ according to (4.14), but with $\xi_1$ and $\eta_1$ having "extremely" large distributions. These are the outlying pairs

In steps 3 and 4 above, the distribution of the scores changes, so as reduce the dependence of the the conclusions on a specific distribution of risky and outlying pairs. We repeat steps 1-4 20 times, and use four distributions for each batch of five simulations according to the following specifications:

Simulations 1 to 5: For $2996 \le i \le 3041$, $\xi_1 \sim (i-2995)t_{10}, \eta_1 \sim (i-2995)t_{11}$; for $3042 \le i \le 3049$, $\xi_1 \sim 2(i-3041)t_3, \eta_1 \sim 2(i-3041)t_4$.

Simulations 6 to 10: For $2996 \le i \le 3041$, $\xi_1 \sim (i-2995)\text{Exp}(0.5), \eta_1 \sim (i-2995)\text{Exp}(1)$; for $3042 \le i \le 3049$, $\xi_1 \sim (i-3041)\text{Exp}(0.1), \eta_1 \sim (i-3041)\text{Exp}(0.5)$;

Simulations 11 to 15: For $2996 \le i \le 3041$, $\xi_1 \sim \frac{2i-2995}{10}\text{Exp}(1), \eta_1 \sim \frac{2i-2995}{11}\text{Exp}(2)$; for $3042 \le i \le 3049$, $\xi_1 \sim \frac{2i-3041}{5}\text{Exp}(1), \eta_1 \sim \frac{2i-3041}{6}\text{Exp}(2)$.

Simulations 16 to 20: For $2996 \le i \le 3041$, $\xi_1 \sim (i-2995)\text{Exp}(0.5), \eta_1 \sim (i-2995)t_{11}$; for $3042 \le i \le 3049$, $\xi_1 \sim (i-3041)\text{Exp}(0.1), \eta_1 \sim (i-3041)t_4$.

We apply Algorithms 1 and 2 to the data generated above. Note that each algorithm estimates the MFPCs and the scores. The estimated MFPCs will be different than those used to generated the data in Step 1. We list the pairs identified as outliers. The target list are pairs $3042, 3043, \ldots, 3049$. We find 54 riskiest pairs and order them from the one with the smallest risk to the one with the highest risk (according to each algorithm). We denote the indexes as $i_1, \ldots, i_{54}$. These indexes will be different for the two algorithms. The pair $(X_{i_1}, Y_{i_1})$ has the the lowest risk out of the 54 pairs. We compute the absolute differences $|i_k - k - 2995|$, $k = 1, \ldots, 54$, and plot them as histograms for both algorithms. If an algorithm performs well, these differences should be small. For an algorithm that detects outliers perfectly and ranks the risks perfectly, they should all be zero. However, due to the random generation of outlying and risky pairs, some of them will not appear to be in these categories because even a $t_3$ distribution can take a value close to zero. However, our experiment should give a reasonable idea how the algorithms perform, as we now report.

In both scenarios A and B, Algorithm 1 identifies five to nine pairs as outlying and Algorithm 2 eight to seventeen pairs. In this sense, Algorithm 1 is closer to our target of seven outlying pairs. However, as shown in Figure 10, Algorithm 2 has an advantage in ranking the risky and outlying pairs, but is more prone to make serious mistakes more often that Algorithm 1. The reader can certainly draw conclusions from the above analysis, but it appears that the additional outliers identification step in Algorithm 2 does not provide a decisive improvement. One might conclude that both algorithms identify outliers and risky pairs in a satisfactory manner, but may result in somewhat different risk rankings, as we have seen in Section 3.

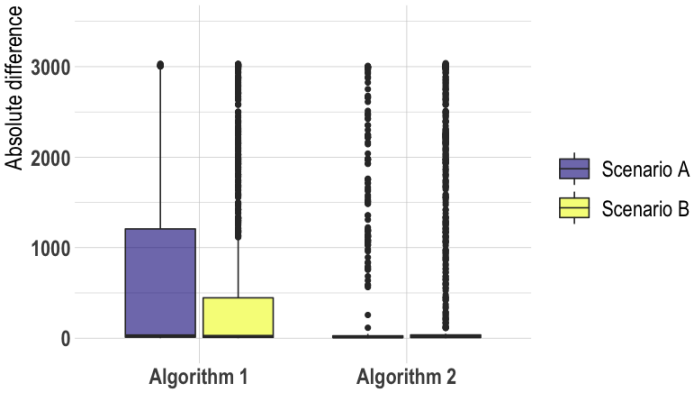


Figure 10: Boxplots of absolute difference for the top 54 risky pairs for both algorithms in two scenarios.

## Acknowledgements

This research was partially supported by the United States National Science Foundation grant DMS–2123761.

## References

Ahmed, Mohiuddin, Mahmood, Abdun Naser and Hu, Jiankun, (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60**,** pp. 19–31.

Amovin-Assagba, Martial, Gannaz, Irène and Jacques, Julien, (2022). Outlier detection in multivariate functional data through a contaminated mixture model. *Computational Statistics & Data Analysis*, 174**,** 107496.

Awan, Mazhar Javed, Farooq, Umar, Babar, Hafiz Muhammad Aqeel, Yasin, Awais, Nobanee, Haitham, Hussain, Muzammil, Hakeem, Owais and Zain, Azlan Mohd, (2021). Real-time DDoS attack detection system using big data approach. *Sustainability*, 13**,** no. 19, 10743.

Berrendero, José R, Justel, Ana and Svarc, Marcela, (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55**,** no. 9, pp. 2619–2634.

Billor, Nedret, Hada, Ali and Velleman, Paul, (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis*, 34**,** pp. 272–298.

Bosq, Dennis, (2000). *Linear Processes in Function Spaces*. Springer.

Chiou, Jeng-Min, Chen, Yu-Ting and Yang, Ya-Fang, (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, pp. 1571–1596.

Czado, Claudia, (2019). *Analyzing Dependent Data with Vine Copulas: A Practical Guide with R*. Springer.

Dai, Wenlin and Genton, Marc G., (2018). Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, 27**,** no. 4, pp. 923–934.

Demarta, Stefano and McNeil, Alexander, (2005). The *t* copula and related copulas. *International Statistical Review*, 73**,** pp. 111–129.

Dong, Shi and Sarem, Mudar, (2019). DDoS attack detection method based on improved KNN with the degree of DDoS attack in software-defined networks. *IEEE Access*, 8**,** pp. 5039–5048.

Ferraty, Frédérick and Vieu, Philippe, (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer.

Fouladi, Ramin Fadaei, Kayatas, Cemil Eren and Anarim, Emin, (2016). Frequency based DDoS attack detection approach using naive Bayes classification. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 104–107. IEEE.

Fouladi, Ramin Fadaei, Seifpoor, Tina and Anarim, Emin, (2013). Frequency characteristics of DoS and DDoS attacks. In *2013 21st Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4. IEEE.

Genest, Christian and Nešlehová, Johanna, (2012). Copulas and copula models. In *Encyclopedia of Environmetrics* (eds El-Shaarawi A.H. and Piegorsch W.W.), 2 edn, volume 2, pp. 541–553. Wiley, Chichester.

Górecki, Tomasz, Krzyśko, Mirosław, Waszak, Łukasz and Wołyński, Waldemar, (2018). Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers*, 59**,** no. 1, pp. 153–182.

Happ, Clara and Greven, Sonja, (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113**,** number 522, pp. 649–659.

Hofert, Marius, Kojadinovic, Ivan, Mächler, Martin and Yan, Jun, (2018). *Elements of Copula Modeling with R*. Springer.

Horváth, Lajos and Kokoszka, Piotr, (2012). *Inference for Functional Data with Applications*, volume 200. Springer Science & Business Media.

Hubert, Mia, Rousseeuw, Peter J and Vanden Branden, Karlien, (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47**,** no. 1, pp. 64–79.

Hussain, Alefiya, Heidemann, John and Papadopoulos, Christos, (2003). A framework for classifying denial of service attacks. In *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 99–110.

Jacques, Julien and Preda, Cristian, (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71**,** pp. 92–106.

Joe, Harry, (2015). *Dependence Modeling with Copulas*. Chapman & Hall.

Kokoszka, Piotr and Reimherr, Matthew, (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC.

Krzyśko, Mirosław and Smaga, Łukasz, (2020). Measuring and testing mutual dependence of multivariate functional data. *Statistics in Transition*, 21**,** no. 3, pp. 21–37.

Krzyśko, Mirosław and Smaga, Łukasz, (2021). Two-sample tests for functional data using characteristic functions. *Austrian Journal of Statistics*, 50**,** no. 4, pp. 53–64.

Liao, Hung-Jen, Lin, Chun-Hung Richard, Lin, Ying-Chih and Tung, Kuang-Yuan, (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, 36**,** no. 1, pp. 16–24.

Modi, Chirag, Patel, Dhiren, Borisaniya, Bhavesh, Patel, Hiren, Patel, Avi and Rajarajan, Muttukrishnan, (2013). A survey of intrusion detection techniques in Cloud. *Journal of Network and Computer Applications*, 36**,** no. 1, pp. 42–57.

Nelsen, Roger, (2006). *An Introduction to Copulas*. Springer.

Nishanth, N. and Mujeeb, A., (2020). Modeling and detection of flooding-based denial-of-service attack in wireless ad hoc network using Bayesian inference. *IEEE Systems Journal*, 15**,** no. 1, pp. 17–26.

Peng, Chen, Xu, Maochao, Xu, Shouhuai and Hu, Taizhong, (2018). Modeling multivariate cybersecurity risks. *Journal of Applied Statistics*, 45**,** no. 15, pp. 2718–2740.

Ramsay, James and Silverman, Bernard, (2005). *Functional Data Analysis*. Springer.

Sambangi, Swathi and Gondi, Lakshmeeswari, (2020). A machine learning approach for DDoS (distributed denial of service) attack detection using multiple linear regression. In *Proceedings*, volume 63, p. 51. MDPI.

Soysal, Murat and Schmidt, Ece Guran, (2010). Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. *Performance Evaluation*, 67**,** no. 6, pp. 451–467.

Wu, Zhijun, Yue, Meng, Li, Douzhe and Xie, Ke, (2015). SEDP-based detection of low-rate DoS attacks. *International Journal of Communication Systems*, 28**,** no. 11, pp. 1772–1788.

Xu, Maochao, Hua, Lei and Xu, Shouhuai, (2017). A vine copula model for predicting the effectiveness of cyber defense early-warning. *Technometrics*, 59**,** no. 4, pp. 508–520.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 23–41,* https://doi.org/10.59170/stattrans-2024-002
Received – 08.01.2023; accepted – 21.04.2023

# Mutual information between Polish subindexes – the use of copula entropy around the time of the COVID-19 pandemic

**Henryk Gurgul**[1], **Robert Syrek**[2]

## Abstract

In this paper, the copula theory is used to describe the dependence structure between variables, while the information theory provides the tools necessary to measure the uncertainty associated with these variables. What both theories have in common is copula entropy, which is strictly related to mutual information.
The findings of this study, focusing on the dependence of the (sub)indexes of the Polish stock market during the pandemic period, may prove useful not only to investors from Poland, but also from other countries, especially Central European, in making investment decisions.
The results of calculating the interdependencies between WIG, sectoral indexes and among sectoral indexes of the Polish economy using copula entropy and Pearson's correlation are quite different.
The source of the basic difference between copula entropy and Pearson's correlation is that the former enables the measurement of nonlinear interdependencies, while the latter not. The interrelations on the stock markets are nonlinear and returns are not normally distributed in general. The use of copulas is also superior in terms of ranking correlation, as it is more general and allows the examination of the structure of dependencies between extreme values.

**JEL Classification:** G15, G19

**Key words:** Polish subindexes, COVID-19 pandemic, mutual information, copula entropy.

## 1. Introduction

At the end of 2019, the COVID-19 pandemic broke out. According to WHO, by September 1 2020 there were 25,327,098 cases of COVID-19. In addition 848,255 deaths were registered across the world. Since the outbreak of the pandemic, governments have tried to restrict the spread of COVID-19.

[1] Department of Applications of Mathematics in Economics, AGH University of Science and Technology in Cracow, Poland. E-mail: henryk.gurgul@gmail.com. ORCID: https://orcid.org/0000-0002-6192-2995.
[2] Institute of Economics, Finance and Management, Jagiellonian University in Cracow, Poland. E-mail: robert.syrek@uj.edu.pl. ORCID: https://orcid.org/0000-0002-8212-8248.

All countries have used different measures in order to protect societies from the pandemic. These measures include stopping production and quarantining people in their own homes. The implementation of these measures has had a great impact on economic development, the economic situation of enterprises and national economies. These economic and social problems are reflected in the growing number of academic contributions. These studies have documented the negative impact of the pandemic (Goodell, 2020 among others) on trade, tourism, transportation, and employment (Leduc and Liu, 2020), even at the beginning. Some contributors have started to compare the effects of the spread of COVID-19 and its consequences to those of an economic crisis (Sharif et al., 2020).

The pandemic has had a great effect on economic and social development as well as the financial markets. Some studies have documented the impact of the pandemic on the returns of financial markets (Ashraf, 2020; Zhang et al., 2020; Aslam et al., 2020 a,b,c) and/or their volatility (Albulescu, 2020; Bakas and Triantafyllou, 2020; Zaremba et al., 2020; Okorie and Lin, 2020).

The risk of contagion between financial markets was also the subject of these inquiries (Akhtaruzzaman, 2020; Goldstein and Pauzner, 2004). Baig et al. (2020) investigated the impact of the pandemic on the liquidity and volatility of the stock market. They established that the increase in confirmed cases and deaths due to the pandemic caused a lack of liquidity, stability and strong volatility on the financial market.

Rizwan et al. (2020) investigated the banking systemic risk in eight important countries. All of them were strongly affected by the pandemic. The authors found that the financial systemic risk of the countries under consideration rose significantly during the pandemic period.

Some studies have concentrated on the performance of stocks in different sectors or different countries. Mazur et al. (2020) examined the return of the healthcare, food, natural gas, and software sectors. They observed that these sectors performed well during the pandemic. However, they also found that the crude petroleum, real estate, entertainment, and hospitality sectors declined noticeably. In addition, these sectors displayed great volatility.

Shehzad et al. (2020) compared the impact of the pandemic on the stock market with that of global financial crises. They established that the American and the European stock markets were affected by the pandemic more strongly, and COVID-19 disturbed economic communication throughout the world and was the source of a financial crisis.

For investors it is very important to analyze the interdependence structure of the stock market. This is important with respect to diversifying investment and building

investment portfolios during the time of the pandemic, which is essential from the point of view of risk management of the financial market taken into account by financial regulators.

In recent years, several researchers have tried to investigate the interdependence among stock markets (Sukcharoen and Leatham, 2016; Long et al., 2016; Qiao et al., 2016; Long et al.,2017a, b; Surya et al., 2018; Alomari et al., 2018; Huang et al., 2019, Kodres and Pritsker 2002, Barberis et al., 2005, Chiang and Zheng, 2010, Wang and Hui, 2018). They have used, among others, the GARCH model, Copula model, Granger causality test, DCC model, and some other models in order to detect the interdependence structure between different stock sectors in the countries under consideration.

Research on the interdependence structure of the stock markets has indentified which sector plays the most important role in a national economy. These studies provide new opportunities for investors to build a proper portfolio of assets (Poynter et al., 2015). However, in Europe there are not many studies concerned with the interdependence structure of the stock sectors during the selected period of the COVID-19 pandemic.

China was the first country that faced COVID-19. This country was the first in the world to implement measures to tackle the pandemic. Taking scientific investment methods into account, investors expect to obtain higher profits and/or reduce investment losses. A very well- known investment strategy is diversification. According to this strategy, assets are distributed to stocks from different sectors. Its main goal is to avoid investment losses caused by investing in closely dependent assets.

The pandemic, which began in 2019 in China, was the source of the greatest recession in economic and social development since the global financial crisis of 2008. Identifying the structure and changes in the interdependence between various sectors during the time of COVID-19 is a very useful piece of advice for investors trying to optimize their investment during the pandemic.

The copula entropy used in this contribution is a combination of copula theory and information theory. The copula function is employed to describe the dependence between variables, and mutual information is used to quantify the dependence. There is a connection between copula theory and information theory, and mutual information can be expressed in terms of copulas, as copula entropy.

One of the first and most important contributions using copula entropy is a paper by Zhao and Liu, 2011. In this research, the copula entropy model was constructed by the copula and the entropy theory. Therefore the copula entropy model combines the advantages of both of them. The used approach is not limited to measuring the linear correlation; it also can describe the nonlinear correlation. It not only measures the

degree of the dependence, but considers the structure. In this paper, the contributors propose copula entropy models with two and three variables to measure dependence in stock markets, which extend the copula theory and are based on Jaynes's information criterion.The research sample is composed of 12 stocks indexes from 12 countries selected by two methods. They chosen three copula functions to represent three different economic situations: recession, boom and interim. Having completed the two experiments, they provided a comparative analysis. The authors proven that three-variable dependence changes across the three economic circles are less obvious than the two-variable dependence.

This study of the dependence of the Polish stock market during the pandemic period may be useful not only for Polish investors in making investment decisions, but also those from other countries, especially Central Europe. Our aim to study interdependencies around the time of the outbreak of COVID-19 seems to be reasonable.

The main task of this study is detection of changes in dependence around event day (13.03.2020 - the day a state of epidemic threat was introduced in Poland). We will prove the dependence of subindexes using the concept of mutual information before and after the event day. By means of mutual information based on copula entropy we aim to check whether the parameters of mutual information are greater before or after the event day. Further research question concerns behaviour of Pearson correlation with respect to the event day.

We will compare results of both measures of dependence before the event day and after the event day and explain possible differences with respect to linear and nonlinear dependence notions.

## 2. Copulas

Sklar (1959) introduced a new class of multivariate cumulative distribution functions, which are multivariate cumulative distributions with uniform margins. Assume that random vector $(X, Y)$ has joint distribution function $F_{XY}(x, y)$ and density $f_{XY}(x, y)$. Let $F_X(x)$ and $F_Y(y)$ be marginal distributions, whereas $f_X(x)$ and $f_Y(y)$ are marginal density functions of $X$ and $Y$, respectively. Sklar's theorem (see Nelsen, 2006) states that there exists function $C$ (called the copula), such that $F(x, y) = C(F_X(x), F_Y(x))$. From this we see that the copula is a function that combines one-dimensional distributions into a multivariate (bivariate is a special case) distribution with uniform margins. Moreover, if marginal distributions are continuous the copula $C$ is unique and the equation holds

$$C(u, v) = F\left(F_X^{-1}(u), F_Y^{-1}(V)\right) \tag{1}$$

where $u, v \in [0,1]$ and $F_X^{-1}, F_Y^{-1}$ are quasi-inverses of the distribution of $F_X$ and $F_Y$, respectively. The density of copula $C$ is the mixed second derivative of $C$ and can be expressed as

$$c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v} = \frac{f_{XY}(x,y)}{f_X(x) \cdot f_Y(y)} \tag{2}$$

The best known classes of copulas are elliptical and Archimedean copulas. The bivariate Gaussian (or normal) copula is an elliptical copula of the form

$$C_\rho^{Ga}(u_1, u_2) = \Phi_\rho\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right) \tag{3}$$

where $\Phi_\rho$ is the cumulative distrubution function of the bivariate standard normal with Pearson's correlation coefficient $\rho$, while $\Phi^{-1}$ is the inverse of the univariate cumulative distribution function of the standard normal.

The other example of elliptical copula is copula *t* which is based on the *t* distribution function and is given by

$$C_{\nu,\rho}^{t}(u_1, u_2) = t_{\nu,\rho}\left(t_\nu^{-1}(u_1), t_\nu^{-1}(u_2)\right). \tag{4}$$

where $t_{\nu,\rho}$ is the cumulative distribution function of the bivariate *t* cumulative distribution function with linear correlation coefficient $\rho$ and $\nu$ degrees of freedom, whereas $t_\nu^{-1}$ is the inverse of the univariate cumulative distribution function of *t* with $\nu$ degrees of freedom.

The other class of copulas is Archimedean copulas, whose construction is based on a special convex and strictly decreasing continuous function called generator (see Nelsen (2006) for details). In Table 1 we present the definitions of the selected copulas and the range of parameters

**Table 1:** Some families of Archimedean copulas

| **name** | $\boldsymbol{C(u,v)}$ | **Range of parameter** |
|---|---|---|
| Frank | $-\frac{1}{\theta}\log\left[1+\frac{(\exp(-\theta u)-1)(\exp(-\theta v)-1)}{\exp(-\theta)-1}\right]$ | $(-\infty,\infty)\backslash\{1\}$ |
| Clayton | $\max([u^{-\theta}+v^{-\theta}-1]^{-\frac{1}{\theta}}, 0)$ | $[-1,\infty)\backslash\{0\}$ |
| Gumbel | $\exp(-\left[(-lnu)^\theta+(-lnv)^\theta\right]^{\frac{1}{\theta}})$ | $[1,\infty)$ |
| BB1 | $\left\{1+\left[(u_1{}^{-\theta}-1)^\delta+(u_2{}^{-\theta}-1)^\delta\right]^{1/\delta}\right\}^{-1/\theta}$ | $\theta \in (0,\infty), \delta \in [1,\infty)$ |

The number of copulas can be easily extended using rotations. In applications, the most frequently used copulas are those rotated 180 degrees, called survival copulas of the form

$$\hat{C}(u,v) = u + v - 1 + C(1-u, 1-v). \tag{5}$$

The main use of copulas is to model dependence. Such concordance measures as Kendall's $\tau$ or Spearman's $\rho$ can be expressed in terms of copulas, but the dependencies between extreme values which can be investigated with copulas are often more interesting. Upper- and lower-tail dependence coeefficients are defined as

$$\lambda_U = \lim_{u \to 1^-} \frac{1-2u+C(u,u)}{1-u} \text{ and } \quad \lambda_L = \lim_{u \to 0^+} \frac{C(u,u)}{u} \tag{6}.$$

The copula C has upper-tail dependence if $\lambda_U \in (0,1]$ and upper-tail independence if $\lambda_U = 0$. The definitions for lower-tail coefficients are analogous. The upper- (lower-) tail dependence coefficients of survival copula are equal to lower- (upper-) tail dependence coefficients.

The Gaussian copula exhibits tail independence for both tails and for $t$ copula $\lambda_U = \lambda_L = 2t_{v+1}\left(-\sqrt{\frac{(v+1)(\rho-1)}{1+\rho}}\right)$. The Archimedean copulas include non-symmetrical cases. The tail-dependence coefficients of the families selected are presented in the table below.

**Table 2:** Tail dependence coefficients of some Archimedean copulas

| **copula** | $\lambda_U$ | $\lambda_L$ |
|---|---|---|
| Frank | 0 | 0 |
| Clayton | 0 | $2^{-1/\theta}$ |
| Gumbel | $2-2^{1/\theta}$ | 0 |
| BB1 | $2-2^{1/\delta}$ | $2^{-1/(\delta\theta)}$ |

The objective of some research, such as that of Ma and Sun (2011), is to present a copula entropy approach based on entropy theory and copula theory to measure the dependence relationship between the financial variables with practical applications.

## 3. Linear correlation vs. mutual information

Most research methods that are concerned with the dependence of the stock markets are based on linear assumptions. Some of them refer to a specific model and parameters. The best known Pearson correlation coefficient can only measure the linear relationship between variables, instead of effectively measuring the nonlinear relationship. Pearson's correlation coefficient is based on the multivariate ellipticity assumption, which does not always hold. This measure will not estimate the dependence between two variables properly when the sample size is not large enough or the dependence relationship is nonlinear.

The rank correlation coefficient can be used in order to estimate the nonlinear dependence relationship between two variables. It has no restriction regarding the distribution of variables. The rank correlation coefficient primarily includes the

Kendall correlation coefficient and the Spearman correlation coefficient, which are examples of concordance measures.

Due to the development of entropy theory and its application, different methods, such as mutual information, have been used in a number of pieces of financial market research (Fiedor, 2014; Fiedor, 2015; Yang et al., 2013; Yang et al., 2014; Kwon and Yang, 2008; Wang and Hui, 2017, Wang et al., 2017 Khan et al., 2007). We briefly present the fundamental concepts of information theory, such as entropy and mutual information.

Entropy is the average amount of information. For discrete random variable $X$ with support set $\mathcal{X}$ is given by

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_m p(x) \tag{7}$$

If the logarithm base is equal to 2, the unit of entropy is bit. For $m = e$ and $m = 10$ we get nat and dit respectively (we omit this parameter in the following formulas).

In the case of a pair of random variables $X$ and $Y$ one can compute conditional entropy $H(Y|X)$, which measures the entropy of variable $Y$ when the values of $X$ are known. This is given by

$$H(Y|X) = -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log\frac{p(x,y)}{p(x)} \tag{8}$$

where $p(x,y) = P(X = x, Y = y)$ and $p(y) = P(Y = y)$. Defining joint entropy by

$$H(X,Y) = -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log p(x,y) \tag{9}$$

conditional entropy can be expressed as $H(Y|X) = H(X,Y) - H(X)$. In this paper, we use mutual information ($MI$) to measure the dependence between variables. Mutual information is given by $MI(X,Y) = H(X) - H(Y|X)$ and for the two given variables $X$ and $Y$, assuming that their respective marginal probability distributions and joint probability distribution are known, and they are, respectively, $p(x)$, $p(y)$, and $p(x, y)$, mutual information can be expressed as

$$MI(X,Y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{10}$$

Mutual information measures the amount of information of $X$ contained in $Y$ and, conversely, the amount of information of $Y$ contained in $X$. In other words, mutual information measures the uncertainty of one variable given knowledge of the other.

This measure has the following properties:

1. $MI(X,Y) \geq 0$
2. $MI(X,X) = H(X)$
3. $MI(X,Y) \leq \min(H(X), H(Y))$.

Although $MI$ is bounded by the entropies of each variables, it is not normalized. Following Joe (1989) we can normalize mutual information using formula $\delta = \sqrt{1-\exp(-2MI)}$. In this way we obtain a normalized index, which is contained in interval $[0,1]$.

For continuous random variables, the definitions above can be reformulated in terms of integrals. In this case, the Shannon entropy is called differential entropy, which unfortunately does not have all the desired properties, such as a discrete version (for example non-negativity). Mutual information is given by

$$MI(X,Y) = \int_y \int_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy. \tag{11}$$

In the literature (for example Jenison and Reale 2004, Ma and Sun 2011, Zhao and Liu, 2011), we can also find the term copula entropy. Using copula entropy, association information and dependence structure information can be measured simultaneously. In addition, copula entropy does not impose constraints on the dimension of multiple variables. Copula entropy can also be used to measure multivariate dependence in different branches of the economy. In general, for N-dimensional copula with density $c(\boldsymbol{u})$, where $\boldsymbol{u} = (u_1, u_2, \dots, u_N)$ copula entropy is defined as

$$H_C(\boldsymbol{u}) = -\int_{[0,1]^N} c(\boldsymbol{u}) \log c(\boldsymbol{u})\, d\boldsymbol{u}, \tag{12}$$

which in the bivariate case takes the form

$$H_C(U,V) = -E[\log c(u,v)] = -\int_0^1 \int_0^1 c(u,v) \log c(u,v)\, dudv. \tag{13}$$

Ma and Sun (2011) show in their paper that mutual information is copula entropy and more specifically $MI = -H_C$. They also implement a method for estimating mutual information in a non-parametric way. Given the density of the copula, one can use numerical integration to obtain $H_C$, or simply as $-\frac{1}{n}\sum_{t=1}^{n} \log c(u_t, v_t)$. This is a general approach, but for some families of copulas there are explicit formulas. For example, in the case of the Gaussian copula with parameter $\rho$, mutual information is equal to $-\frac{1}{2}\ln(1-\rho^2)$.

To summarize, dependence measurement using mutual information expressed in terms of copulas has many advantages. It is not limited to measuring linear correlations; it can also capture a nonlinear correlation. It not only measures the degree of the dependence, but also considers the dependence structure which is more than correlation. Moreover, there is no assumption about the ellipticity of marginal and joint distribution. It even allows the dependence of variables with different cumulative distribution functions to be modelled. Although in this paper we consider bivariate copulas, extension to the multidimensional case is obvious.

## 4. Data and empirical results

We consider the closing prices of 14 sectoral indexes between January 3, 2018 and May 13, 2022 (the day the state of epidemic in Poland was lifted). In addition, the largest index from the Warsaw Stock Exchange WIG is considered. On March 11, 2020 the World Health Organization (WHO) announced the COVID-19 pandemic. We divide the time series of logarithmic returns with the date 13 March, 2020. On that day a state of epidemic threat was introduced in Poland (we refer to this as the event day). For all the series we computed descriptive statistics (mean, standard deviation, kurtosis and skewness given in Table 3).

We conducted a Ljung-Box test of lack of autocorrelation and Jarque-Bera test of normality. To save space we present their quartiles (all results of the computations are available upon request).

**Table 3:** Descriptive statistics of returns

| **before event** | | | | |
|---|---|---|---|---|
| Specification | mean | sd | kurtosis | skewness |
| min | -0.21 | 0.92 | 5.31 | -2.89 |
| 1st quartile | -0.17 | 1.23 | 7.83 | -2.02 |
| median | -0.13 | 1.63 | 12.14 | -1.37 |
| 3rd quartile | -0.03 | 1.90 | 23.68 | -0.77 |
| max | 0.14 | 2.58 | 29.75 | -0.52 |
| **after event** | | | | |
| | mean | sd | kurtosis | skewness |
| min | -0.08 | 1.26 | 3.78 | -5.48 |
| 1st quartile | 0.05 | 1.97 | 6.25 | -0.30 |
| median | 0.10 | 2.26 | 6.49 | 0.02 |
| 3rd quarile | 0.14 | 2.85 | 9.04 | 0.38 |
| max | 0.19 | 3.50 | 88.23 | 1.08 |

The results of the first test are mixed, but we reject null of normality for all the series. This is also the case for the WIG series.

### Main index – subindex dependence

To investigate the dependence between the WIG index and the sectoral indexes we compute mutual information for all pairs main index - subindex before and after the event day using copula entropy.

First, we filter our time series using Vector Autoregression models for conditional means, GARCH type models for conditional variance and skew t for conditional distribution. Given the estimated models we computed the probability integral transform. Tenzer and Elidan (2016) established a monotonic relationship between the mutual information and the copula dependence parameter. We limit the set of potential

copulas to selected families, but allow their rotated versions (survival copulas). Using the Bayesian information criterion, we choose bivariate copulas that fit the best.

In most cases, the fitted copulas belong to a class of asymmetric (BB1, survival Gubmel, survival BB1) copulas with a different structure in the upper and lower tails. The computed tail dependence coefficients are not smaller after the event for 9 and 12 subindexes, respectively. Given the densities of the estimated copulas we compute mutual information and the corresponding parameter $\delta$ for all pairs under investigation. We present this in Figure 1 (red before, green after).


**Figure 1:** Parameter $\delta$ for pairs index-subindex

Both before and after the event the weakest and strongest dependence is observed for PHA and BANK, respectively.

In 9 cases out of 14 delta values are higher after the event (the exceptions are the sectors BANK, CLO, GAM, MIN, OIL). The three largest percentage increases in the delta parameter are observed for MED (32.5%), PHA (25.3%) and AUT (21.5%), the largest decreases for MIN (-21.3%), BAN (-7.1%) and GAM (-4.6).

For the purpose of comparison we computed linear correlation coefficients. Before the event, the weakest and strongest dependence is observed for MED and BANK, whereas after PHA and BANK. Only in 3 cases (AUD, IT, MED) does the correlation increase after the event, with the highest percentage increase for IT (14.1%) and largest decrease for MIN(-27.3%).

## Dependence of subindexes

We repeat the procedure for all pairs of subindexes. In most cases the copula that fits the best is a rotated version of Gumbel (63 cases before and 50 cases after the event), which is dependent in the lower tail and independent in the upper one. The number of symmetric copulas increases, and most of them are Gaussian copulas with tail

independence. However, in 50 cases the number of estimated lower-tail dependence coefficients increases. Upper-tail dependence coefficients remain the same in 72 cases because of elliptical copulas and Archimedean copulas with upper-tail independence. For the subsectors GAM, MED and MIN, the lower-tail dependence coefficients do not increase with the other sectors in most cases, whereas for FOOD and OIL they do. In the left diagram of Figure 2, we present pairs of sectors for which lower-tail dependence coefficients do not decrease (yellow) and decrease (grey). On the right, we present pairs of sectors for which upper-tail dependence increases (yellow) and does not increase (grey).

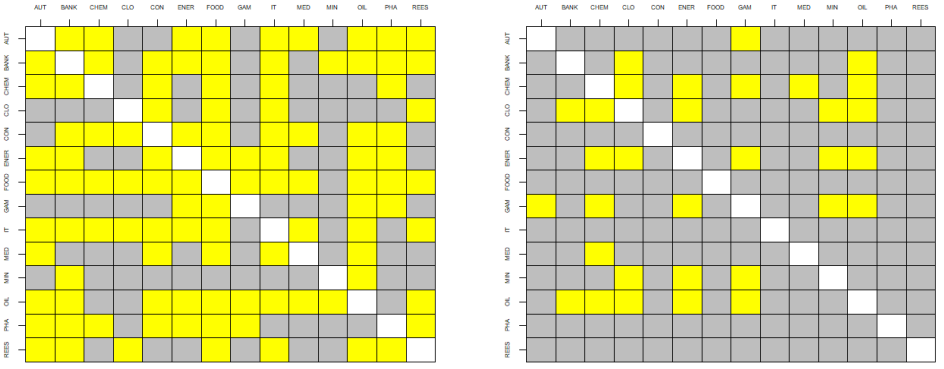


**Figure 2:** Designation of pairs with tail-dependence coefficient changes

Despite the lack of economic justification, we compute mutual information and cooresponding paramerer $\delta$ for all pairs and for both subperiods (see Figure 3).

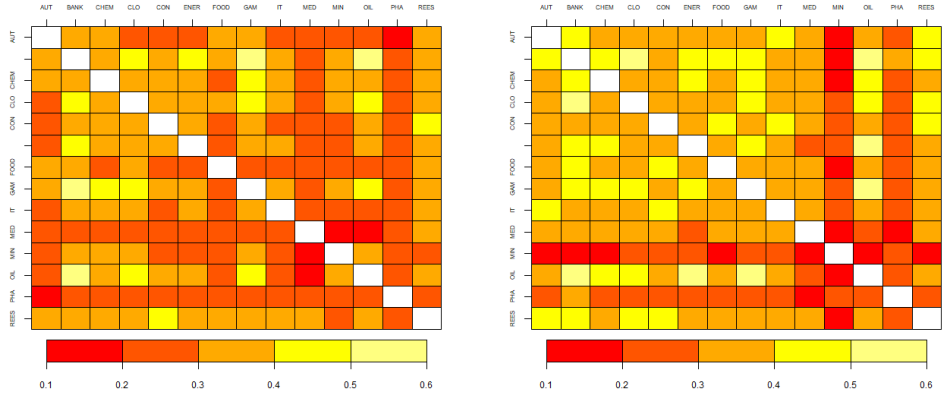


**Figure 3:** Heat map of parameters delta before (left) and after (right) the event

In Table 4 we present the 3 weakest and 3 strongest relationships before and after the event.

**Table 4:** Selected weakest and strongest relationships

| before event | | after event | |
|---|---|---|---|
| **pair** | **value** | **pair** | **value** |
| MIN - MED | 0.1911 | REES - MIN | 0.1640 |
| OIL - MED | 0.1928 | MIN - CHEM | 0.1659 |
| PHA - AUT | 0.1990 | MIN - BANK | 0.1716 |
| … | | … | |
| CLO - BANK | 0.5000 | CLO - BANK | 0.5149 |
| OIL - BANK | 0.5078 | OIL - GAM | 0.5308 |
| GAM - BANK | 0.5606 | OIL - BANK | 0.5909 |

In 69 cases dependence increases, which accounts for over 75% of all cases. In Figure 4, we present the heat map of the percentage changes of delta coefficients.

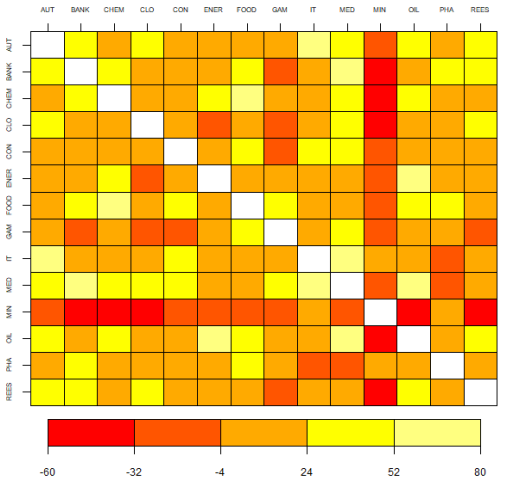


**Figure 4:** Heat map of percentage changes of $\delta$

From Figure 4 we notice that mining is the sector with the lowest percentage changes in dependencies. Table 5 contains the three smallest and three largest percentage changes.

**Table 5:** Selected smallest and largest percentage changes of $\delta$

| pair | percentage change |
|---|---|
| MIN - BANK | -53.7 |
| MIN - CHEM | -52.1 |
| MIN - REES | -40.6 |
| .... | |
| MED - IT | 59.7 |
| FOOD - CHEM | 68.1 |
| IT - AUT | 74.8 |

Given $\delta_{ij}$, the dependence parameter between $i$ and $j$ subindexes, we compute $S_i = \sum_j \delta_{ij}$, which reflects the sum of parameters $\delta$ of certain subindex with all of the other subindexes.

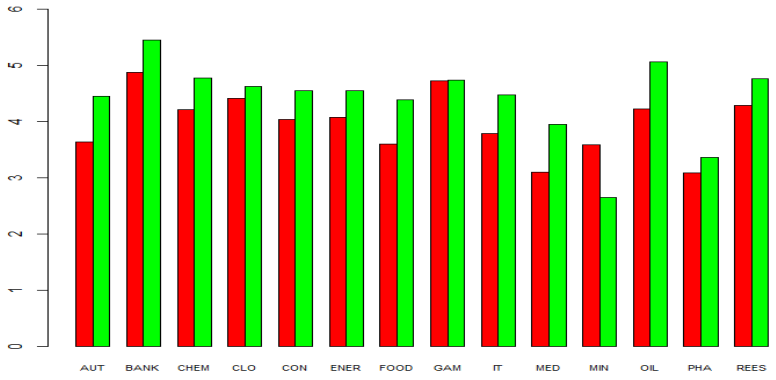


**Figure 5:** $S_i$ of each sector before (red) and after the event (green)

We can see from Figure 5 that only in the case of the mining sector $S_i$ is greater before the event than after with a drop of about 47%. The increase in the gaming sector is small (about 0.25%) while the greatest change is observed for the media sector (about 71%).

Again, for the purposes of comparison we computed the linear correlation coefficients between the returns of the subindexes. In 24 cases the dependence after the event is greater than dependence before the event. This is contrary to the results by mutual information. However, the correlation coefficient was calculated for returns. Moreover, the smallest value of correlation coefficient before the event is greater than

the smallest one after, and the maximum value of the correlation coefficient is greater before the event than after. These results are presented in the heat maps (Figure 6):

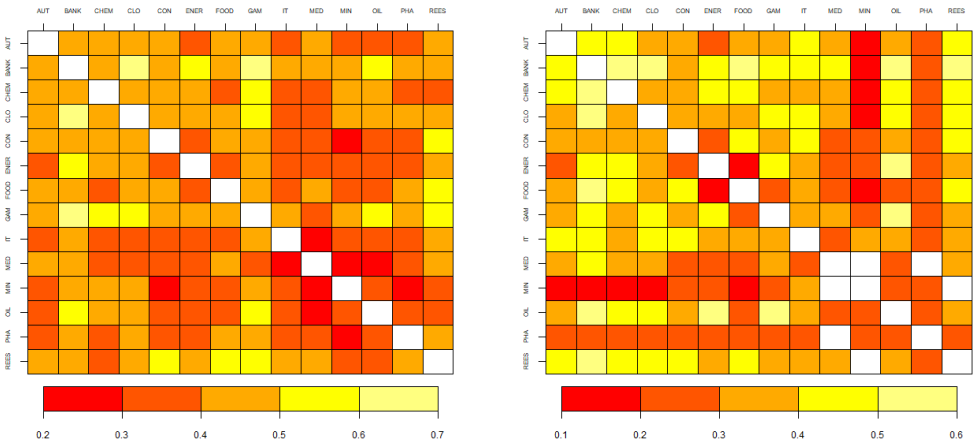


**Figure 6:** Heat maps of correlation coefficients before (left) and after (right) the event

From these pictures we notice that after the event the smallest values of correlation coefficients are when we consider the PHA or MIN sectors as a member of a pair, which is partly in line with the results from mutual information. In Table 6 we again present the 3 weakest and 3 strongest dependencies. These results are similar to those from Table 4.

**Table 6:** Selected weakest and strongest relationships measured with linear correlation coefficient.

| **before event** | | **after event** | |
|---|---|---|---|
| **pair** | **value** | **pair** | **value** |
| OIL - MED | 0.2253 | MIN - MED | 0.1320 |
| MED - IT | 0.2476 | REES - MIN | 0.1468 |
| MIN - MED | 0.2702 | PHA - MED | 0.1487 |
| … | | … | |
| GAM - CLO | 0.5509 | OIL - GAM | 0.5297 |
| CLO - BANK | 0.6253 | OIL - BANK | 0.5501 |
| GAM - BANK | 0.6671 | CLO - BANK | 0.5595 |

The last confirmation (Figure 7) of difference between mutual information and linear correlation results is the sum of the correlation coefficients of certain subindex with all of the other subindexes (red before event, green after event). The largest drop in these values is observed for MIN (about 43%) and PHA (about 37%). The only positive change is observed for the IT sector with the value of 3.7%.

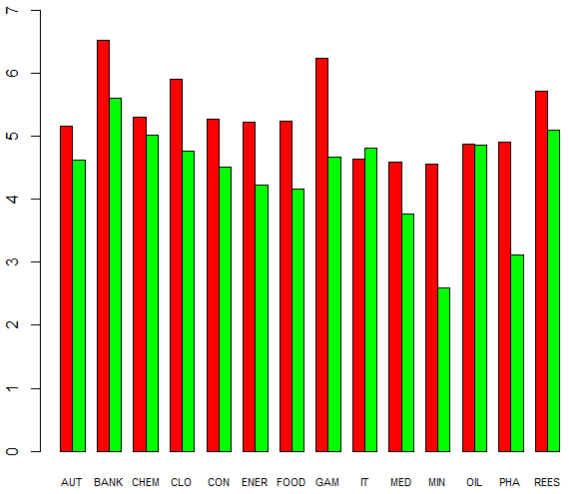


**Figure 7:** Sum of correlation coefficients of each sector before and after the event

## 5. Conclusions

The main goal of this contribution was to detect changes in dependence around the event day. The event day was 13.03.2020. On that day a state of epidemic threat was introduced in Poland. In the first part of this empirical study we examined the dependence between the WIG index and (14) sectoral subindexes. We calculated mutual information and their normalized value for all pairs main index – subindex before and after the event day using copula entropy. In most cases dependence parameters were higher after the event day than before this day.

Then we checked the dependence of subindexes using the concept of mutual information before and after the event day. In all cases, except for the subsector, mining dependence parameters were greater after the event day than before the event day. A quite different picture emerges from a linear correlation analysis. In almost all subsectors, the sum of the Pearson correlation coefficients before the event day is larger than after the event day.

The COVID-19 pandemic was reflected not only in a crisis in health systems, but also in an economic crisis, among other things. Past experiences, e.g. the Global Financial Crisis of 2007 and 2008, have convinced us that in times of crisis dependence between economic variables, especially variables from stock exchanges become stronger. The results before and after the event day, which are based on mutual information, are in line with our expectations formulated by observing crises in the

past. On the contrary, Pearson's correlation delivers different results. This may be caused by an increase in nonlinear dependencies between subsectors of the Warsaw Stock Exchange after the event day. Nonlinearities after the event day may be incorrectly categorized as independent by Pearson' correlation.

An interesting question in future research will be a comparison of pandemic outbreak of the mutual entropy and correlation results for subindexes of developed capital markets and a comparison of these dependence measures with results for Polish subindexes (and/or) subindexes of other emerging markets.

The second research question will be the impact of the size of subsectors on the dependence measures under study.

## References

Akhtaruzzaman, M., Boubaker, S., Sensoy, A., (2020). Financial contagion during COVID–19 crisis. *Finance Research Letters*, vol. 38, Article ID 101604.

Albulescu, C. T., (2020). COVID-19 and the United States financial markets' volatility. *Finance Research Letters*, vol. 38, Article ID 101699.

Alomari, M., Power, D. M., Tantisantiwong, N., (2018). Determinants of equity return correlations: a case study of the Amman Stock Exchange. *Review of Quantitative Finance and Accounting*, 50(1), pp. 33–66.

Ashraf, B. N., (2020). Stock markets' reaction to COVID-19: cases or fatalities? *Research in International Business and Finance*, vol. 54, Article ID 101249.

Aslam, F., Awan, T. M., Syed, J. H., Kashif, A., Parveen, M., (2020a). Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. *Humanities and Social Sciences Communications*, vol. 7, p. 23.

Aslam, F., Mohti, W., Ferreira, P., (2020b). Evidence of intraday multifractality in European stock markets during the recent coronavirus (COVID-19) outbreak. *International Journal of Financial Studies*, 8(2), p. 31.

Aslam, F., Aziz, S., Nguyen, D. K., Mughal, K. S., Khan, M., (2020c). On the efficiency of foreign exchange markets in times of the COVID-19 pandemic. *Technological Forecasting and Social Change*, vol. 161, Article ID 120261

Baig, S., Butt, H, A., Haroon, O., Rizvi S., (2020). Deaths, panic, lockdowns and US equity markets: the case of COVID-19 pandemic. *Finance Research Letters*, vol. 38, Article ID 101701.

Bakas, D., Triantafyllou, A., (2020). Commodity price volatility and the economic uncertainty of pandemics. *Economics Letters*, vol. 193, Article ID 109283.

Barberis, N., Shleifer, A., Wurgler, J., (2005). Comovement. *Journal of Financial Economics*, 75 (2), pp. 283–317.

Chiang, T C., Zheng, D., (2010). An empirical analysis of herd behavior in global stock markets. *Journal of Banking & Finance*, 34(8), pp. 1911–1921.

Cerqueti, R., Rotundo, G., Ausloos, M., (2018). Investigating the Configurations in Cross-Shareholding: A Joint Copula-Entropy Approach. *Entropy*, 20(2), pp.134–134. https://doi.org/10.3390/e20020134.

Fiedor, P., (2014). Networks in financial markets based on the mutual information rate. *Physical Review E*, 89(5), Article ID 052801.

Fiedor, P., (2015). Mutual information-based hierarchies on Warsaw Stock Exchange. *Acta Physica Polonica A*, 127(3a), A–33.

Goldestein, I., Pauzner, A., (2004). Contagion of self-fulfilling financial crises due to diversification of investment portfolios. *Journal of Economic Theory*, 119(1), pp 151–183.

Goodell, W., (2020). COVID-19 and finance: agendas for future research. *Finance Research Letters*, vol. 35, Article ID 101512.

Huang, Ji C., Cao, Y., Hu, S., (2019). The network structure of Chinese finance market through the method of complex network and random matrix theory. *Concurrency and Computation: Practice and Experience*, 31(9), Article ID e4877.

Jenison, R., L., Reale, R. A., (2004). The Shape of Neural Dependence. *Neural Computation*, 16, pp. 665–672.

Joe, H., (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405), pp. 157–164.

Khan, S., Bandyopadhyay, S.,. Ganguly, A. R., et al., (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E,* 76(2), Article ID 026209.

Kodres, E., Pritsker, M., (2002). A rational expectations model of financial contagion. *The Journal of Finance*, 57(2), pp. 769–799.

Kwon, O., Yang, J-S., (2008). Information flow between composite stock index and individual stocks. *Physica A: Statistical Mechanics and Its Applications*, 387(12), pp. 2851–2856.

Leduc, S., Liu, Z., (2020). The uncertainty channel of the coronavirus. *FRBSF Economic Letter*, 7, pp. 1–5.

Long, W., Guan, L., Shen, J., Song, L., Cui, L., (2017a). A complex network for studying the transmission mechanisms in stock market. *Physica A: Statistical Mechanics and Its Applications*, 484, pp. 345–357.

Long, W., Tang, Y., Cao, D., (2016). Correlation analysis of industry sectors in China's stock markets based on interval data. *Filomat*, 30(15), pp. 3999–4013.

Long, H., Zhang, J., Tang, N., (2017b). Does network topology influence systemic risk contribution? a perspective from the industry indices in Chinese stock market. *PLoS One*, 12(7), Article ID e0180382.

Mazur, M. D., Vega, M., (2020). COVID-19 and the March 2020 stock market crash. evidence from S&P1500. *Finance Research Letters*, vol. 38, Article ID 101690.

Ma, J., Sun, Z., (2011). Mutual Information Is Copula Entropy. *Tsinghua Science & Technology*, 16(1), pp. 51–54.

Nelsen, R. B., (2006). An Introduction to Copulas. 2nd Edition, *Springer*, New York.

Okorie, D. I., Lin B. Q., (2020). Stock markets and the COVID-19 fractal contagion effects. *Finance Research Letters*, vol. 38, Article ID 101640.

Qiao, H., Xia, Y., Li Y., (2016). Can network linkage effects determine return? evidence from Chinese stock market. *PLoS One*, 11(6), Article ID e0156784.

Poynter, J. G., Winder, J. P., Tai, T., (2015). An analysis of comovements in industrial sector indices over the last 30 years. *Review of Quantitative Finance and Accounting*, 44(1), pp. 69–88.

Rizwan, S., Ahmad, G., Ashraf, D., (2020) Systemic risk: the impact of COVID-19. *Finance Research Letters*, vol. 36, Article ID 101682.

Scharfstein, D. S., Stein J. C., (1990). Herd behavior and investment. *American Economic Review*, 80(3), pp. 465–479.

Sharif, A., Aloui, C., Yarovaya, L., (2020). COVID-19 pandemic, oil prices, stock market, geopolitical risk and policy uncertainty nexus in the US economy: fresh evidence from the wavelet-based approach. *International Review of Financial Analysis*, vol. 70, Article ID 101496.

Shehzad, K., Xiaoxing, L., Kazouz, H., (2020). COVID-19's disasters are perilous than global financial crisis: a rumor or fact? Finance Research Letters, vol. 36, Article ID 101669.

Sklar, A., (1959). Functions de Repartition an Dimension Set Leursmarges. *Publications de L'In-stitut de Statistique de L'Universite de Paris.*

Sukcharoen, K., Leatham, D. J., (2016). Dependence and extreme correlation among US industry sectors. *Studies in Economics and Finance*, 33(1), pp. 26–49.

Surya, C., Natasha, G., Natasha, G., (2018). Is there any sectoral cointegration in Indonesia equity market? *International Research Journal of Business Studies*, 10(3) pp. 159–172.

Tenzer, Y., Elidan, G., (2016). On the Monotonicity of Copula Entropy. arXiv:1611.06714v1 [math.ST].

Yang, C., Chen, Y., Hao, W., Shen, Y., Tang, M., Niu L., (2014). Effects of financial crisis on the industry sector of Chinese stock market-from a perspective of complex network. *Modern Physics Letters B*, 28(13), Article ID 1450102.

Yang, C., Shen, Y., Xia, B., (2013). Evolution of Shanghai stock market based on maximal spanning trees. *Modern Physics Letters B*, 27(03), Article ID 1350022.

Wang, Y., Yue, J., Liu, S., Wang, L., (2017). Copula Entropy coupled with Wavelet Neural Network Model for Hydrological Prediction. *IOP Conf. Series: Earth and Environmental Science*, 113(2018), 012160, doi :10.1088/1755-1315/113/1/012160

Wang, X. D., Hui, X. F., (2018). Cross-sectoral information transfer in the Chinese stock market around its crash in 2015. *Entropy*, 20(9), pp. 1–14.

Zaremba, A., Kizys, R., Aharon, D. Y., Demir, E., (2020). Infected markets: novel coronavirus, government interventions, and stock return volatility around the Globe. *Finance Research Letters*, vol. 35, Article ID 101597.

Zhang, D, Hu, M., Ji, Q., (2020). Financial markets under the global pandemic of COVID-19. *Finance Research Letters*, vol. 36, Article ID 101528.

Zhao, N., Lin, W., (2011). A copula entropy approach to correlation measurement at the country level. *Appl. Math. Comput*, 218(2), pp. 628–642.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 43–62,* https://doi.org/10.59170/stattrans-2024-003
Received – 10.06.2020; accepted – 06.06.2023

# Skew Log-Logistic distribution: properties and application

**Arjun Kumar Gaire**[1], **Yogendra Bahadur Gurung**[2]

## Abstract

This paper introduces a novel three-parameter skew-log-logistic distribution. The research involves the development of a new random variable based on Azzalini and Capitanio's (2013) proposition. Additionally, various statistical properties of this distribution are explored. The paper presents a maximum likelihood method for estimating the distribution's parameters. The density function exhibits unimodality with heavy right tails, while the hazard function exhibits rapid increase, unimodality, and slow decrease, resulting in a right-skewed curve. Furthermore, four real datasets are utilized to assess the applicability of this new distribution. The AIC and BIC criteria are employed to assess the goodness of fit, revealing that the new distribution offers greater flexibility compared to the baseline distribution.

**Key words:** Log-Logistic, skew, marriage, menarche, age-specific fertility rate.

## 1. Introduction

Different families of distribution created from the baseline distribution by using different mathematical techniques have attracted the interest of statisticians and other scholars. In the literature, univariate probability distributions have been modified by adding extra parameters such as shape, scale, or location in the existing distribution, the primary aim of such extension, generalization, and modification of the existing distribution is to generate a more flexible distribution. Such new distributions have been applied to fit a distribution pattern of real-world problems such as in finance, economics, physics, biostatistics, actuarial science, reliability analysis, engineering, and many more fields. In this study, a new random variable from the application of Azzalini and Capitanio's proposition (Azzalini and Capitanio, 2013) has been introduced. For this, the Log-Logistic (LLog) distribution is chosen as a base distribution. Heavy-tailed distribution is always desired by the researcher to capture the right-tailed skewed data. This research is motivated to find the distribution to capture the unusual data or outliers present in the real dataset. Four real data sets of the age of the Nepalese mother at the birth of a child, the waiting time of customers at the bank before receiving the service, the age at first marriage of Nepalese females, and the age at

[1] Khwopa Engineering College, Bhaktapur, Nepal. E-mail: arjun.gaire@gmail.com, ORCID: https://orcid.org/0000-0002-1958-9797.
[2] Central Department of Population Studies, Kirtipur, Nepal. E-mail: gurungyb@gmail.com.

menarche of Nepalese girls have been applied to test the suitability and flexibility of the proposed distribution.

The rest of the paper is organized as follows. In Section 2, a brief review of LLog distribution has been presented. In Section 3 a new distribution, called hereafter, 'Skew-Log-Logistic' (SLLog) distribution, is formulated and some statistical properties have been derived. Section 4 includes the methods of parameter estimation and Section 5 illustrates the application and validity of a model by using the four real data sets. Finally, Section 6 concludes the paper.

## 2. Log Logistic Distribution

The LLog distribution is a popular logistic distribution, which was initially developed to model population growth by Verhulst (1838) as cited in (Tahir et al., 2014). It is a continuous distribution with a uni-model failure rate function for a non-negative random variable. If T has a logistic distribution, then $X = e^T$ has LLog distribution. It is popularly known as Fisk-distribution in economics (Fisk, 1961). This distribution is applicable for modeling in various real-world situations, viz.: wealth and income (Fisk, 1961); economics and actuarial sciences (Kleiber and Kotz, 2003); flow data in hydrology (Ashkar and Mahdi, 2006) and 'time following a heart transplantation' in biostatistics (Collet, 2015). Similarly, Yilmaz et al. (2011) used it to estimate the seismic risk and earthquake occurrence probabilities. Further, Tahir et al. (2014) applied it to study the reliability analysis. Furthermore, it is used by Surendran and Tota-Maharaj, (2015) for modeling daily water consumption, estimation, and forecasting. So, LLog is a widely applicable model in different walks of life.

The probability density function (PDF) and the cumulative distribution function (CDF) of the three-parameter LLog distribution are given as

$$g(x) = \frac{\alpha}{\beta} \frac{\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^2}, \text{ for x} > \gamma \tag{1}$$

$$G(x) = \frac{\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} \tag{2}$$

where $\alpha > 0$ is a shape parameter, $\beta > 0$ is a scale parameter and $\gamma$ is a threshold or location parameter. The random variables under study in the different situations have positive values and the minimum cutoff value of these random variables is greater than zero, such as the minimum age of the mother at the birth of a child. Here, we consider the third threshold or location parameter of the LLog distribution.

The basic properties of this distribution are studied by Kleiber and Kotz (2003), Lawless (2003), and Ashkar and Mahdi (2006). The $k^{th}$ order moments of two-parameter LLog

distribution are derived and studied by Tadikamalla (1980) for $\alpha > k$ as

$$E(x^k) = \beta^k B\left(1+\frac{k}{\alpha}, 1-\frac{k}{\alpha}\right) = \beta^k \frac{\Gamma\left(1+\frac{k}{\alpha}\right)\Gamma\left(1-\frac{k}{\alpha}\right)}{\Gamma(2)} = \frac{k\pi\beta^k}{\alpha sin\frac{k\pi}{\alpha}} \tag{3}$$

where $B(a,b)$ is the Beta function defined as $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$. Also, the value of the Beta function is computed by using the relation as $B(a,b) = \frac{\Gamma a\ \Gamma b}{\Gamma(a+b)}$.

In particular, the mean and variance of the two-parameter LLog distribution are given as

$$\text{Mean} = \frac{\pi\beta}{\alpha sin\frac{\pi}{\alpha}} \text{ for } \alpha > 1 \text{ and Variance } = \frac{2\pi\beta^2}{\alpha sin\frac{2\pi}{\alpha}} - \left(\frac{\pi\beta}{\alpha sin\frac{\pi}{\alpha}}\right)^2 \text{ for } \alpha > 2.$$

## 3. Skew Log-Logistic Distribution

The Normal distribution was extended to the Skew-Normal distribution by adding an asymmetry parameter $\lambda > 0$ (Azzalini, 1985, 2005). The PDF of a Skew-Normal distribution was derived by using the relation expressed in Equation (4).

$$f(z) = 2\,\phi(z)\,\Phi(\lambda z),\ z \in R,\ \lambda \in R \tag{4}$$

Here, $\phi(z)$ and $\Phi(z)$ are the PDF and CDF of Standard Normal distribution. The general formula for the construction of a skew-symmetrical distribution other than the Standard Normal distribution proposed by Azzalini and Capitanio (2013) is as:

$$f(x) = 2\,g(x)\,G(x),\ x \in R \tag{5}$$

where g(x) and G(x) are the PDF and CDF of any baseline distribution. Gupta et al. (2002) introduced Skew-uniform, Skew-t, Skew-Cauchy, Skew-Laplace, and Skew-logistic distributions. Later, Nadarajah (2009) studied in detail about the Skew-Logistic distribution. The base distributions chosen in all of these cases were a symmetrical distribution about the origin. However, Shaw and Buckley (2007) claimed to choose any distribution other than the symmetrical one (p. 15). Thus, in this research, the LLog distribution is chosen as a base distribution that is already positively skewed. Since the distributions proposed by different researchers are unable to catch the extreme value of data. We hope this construction of a heavy-tailed distribution could catch the unusual extreme value that exists in the data. [Note: Some or part of this research is published as a preliminary result in proceeding (Gaire et al., 2019)].

The LLog distribution is chosen as the base distribution because it has been preferred by different researchers in their generalization, modification and extension due to the flexible nature of both PDF and hazard rate functions. Different forms of generalization of the LLog distributions are found in literature used by different scholars. Some of frequently used distributions are exponentiated LLog distribution (Rosaiah et al., 2006); Beta LLog distribu-

tion (Lemonte, 2012) by using the generator introduced by Eugene et al. (2002) and Jones (2004); Kumaraswamy LLog distribution proposed by De-Santana et al. (2012) by using the relationship provided by Cordeiro and Castro (2011); Transmuted LLog distribution introduced by Aryal (2013) using the concept of the quadratic transmutation rank map of Shaw and Buckley (2007); Marshall-Olkin Extended LLog distribution proposed and studied by Gui (2013) using the concept of Marshall and Olkin (1997); Zografos-Balakrishnan LLog distribution introduced and studied by Hamedani (2013) based on the concept of Zografos-Balakrishnan generalized distribution (Zografos and Balakrishnan, 2009); McDonald LLog distribution proposed and studied by Tahir et al. (2014) using the concept of Alexander et al. (2012); Extended LLog distribution studied and presented by Lima and Cordeiro (2017) using an exponentiated generalized class of distribution of Cordeiro et al. (2013). Similarly, Additive Weibull LLog distribution has been introduced by Hemeda (2018) using the concept suggested by Hassan and Hemeda (2016); Transmuted generalized LLog distribution was studied by Adeyinka and Olapade (2019). At this juncture, the SLLog distribution is introduced and formulated; further some structural properties of the distribution are derived, along with a method of parameter estimation, and applied to four real data sets for model validity.

### 3.1. Probability Density Function of the SLLog Distribution

In this section, the PDF of the SLLog distribution is introduced. After substituting the values of $g(x)$ and $G(x)$ of the LLog distribution in Equation (5) we obtained the PDF of the SLLog distribution in Equation (6) as:

$$f(x) = \frac{2\alpha}{\beta} \frac{\left(\frac{x-\gamma}{\beta}\right)^{2\alpha-1}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{3}}, \text{ for x} > \gamma \tag{6}$$

Here, $f(x)$ is a probability density function since the total probability under a given range is unity. Figure 1 depicts the plots of PDF of the distribution for the selected values of parameters. The graph shows that the PDF is right-skewed for selected values of parameters.

### 3.2. Cumulative Distribution Function of SLLog Distribution

The CDF of the SLLog distribution is defined as:

$$F(x) = \int_{\gamma}^{x} f(x)dx = \int_{\gamma}^{x} \frac{2\alpha}{\beta} \frac{\left(\frac{x-\gamma}{\beta}\right)^{2\alpha-1}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{3}} dx$$

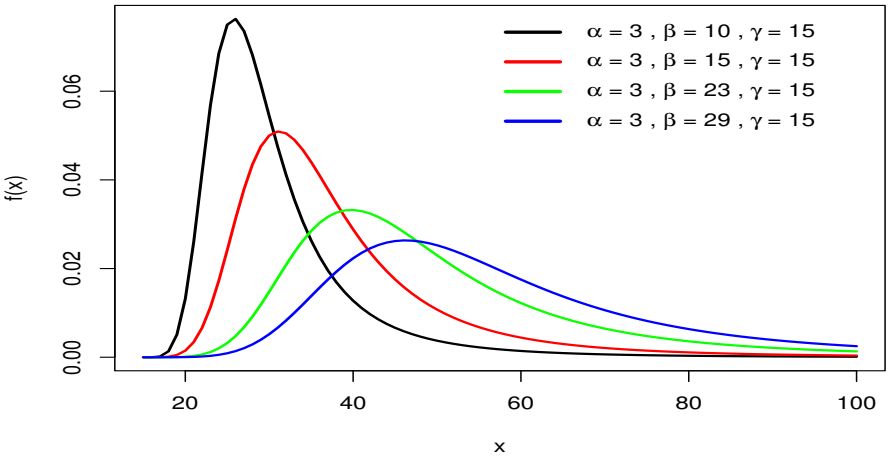


Figure 1: Plots of PDF of SLLog distribution for selected values of parameters

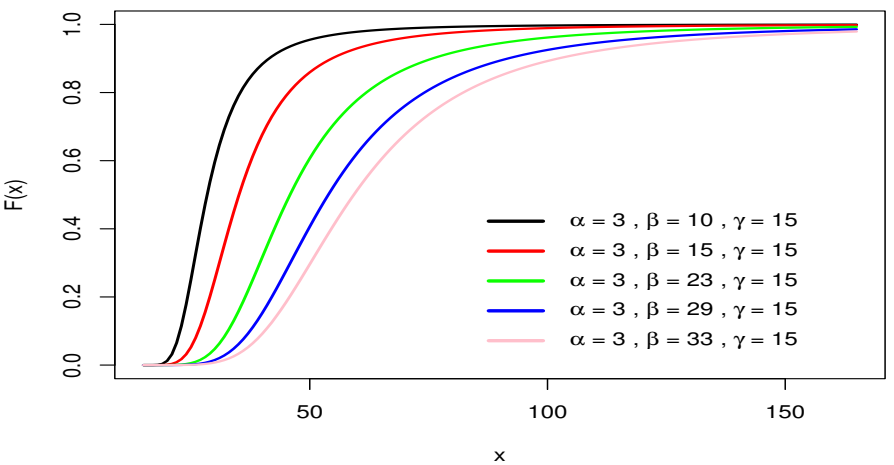


Figure 2: Graph of CDF for the selected values of parameters

On simple calculation, it gives the value of F(x) as:

$$F(x) = 1 - \frac{2}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} + \frac{1}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{2}}, \text{ for x} > \gamma \tag{7}$$

The graph of CDF of SLLog distribution has been presented in Figure 2 for the selected values of the parameters $\alpha$, $\beta$, and fixed value of $\gamma = 15$. The graph is monotonically increasing and the maximum value is 1 for a different set of parameters selected.

### 3.3. Moments About Origin of the SLLog Distribution

To calculate the $k^{th}$ order moments of the SLLog distribution about the origin, first, consider the third parameter $\gamma = 0$, then the density function (6) becomes,

$$f(x) = \frac{2\alpha}{\beta} \frac{\left(\frac{x}{\beta}\right)^{2\alpha-1}}{\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{3}}, \text{ for x} > 0$$

Now, the $k^{th}$ order moments of the SLLog distribution about the origin is

$$E(x^k) = \int_0^\infty X^k f(x)dx = \int_0^\infty x^k \frac{2\alpha}{\beta} \frac{\left(\frac{x}{\beta}\right)^{2\alpha-1}}{\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{3}} dx = 2\,\beta^k B\left(2+\frac{k}{\alpha}, 1-\frac{k}{\alpha}\right)$$

By using the relation of integration from Gradshteyn and Ryzhik (2000),

$$E(x^k) = 2\,\beta^k \frac{\Gamma\left(2+\frac{k}{\alpha}\right)\Gamma\left(1-\frac{k}{\alpha}\right)}{\Gamma 3}$$

Here, it is to be noted that moments of the SLLog distribution are only defined for $\alpha > k$ as.

$$E(x^k) = \frac{\pi(k+\alpha)\beta^k}{\alpha^2 sin\frac{k\pi}{\alpha}} \tag{8}$$

In particular, the mean and variance of the SLLog Distribution are given in Equation (9).

$$\text{Mean} = \frac{\pi(\alpha+1)\beta}{\alpha^2 sin\frac{\pi}{\alpha}} \text{ and Variance } = \frac{\pi(\alpha+2)\beta^2}{\alpha^2 sin\frac{2\pi}{\alpha}} - \left(\frac{\pi(\alpha+1)\beta}{\alpha^2 sin\frac{\pi}{\alpha}}\right)^2 \tag{9}$$

Thus, the value of the mean of SLLog distribution for $\alpha = 3$ and $\beta = 10$ is 16.1252. Table 1 gives the value of the first four moments of distribution about the origin for different values of parameters. These moments can be used to compute the value of skewness and kurtosis of the distribution. The values of moments are increased with the increase in the value of parameters.

Table 1: Value of first four moments about the origin for different values of parameters

| Parameters | $K=1$ | $K=2$ | $K=3$ | $K=4$ |
|---|---|---|---|---|
| $\alpha=8,\ \beta=10$ | 11.54 | 138.84 | 1753.35 | 23571.43 |
| $\alpha=9,\ \beta=11$ | 12.47 | 160.62 | 2146.27 | 29992.96 |
| $\alpha=10,\ \beta=12$ | 13.42 | 184.72 | 2617.31 | 38367.2 |
| $\alpha=11,\ \beta=13$ | 14.37 | 211.02 | 3170.38 | 48922.32 |

### 3.4. Random Number Generation and Quantile Function of the SLLog Distribution

A set of random numbers can be generated by using the method of inversion from the CDF of the SLLog distribution. For this, let, $F(x) = U$, where, the function $U$ follows the uniform distribution in an interval $[0,1]$ as

$$1 - \frac{2}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} + \frac{1}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{2}} = U$$

$$\frac{2}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} - \frac{1}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{2}} = 1 - U$$

Let, $\left(\frac{x-\gamma}{\beta}\right)^{\alpha}$ = Z. Then, it leads to: $Z^2(1-U) - 2UZ - U = 0$

This is quadratic in $Z$. After simple calculation, this becomes

$$Z = \frac{U \pm \sqrt{U}}{1-U}$$

Here, $U < 1$ so the term becomes negative and this negative term is not included in further analysis. Thus, the value of the random variable is given as

$$X = \gamma + \beta \left(\frac{U+\sqrt{U}}{1-U}\right)^{\frac{1}{\alpha}} \tag{10}$$

For the known value of parameters $\alpha$, $\beta$ and $\gamma$, one can generate a set of random numbers $X$ by using Equation (10). Similarly, by choosing the suitable value of $U$ in Equation (10) one can also get the different values of quantiles such as the first, second, and third quartiles obtained by setting $U = \frac{1}{4}$, $U = \frac{1}{2}$, and $U = \frac{3}{4}$ respectively.

### 3.5. Reliability Analysis of the SLLog Distribution

The reliability function $R(x)$ as defined by Rodriguez (2010) is simply the complement of the CDF. It is also the probability that a random variable $X$ will take a value greater than a number $x$ or the probability of an item not failing before some time $x$. So, it is defined as $R(X) = Prob(X > x) = 1 - Prob(X \leq x) = 1 - F(x)$

The graph of the reliability function of the SLLog distribution is presented in Figure 3 and the expression is given in Equation (11). The graph of the reliability function is decreasing with respect to increase in the value of variable $X$.

$$R(x) = \frac{2}{1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} - \frac{1}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{2}} = \frac{1+2\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{2}} \tag{11}$$

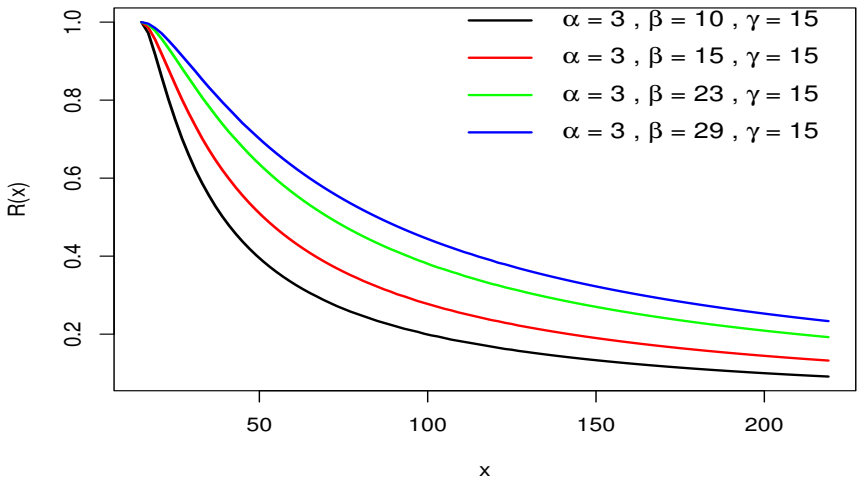


Figure 3: The plot of the reliability function for selected values of parameters

### 3.6. Hazard, Inverse Hazard, and Cumulative Hazard Rate Function

The other characteristics of interest of a random variable are the hazard and inverse hazard rate function defined as $h(x) = \frac{f(x)}{1-F(x)}$, and $rh(x) = \frac{f(x)}{F(x)}$. Thus, the hazard rate function for the SLLog distribution, which is the conditional probability of failure, given that it has survived up to the time $x$ is given in Equation (12), and the graph of the hazard rate function is presented in Figure 4. Similarly, the inverse hazard rate function defined by (Barlow et al., 1963) for SLLog is present in Equation (13). The hazard function increases fast along with the uni-modality and decreases slowly creating a right skew curve.

$$h(x) = \frac{2\alpha}{\beta} \frac{\left(\frac{x-\gamma}{\beta}\right)^{2\alpha-1}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)\left(1+2\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)} \tag{12}$$

Similarly, the Inverse hazard rate function of the SLLog distribution is given as:

$$rh(x) = \frac{2\alpha}{\beta} \frac{1}{\left(\left(\frac{x-\gamma}{\beta}\right)+\left(\frac{x-\gamma}{\beta}\right)^{\alpha+1}\right)} \tag{13}$$

Furthermore, the cumulative hazard rate function of the SLLog distribution is defined by $H(x) = -ln(R(x))$ as given in Equation (14) and the graph is increasing with respect to the increase in values of variable $X$, which has been depicted in Figure 5.

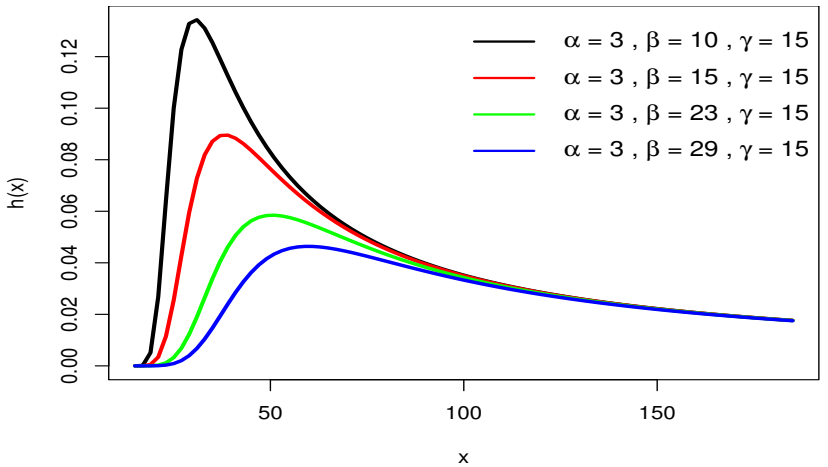


Figure 4: Plot of hazard rate function for the selected value of parameters

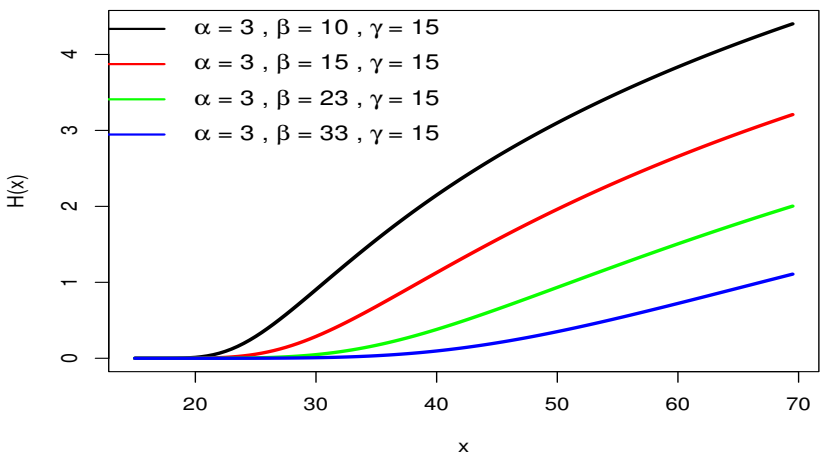


Figure 5: Graph of cumulative hazard rate function for selected values of parameters

$$H(x) = -ln(R(x)) = -ln\left(\frac{1+2\left(\frac{x-\gamma}{\beta}\right)^{\alpha}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{2}}\right)$$

$$H(x) = 2\,ln\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right) - ln\left(1+2\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right) \tag{14}$$

### 3.7. Entropy Measure of SLLog Distribution

Entropy is defined as the measure of the variation of the uncertainty of a random variable which is used in various situations in science and engineering. Different forms of the entropy are studied and compared, here we only derived the expression of two types of entropy Renyi entropy and q-entropy of the SLLog distribution.

#### 3.7.1 Reny Entropy

First of all, for the SLLog random variable $X$ with PDF $f(x)$, the Renyi entropy as defined by (Renyi, 1961) which has a similar role of kurtosis to measure and compare the shapes of densities is given as

$$I_R(\rho) = \frac{1}{1-\rho}\, ln\left(\int (f(x))^\rho dx\right)$$

Where $\rho > 0$ and $\rho \neq 1$ and $\rho$ is a real non-integer. And the integral is computed as,

$$\int_\gamma^\infty (f(x))^\rho dx = \int_\gamma^\infty \left(\frac{2\alpha}{\beta}\right)^\rho \frac{\left(\frac{x-\gamma}{\beta}\right)^{(2\alpha-1)\rho}}{\left(1+\left(\frac{x-\gamma}{\beta}\right)^\alpha\right)^{3\rho}}\, dx$$

$$= 2^\rho \left(\frac{\alpha}{\beta}\right)^{\rho-1} B\left(\frac{2\rho\alpha-\rho+1}{\alpha}, \frac{4\rho-2\rho\alpha-1}{\alpha}\right)$$

$$\int_\gamma^\infty (f(x))^\rho dx = 2^\rho \left(\frac{\alpha}{\beta}\right)^{\rho-1} \frac{\Gamma\left(\frac{2\rho\alpha-\rho+1}{\alpha}\Gamma\frac{4\rho-2\rho\alpha-1}{\alpha}\right)}{\Gamma(3\rho)}$$

Therefore, the Renyi entropy of the SLLog distribution can be expressed as

$$I_R(\rho) = \frac{1}{1-\rho} ln\left(2^\rho \left(\frac{\alpha}{\beta}\right)^{\rho-1} \frac{\Gamma\left(\frac{2\rho\alpha-\rho+1}{\alpha}\Gamma\frac{4\rho-2\rho\alpha-1}{\alpha}\right)}{\Gamma(3\rho)}\right) \tag{15}$$

#### 3.7.2 q-Entropy

For the SLLog random variable $X$ with PDF $f(x)$, the q-entropy as defined and introduced by Havarda and Charvat (1967) and later applied to physical problems by Tsallis (1988) is defined as

$$I_R(q) = \frac{1}{1-q}\left(1 - \int (f(x))^q dx\right)$$

Where $q > 0$ and $q \neq 1$ and $q$ is a real non-integer.

Therefore, after using the expression of Equation (15) with replace of $\rho$ by $q$ the q-entropy of the SLLog distribution can be expressed as

$$I_R(q) = \frac{1}{1-q}\left(1 - 2^q\left(\frac{\alpha}{\beta}\right)^{q-1}\ \frac{\Gamma\left(\frac{2q\alpha-q+1}{\alpha}\Gamma\frac{4q-2q\alpha-1}{\alpha}\right)}{\Gamma(3q)}\right) \tag{16}$$

Entropy is the average amount of information conveyed by an event when considering all possible outcomes or events drawn from the probability distribution. It is also used to measure disorder. It is also used to measure the variation of the uncertainty of a random variable in various situations in science and engineering.

## 4. Method of Parameter Estimation

To estimate the parameters involved in the SLLog distribution, the expression is derived by using the maximum likelihood estimates (MLEs) method. Let $X_1, X_2, \ldots, X_n$ be a set of $n$ samples drawn from a SLLog distribution. Then the likelihood function of this distribution is given by

$$L = \left(\frac{2\alpha}{\beta}\right)^n \prod_{i=1}^{n}\left(\frac{\left(\frac{x_i-\gamma}{\beta}\right)^{2\alpha-1}}{\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)^3}\right) \tag{17}$$

Therefore, the log-likelihood function of the SLLog distribution becomes

$$lnL = n\,ln\left(\frac{2\alpha}{\beta}\right) + (2\alpha-1)\sum_{i=1}^{n} ln\left(\frac{x_i-\gamma}{\beta}\right) - 3\sum_{i-1}^{n} ln\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right) \tag{18}$$

The components of the score vector to estimate the parameters associated with the SLLog distribution are given by

$$\frac{\partial \ln L}{\partial \alpha} = \frac{n}{\alpha} + 2\sum_{i=1}^{n} ln\left(\frac{x_i-\gamma}{\beta}\right) - 3\sum_{i-1}^{n}\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} ln\left(\frac{x_i-\gamma}{\beta}\right) \tag{19}$$

$$\frac{\partial \ln L}{\partial \beta} = -\frac{n}{\beta} - n\left(\frac{2\alpha-1}{\beta}\right) + \frac{3\alpha}{\beta}\sum_{i-1}^{n}\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} \tag{20}$$

$$\frac{\partial \ln L}{\partial \gamma} = -\left(\frac{2\alpha-1}{\beta}\right)\sum_{i=1}^{n}\left(\frac{x_i-\gamma}{\beta}\right)^{-1} + \frac{3\alpha}{\beta}\sum_{i-1}^{n}\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha-1} \tag{21}$$

By solving the nonlinear system of equations simultaneously using suitable numerical methods by setting the score vector to zero we obtain the value of parameters $\alpha$, $\beta$ and $\gamma$ of the SLLog distribution.

## 5. Application of SLLog Distribution

To test the potentiality of the proposed SLLog distribution, four real data sets are presented. To test the validity and suitability of the proposed models, Akaike's Information Criteria (AIC), and Bayesian Information Criteria (BIC) at the maximum value of Negative Log-likelihood (NLL) of probability distributions have been applied. The formulas of AIC and BIC for the fitted models are given as

$$AIC = 2k - 2\,lnL \tag{22}$$

$$BIC = k\,ln(n) - 2\,lnL \tag{23}$$

where $k$ is the number of parameters associated with the probability distribution. $n$ is the number of observations and $lnL$ is the log-likelihood function at the maximum likelihood estimate of that distribution.

The first data set is taken from the Nepal Demographic and Health Survey (NDHS, 2022). Different demographers and researchers used different right-skewed probability distribution models to test the goodness of fit of Age-Specific Fertility Rates (ASFRs) of different countries viz. Peristera and Kostaki (2007) used the Normal mixture model to capture both traditional and modern distorted ASFRs. Mazzuco and Scarpa (2011) applied a flexible generalized skew Normal distribution to fit the fertility pattern of countries that experienced a bimodal-fertility schedule eg. the USA, the UK, Ireland, and countries that keep a classic fertility pattern viz. Italy and the Czech Republic. Gaire and Aryal (2015) applied inverse Gaussian model to describe the distribution pattern of ASFRs of Nepalese mothers. Asili et al. (2014) used skew-logistic probability to fit ASFRs of Italy and the same model was applied to fit the ASFRs of India by Mishra et al. (2017). A polynomial model was used by Gaire et al. (2022). In this paper, the proposed SLLog model is applied to the age of the mother at the birth of a child to fit ASFRs of Nepal, and the results are compared with baseline distribution which are presented in Table 2.

Table 2: Parameter estimation and different test statistics for the age of the mother at the birth of a child

| Distribution | Parameters | | | NLL | AIC | BIC |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\gamma$ | | | |
| LLog | 24.612 | 0.321 | 17.445 | -602.39 | 1210.77 | 1210.61 |
| SLLog | 5.958 | 22.914 | 0.000 | -28.280 | 60.360 | 60.252 |

The second data set consists of 100 observations of the waiting time (minutes) of a customer at the bank before receiving the service and it has been taken from Ghitany et al. (2008) and recently the same data was applied to Skew-Lomax distribution by Gaire (2022). The values of parameters and the result of test statistics have been presented in Table 3.

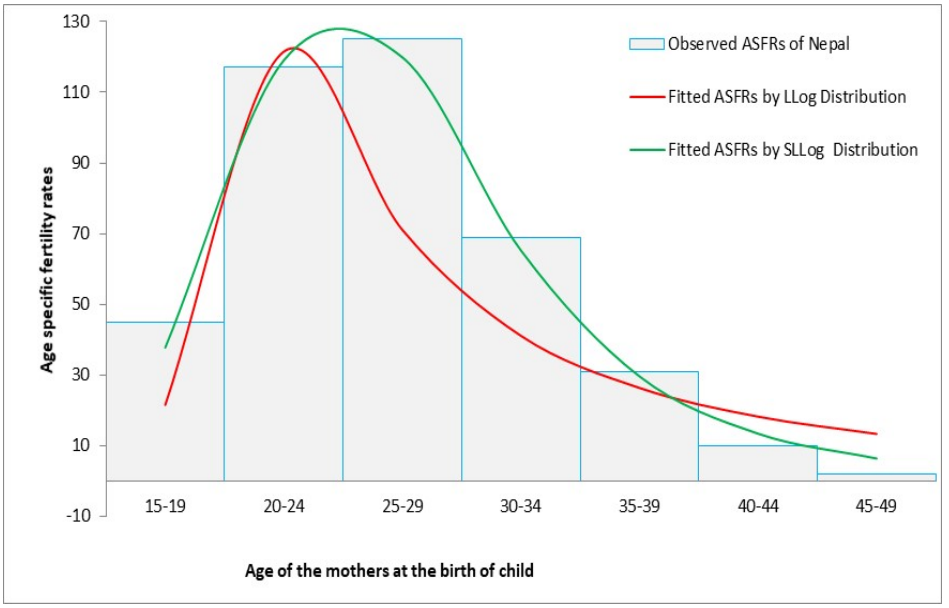


Figure 6: Empirical and fitted ASFRs of Nepalese Mothers

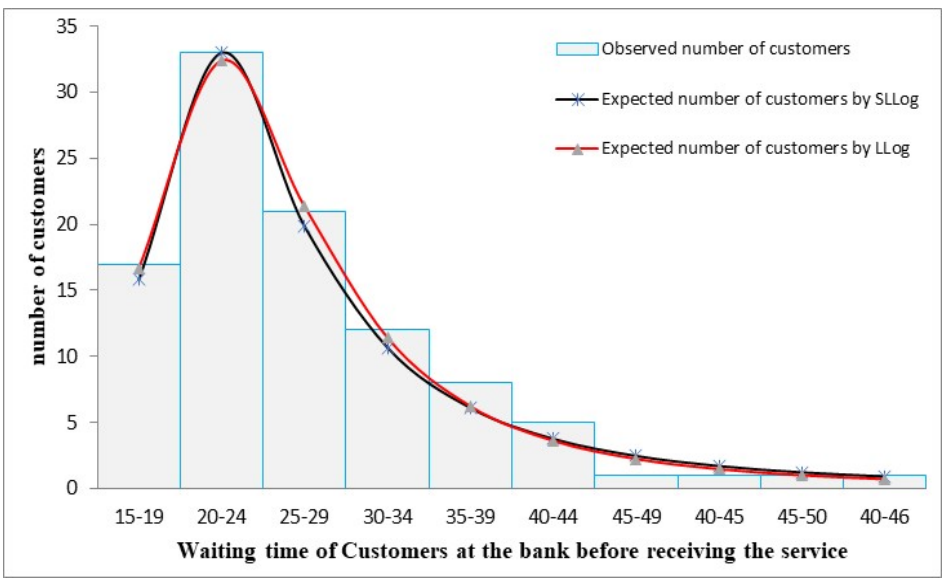


Figure 7: Empirical and fitted number of customers waiting for a service at the bank

The third data set consists of 10,631 data of age at first marriage of Nepalese women taken from (NDHS, 2022). The values of the estimated parameters along with the test statistics have been presented in Table 4.

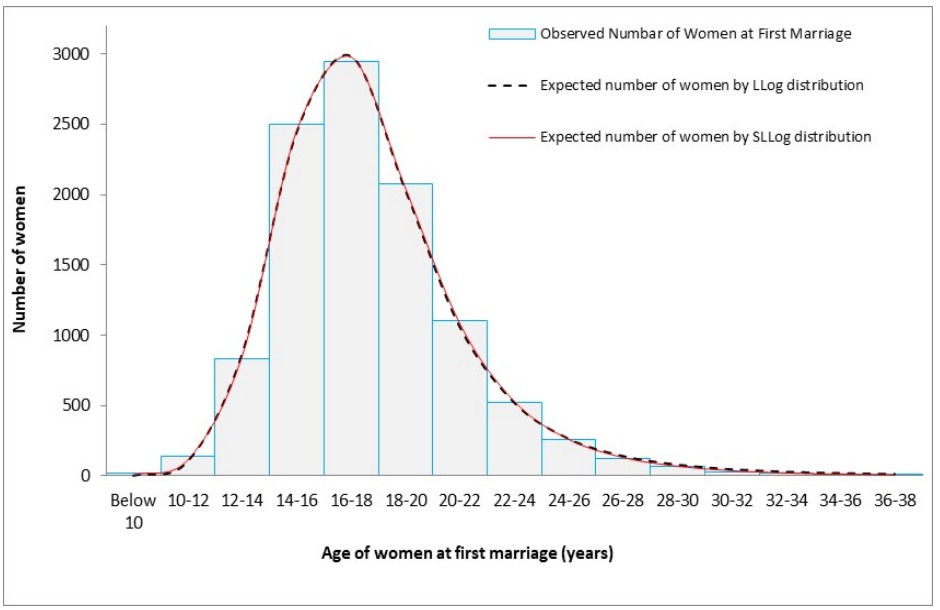


Figure 8: Empirical and fitted number of Nepalese women with age at first marriage

Table 3: Parameter Estimation and different test statistics for waiting time of customers

| Distribution | Parameters | | | NLL | AIC | BIC |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\gamma$ | | | |
| LLog | 2.185 | 7.935 | 0.198 | -44.135 | 94.270 | 95.178 |
| SLLog | 1.811 | 5.031 | 0.000 | -43.703 | 93.407 | 94.315 |

Finally, the fourth data set consists of 14,349 data on the age of girls at menarche and has been taken from (NDHS, 2022). The value of the estimated parameter along with the test statistics have been presented in Table 5.

Table 4: Parameter estimation and different test statistics for age at first marriage

| Distribution | Parameters | | | NLL | AIC | BIC |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\gamma$ | | | |
| LLog | 4.694 | 8.349 | 8.883 | -80.837 | 167.674 | 169.799 |
| SLLog | 7.994 | 14.946 | 0.522 | -74.264 | 154.528 | 156.446 |

In general smaller values of NLL, AIC, and BIC values of goodness of fit of the probability distribution suggest the best fit to the data. The values of AIC and BIC at the maximum likelihood estimate for the proposed SLLog distribution are lower than that of the LLog distribution for all four data sets. This clearly showed that the proposed model is flexible enough to fit the data better than that of the base distribution.

Table 5: Parameter estimation and different test statistics for the age of girls at menarche

| Distribution | Parameters | | | NLL | AIC | BIC |
|---|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\gamma$ | | | |
| LLog | 7.420 | 6.471 | 7.677 | -63.30 | 132.60 | 134.77 |
| SLLog | 13.644 | 13.269 | 0.000 | -49.985 | 105.970 | 107.88 |

## 6. Conclusions

A new three-parameter skew probability distribution model has been formulated as the SLLog distribution. Some of the statistical properties of the distribution have been studied. The parameter estimation method is discussed by using maximum likelihood. To test the suitability and validity of the proposed model four real data sets, viz. age of the Nepalese mother at the birth of a child, the waiting time of the customer before receiving the service, the age at first marriage of Nepalese female, and the age of Nepalese girls at menarche have been used. The AIC and BIC test criteria have been applied to test the validity and suitability of the model obtained at the maximum value of negative log-likelihood of the probability distribution. The observed values of AIC and BIC show that the proposed distribution is more flexible than the baseline distribution to fit the pattern of these real data sets.

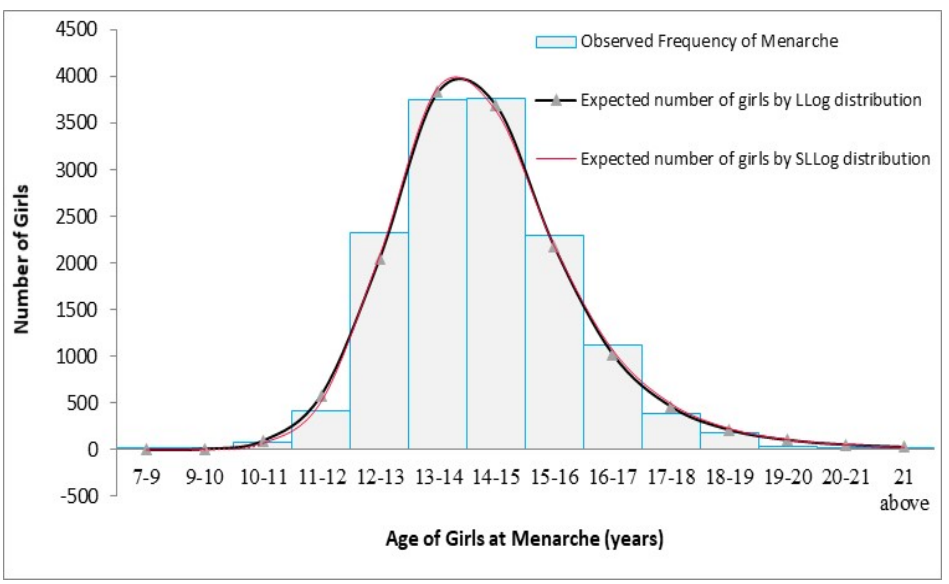


Figure 9: Empirical and fitted number of Nepalese girls at the age of menarche

## Acknowledgment

The first author would like to acknowledge the travel grant support 2019 by UGC Nepal to present the research at an international conference. Both authors wish to thank the Chief Editor, the Editors, and anonymous referees for their comments and suggestions which have greatly improved our manuscript.

## References

Adeyinka, F. S., and Olapade, A. K., (2019). On transmuted four parameters generalized Log-Logistic distribution. *International Journal of Statistical Distributions and Applications*, 5(2), pp. 32–37.

Alexander, C., Cordeiro, G. M., Ortega, E. M. and Sarabia, J. M. (2012). Generalized Beta-generated distributions. *Computational Statistics and Data Analysis*, 56(6), pp. 1880–1897.

Aryal, G. R., (2013). Transmuted log-logistic distribution. *Journal of Statistics Applications and Probability*, 2(1), pp. 11–20.

Ashkar, F., Mahdi, S., (2006). Fitting the log-logistic distribution by generalized moments. *Journal of Hydrology*, 328(3-4), pp. 694–703.

Asili, S., Rezaei, S. and Najjar, L., (2014). Using skew-logistic probability density function as a model for age-specific fertility rate pattern. *BioMed Research International*, 10, pp. 1–5.

Azzalini A., (1985). A class of distributions that includes the normal ones. *Scandinavian Journal of Statistics*, pp. 171–178.

Azzalini A., (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2), pp. 159–188.

Azzalini, A., Capitanio, A., (2013). *The skew-normal and related families* (Vol. 3), London: Cambridge University Press.

Barlow, R. E., Marshall, A. W. and Proschan, F., (1963). Properties of probability distributions with monotone hazard rate. *The Annals of Mathematical Statistics*, 34(2), pp. 375–389.

Collett, D., (2015). *Modeling survival data in medical research*. Boca Raton, Florida USA: CRC press.

Cordeiro, G. M., De-Castro M., (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, 81, pp. 883–898.

Cordeiro, G. M., Ortega, E. M. and Da-Cunha, D. C., (2013). The exponentiated generalized class of distributions. *Journal of Data Science*, 11(1), pp. 1–27.

De-Santana, T. V. F., Ortega, E. M., Cordeiro, G. M. and Silva, G. O., (2012). The Kumaraswamy-log-logistic distribution. *Journal of Statistical Theory and Applications*, 11(3), pp. 265–291.

Eugene, N., Lee, C. and Famoye, F., (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and Methods*, 31(4), pp. 497–512.

Fisk, P. R., (1961). The graduation of income distribution. *Econometrica*, 29(2), pp. 171–185.

Gaire, A. K., Aryal, R., (2015). Inverse Gaussian model to describe the distribution of age-specific fertility rates of Nepal. *Journal of Institute of Science and Technology*, 20(2), pp. 80–83.

Gaire, A. K., Thapa G. B. and KC, S., (2019). *Preliminary results of Skew Log-logistic distribution, properties, and application. Proceeding of the 2nd International Conference on Earthquake Engineering and Post Disaster Reconstruction Planning*, 25–27 April 2019, Bhaktapur, Nepal, pp. 37–43.

Gaire, A. K., Thapa, G. B. and KC, S., (2022). Mathematical modeling of age-specific fertility rates of Nepali mothers. *Pakistan Journal of Statistics and Operation Research*, 18(2), pp. 417–426. https://doi.org/10.18187/pjsor.v18i2.3319.

Gaire, A. K., (2022). Skew Lomax distribution, parameter estimation, its properties, and applications. *Journal of Science and Engineering*, 10, pp. 1–11.

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

Gradshteyn, I. S., Ryzhik, I. M., (2000). *Table of integrals, series, and products*, San Diego, CA: Academic Press.

Gupta, A. K., Chang, F. C. and Huang, W. J., (2002). Some skew-symmetric models. *Random Operators and Stochastic Equations*, 10(2), pp. 133–140.

Gui, W., (2013). Marshall-Olkin extended log-logistic distribution and its application in minification processes. *Applied Mathematical Science*, 7(80), pp. 3947–3961.

Harvda, J., Charvat, F., (1967). Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika*, 3(1), pp. 30–35.

Hassan, A. S., Hemeda, S. E., (2016). The additive Weibull-G family of probability distributions. *International Journals of Mathematics and Its Applications*, 4(2), pp. 151-164.

Hemeda, S., (2018). Additive Weibull Log Logistic distribution: Properties and application. *Journal of Advanced Research in Applied Mathematics and Statistics*, 3(4), pp. 8–15.

Hamedani, G., (2013). The Zografos-Balakrishnan log-logistic distribution: Properties and applications. *Journal of Statistical Theory and Applications*, 12(3), pp. 225–244.

Jones, M., (2004). Families of distributions arising from distributions of order statistics. *Test*, 13(1), pp. 1–43.

Kleiber, C., Kotz, S., (2003). *Statistical size distributions in economics and actuarial sciences* (Vol. 470). New York: John Wiley and Sons.

Lawless, J. F., (2003). *Statistical models and methods for lifetime data*. Vol. 362. New York: John Wiley and Sons.

Lemonte, A. J., (2014). The Beta log-logistic distribution. *Brazilian Journal of Probability and Statistics*, 28(3), pp. 313–332.

Lima, S. R., Cordeiro, G. M., (2017). The extended Log-Logistic distribution: Properties and application. *Anais da Academia Brasileira de Ciências*, 89(1), pp. 3–17.

Marshall, A. W., Olkin, I., (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, 84(3), pp. 641–652.

Mazzuco, S., Scarpa, B., (2011). *Fitting an age-specific fertility rate by Skew-symmetrical probability density function, the University of Padova, Working paper Series*, Italy, 10, pp. 1–18.

Mishra, R., Singh, K. K. and Singh, A., (2017). A model for age-specific fertility rate pattern of India using skew-logistic distribution function. *American Journal of Theoretical and Applied Statistics*, 6(1), pp. 32–37.

Nadarajah, S., (2009). The skew logistic distribution. *Advances in Statistical Analysis*, 93(2), pp. 187–203.

NDHS, (2022). *Nepal Demographic and Health Survey 2022: Key Indicators Report.* Kathmandu, Nepal: Ministry of Health and Population; New ERA; and ICF., Nepal.

Peristera, P., Kostaki, A., (2007). Modeling fertility in modern populations. *Demographic Research*, 16, pp. 141–194.

Renyi, A., (1961). *On measures of entropy and information, Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, 1, pp. 547–561, Barkeley: The University of California Press.

Rodrıguez, G., (2010). *Parametric survival models*. New Jersey: Rapport technique, Princeton University.

Rosaiah, K., Nagarjuna, K. M., Kumar, D. C. U. S. and Rao, B. S., (2014). Exponential-log-logistic additive failure rate model. *Int J Sci Res Publ.*, 4(3), pp. 1–5.

Shaw, W. T., Buckley, I. R., (2007). *The alchemy of probability distributions: Beyond Gram-Charlier expansions, and a skew-kurtotic-normal distribution from a rank transmutation map*. arXiv preprint arXiv:0901.0434.

Surendran, S., Tota-Maharaj, K., (2015). *Log logistic distribution to model water demand data*. Procedia Engineering, 119, pp. 798–802.

Tadikamalla, P. R., (1980). A look at the Burr and related distributions. *International Statistical Review/Revue Internationale de Statistique*, pp. 337–344.

Tahir, M. H., Mansoor, M., Zubair, M. and Hamedani, G., (2014). McDonald log-logistic distribution with an application to breast cancer data. *Journal of Statistical Theory and Applications*, 13(1), pp. 65–82.

Tsallis, C., (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52, pp. 479–487.

Yilmaz, V., Erişoğlu, M. and Çelik, H. E., (2011). *Probabilistic prediction of the next earthquake in the NAFZ (North Anatolian Fault Zone)*, Turkey: Doğuş Üniversitesi Dergisi, 5(2), pp. 243–250.

Zografos, K., Balakrishnan, N., (2009). On families of Beta-and generalized Gamma-generated distributions and associated inference. *Statistical Methodology*, 6(4), pp. 344–362.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 63–76,* https://doi.org/10.59170/stattrans-2024-004
Received – 01.09.2021; accepted – 12.07.2023

# A chain ratio-type exponential estimator for population mean in double sampling

**Nirupama Sahoo**[1], **Sananda Kumar Jhankar**[2]

## Abstract

In this paper we have proposed an efficient ratio-type exponential estimator for estimating the population mean of the study variable, by incorporating two auxiliary variables in two-phase (double) sampling. The bias and the mean square error of the proposed estimator have been obtained up to the first order of approximation. The newly proposed estimator offers more precision in comparison to other competing estimators, theoretically as well as empirically, by considering a known value of some population parameter.

**Key words:** two-phase sampling, auxiliary variables, study variable, bias, mean square error, percent relative efficiency.

## 1. Introduction

Consider a finite population $U = (U_1, U_2, \dots\dots\dots\dots\dots\dots U_N)$ of N units. Let $\bar{X}, \bar{Y}$ and $\bar{Z}$ denote the population mean, $C_x$, $C_y$ and $C_z$ denote the coefficient of variation, $\rho_{yx}$, $\rho_{yz}$ and $\rho_{xz}$ denote the correlation coefficient. Let Y be the study variable and X and Z be the auxiliary variables with corresponding value $y_i, x_i, z_i$ $(i = 1,2, \dots\dots\dots N)$. The problem is to estimate $\bar{Y}$ in the presence of two auxiliary variable x and z.

Let $S_y^2 = \sum_{i=1}^{n}(y_i - \bar{Y})^2/(N-1)$ and $S_x^2 = \sum_{i=1}^{n}(x_i - \bar{X})^2/(N-1) S_z^2 = \sum_{i=1}^{n}(z_i - \bar{Z})^2/(N-1)$ and let $C_y = S_y/\bar{Y}$ and $C_x = S_x/\bar{X}\, C_z = S_z/\bar{Z}$ be the coefficients of variation of $y$, $x$ and $z$ respectively. $f_1 = \left(\frac{1}{n} - \frac{1}{N}\right) = \left(\frac{(1-f)}{n}\right)$, $f_2 = \left(\frac{1}{n'} - \frac{1}{N}\right) = \left(\frac{(1-f')}{n'}\right)$, $f_3 = f_1 - f_2 = \left(\frac{(1-f'')}{n}\right)$

where $f = \frac{n}{N}$, $f' = \frac{n'}{N}$ and $f'' = \frac{n}{n'}$

$$v(\bar{y}) = f_1 \bar{Y}^2 C_y^2$$

[1] Gangadhar Meher University, Amruta Vihar, Sambalpur, Odisha, India. E-mail: nirustatistics@gmail.com. ORCID: https://orcid.org/0000-0002-3244-3557.

[2] Gangadhar Meher University, Amruta Vihar, Sambalpur, Odisha, India. E-mail: sanandajhankar10@gmail.com.

## 2. Estimator with single auxiliary variable

When the population mean of the auxiliary variable $x$ is not known, Sukhatme (1962) defined the two-phase sampling ratio estimator for population mean $\bar{Y}$ as

$$t_1 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right) \tag{1}$$

$$MSE(t_1) = \bar{Y}^2\left[f_1 C_y^2 + f_3\left(C_x^2 - 2\rho_{yx}C_yC_x\right)\right] \tag{2}$$

The usual regression estimator in two-phase sampling is defined as

$$t_2 = \bar{y} + b_{yx}(n)(\bar{x}' - \bar{x}) \tag{3}$$

$$MSE(t_2) = \bar{Y}^2 C_y^2\left[f_1\left(1 - \rho_{yx}^2\right) + f_2\rho_{yx}^2\right] \tag{4}$$

Singh and Vishwakarma (2007) suggested exponential ratio and product type estimator for $\bar{Y}$as

$$t_3 = \bar{y}exp\left(\frac{\bar{x}'-\bar{x}}{\bar{x}'+\bar{x}}\right) \tag{5}$$

$$t_4 = \bar{y}exp\left(\frac{x-\bar{x}'}{x+\bar{x}'}\right) \tag{6}$$

The MSEs of the estimators $t_3$ and $t_4$ respectively are

$$MSE(t_3) = \bar{Y}^2\left[f_1 C_y^2 + \frac{f_3}{4}\left(C_x^2 - 4\rho_{yx}C_yC_x\right)\right] \tag{7}$$

$$MSE(t_4) = \bar{Y}^2\left[f_1 C_y^2 + \frac{f_3}{4}\left(C_x^2 + 4\rho_{yx}C_yC_x\right)\right] \tag{8}$$

## 3. Estimator with two auxiliary variables

Chand (1975) suggested a chain ratio-type estimator for the population mean $\bar{Y}$ defined as

$$t_5 = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)\left(\frac{\bar{Z}}{\bar{z}'}\right);\ \bar{x} \neq 0\ , \bar{z} \neq 0 \tag{9}$$

$$MSE(t_5) = \bar{Y}^2\left[f_1 C_y^2 + f_3\left(C_x^2 - 2\rho_{yx}C_yC_x\right) + f_2\left(C_z^2 - 2\rho_{yz}C_yC_z\right)\right] \tag{10}$$

Kiregyera (1980, 1984) suggested some modification of Chand (1975) and proposed ratio to regression estimator of population mean given as

$$t_6 = \bar{y} + b_{yx}(n)\left(\frac{\bar{x}'}{\bar{z}'}\bar{Z} - \bar{x}\right) \tag{11}$$

where $b_{yx}(n)$: Sample regresssion coefficient of $y$ on $x$ based on $s$ (Sub-sample)

$\rho_{yx}, \rho_{yz}, \rho_{xz}$ : Population correlation coefficient between the variables.

The MSE of estimator $t_6$ is as follows:

$$MSE(t_6) = \bar{Y}^2 C_y^2\left[f_1\{1-\rho_{yx}^2\} + f_2\left\{\rho_{yx}^2 + \rho_{yx}^2\frac{C_z^2}{C_x^2} - 2\rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right\}\right] \tag{12}$$

Singh and Khalid (2015) suggested the following estimator:

$$t_7 = \bar{y}exp\left(\frac{\bar{x}'\left(\frac{\bar{z}^*}{\bar{Z}}\right)-\bar{x}}{\bar{x}'\left(\frac{\bar{z}^*}{\bar{Z}}\right)+\bar{x}}\right) \tag{13}$$

where

$$\bar{z}^* = \frac{(N\bar{Z}-n'\bar{z}')}{(N-n')} \text{ and } k = \frac{n'}{(N-n')}$$

The required mean square error of the estimator $t_7$ is

$$MSE(t_7) = \bar{Y}^2\left[f_1 C_y^2 + f_2\left(\frac{k^2}{4}C_z^2 - k\rho_{yz}C_yC_z\right) + \frac{f_3}{4}\left(C_x^2 - 4\rho_{yx}C_yC_x\right)\right] \tag{14}$$

Singh and Choudhury (2012) developed the following exponential chain-type ratio estimators of $\bar{Y}$ under double sampling as

$$t_8 = \bar{y}exp\left(\frac{\left(\frac{\bar{x}'}{\bar{z}'}\right)\bar{Z}-\bar{x}}{\left(\frac{\bar{x}'}{\bar{z}'}\right)\bar{Z}+\bar{x}}\right) \tag{15}$$

The MSE of the estimator $t_8$ is

$$MSE(t_8) = \bar{Y}^2\left[f_1 C_y^2 + \frac{1}{4}(f_3 C_x^2 + f_2 C_z^2) - \left(f_3\rho_{yx}C_yC_x + f_2\rho_{yz}C_yC_z\right)\right] \tag{16}$$

Motivated by Singh and Vishwakarma (2007), Yadav, Singh and Chatterjee (2013) suggested a class of chain ratio exponential type estimator for population mean $\bar{Y}$ using information on two auxiliary variables x and z in two phase or double sampling as

$$t_9 = \bar{y} exp\left[\frac{\hat{\bar{X}}_{rd}-\bar{x}}{\hat{\bar{X}}_{rd}+\bar{x}}\right] \tag{17}$$

where

$$\hat{\bar{X}}_{rd} = \frac{\bar{x}'}{(a\bar{z}'+b)}(a\bar{Z}+b)$$

Thus, the above estimator can be expressed as

$$t_9 = \bar{y} exp\left[\frac{\frac{\bar{x}'}{(a\bar{z}'+b)}(a\bar{Z}+b)-\bar{x}}{\frac{\bar{x}'}{(a\bar{z}'+b)}(a\bar{Z}+b)+\bar{x}}\right] \tag{18}$$

The mean square error of the estimator $t_9$ is

$$MSE\ (t_9) = \bar{Y}^2\left[f_1 C_y^2 + \frac{f_3}{4}\left(C_x^2 - 4\rho_{yx}C_yC_x\right) - f_2\rho_{yz}^2 C_y^2\right] \tag{19}$$

## 4. The suggested class of estimator

Following the previously discussed estimation procedures for two-phase sampling, we have proposed an efficient ratio-type exponential estimator

$$t_{10} = \bar{y} exp\left[k_1\left\{\frac{\bar{x}'-\bar{x}}{\bar{x}'+\bar{x}}\right\} + k_2\left\{\frac{\bar{z}'-\bar{Z}}{\bar{z}'+\bar{Z}}\right\}\right] \tag{20}$$

Where $k_1$ and $k_2$ are unknown constants. The values of $k_1$ and $k_2$ can be determined by the principle of optimality conditions.

To obtain the mean square error of the proposed estimator $t_{10}$, we consider

$$\bar{y} = \bar{Y}(1+e_0), \bar{x} = \bar{X}(1+e_1)\ ,\quad \bar{x}' = \bar{X}(1+e_2)\ ,\ \bar{z}' = \bar{Z}(1+e_3)$$

Such that

$$E(e_0) = E(e_1) = E(e_2) = E(e_3) = 0$$

$$E(e_0^2) = \left(\frac{1-f}{n}\right) C_y^2 \ , \ E(e_1^2) = \left(\frac{1-f}{n}\right) C_x^2$$

$$E(e_2^2) = \left(\frac{1-f'}{n'}\right) C_x^2, \ E(e_3^2) = \left(\frac{1-f'}{n'}\right) C_z^2$$

$$E(e_0e_1) = \left(\frac{1-f}{n}\right) \rho_{yx} C_y C_x, \ E(e_0e_2) = \left(\frac{1-f'}{n'}\right) \rho_{yx} C_y C_x$$

$$E(e_0e_3) = \left(\frac{1-f'}{n'}\right) \rho_{yz} C_y C_z, \ E(e_1e_2) = \left(\frac{1-f'}{n'}\right) C_x^2$$

$$E(e_1e_3) = \left(\frac{1-f'}{n'}\right) \rho_{xz} C_x C_z, \ E(e_2e_3) = \left(\frac{1-f'}{n'}\right) \rho_{xz} C_x C_z$$

The mean square error of the proposed estimator $t_{10}$ is as follows:

$$t_{10} = \bar{y} exp\left[k_1 \left\{\frac{e_2 - e_1}{2 + e_1 + e_2}\right\} + k_2 \left\{\frac{e_3}{2 + e_3}\right\}\right]$$
$$= \bar{y} exp\left[k_1 \{e_2 - e_1\} \frac{1}{2} \left(1 - \frac{e_1}{2} - \frac{e_2}{2} + \frac{e_1^2}{4} + \frac{e_2^2}{4}\right) + k_2 \left\{e_3 \frac{1}{2} \left(1 - \frac{e_3}{2} + \frac{e_3^2}{4}\right)\right\}\right]$$
$$= \bar{y} exp\left[k_1 \left\{\frac{e_2}{2} - \frac{e_1 e_2}{4} - \frac{e_2^2}{4} - \frac{e_1}{2} + \frac{e_1^2}{4} + \frac{e_1 e_2}{4}\right\} + k_2 \left\{\frac{e_3}{2} - \frac{e_3^2}{4}\right\}\right]$$

$$= \bar{Y}(1 + e_0) \left[1 + k_1 \left\{\frac{e_2}{2} - \frac{e_1 e_2}{4} - \frac{e_2^2}{4} - \frac{e_1}{2} + \frac{e_1^2}{4} + \frac{e_1 e_2}{4}\right\} + k_2 \left\{\frac{e_3}{2} - \frac{e_3^2}{4}\right\} + \left\{k_1 \left(\frac{e_2}{2} - \frac{e_1 e_2}{4} - \frac{e_2^2}{4} - \frac{e_1}{2} + \frac{e_1^2}{4} + \frac{e_1 e_2}{4}\right)\right\}^2 + \left\{k_2 \left(\frac{e_3}{2} - \frac{e_3^2}{4}\right)\right\}^2\right]$$

Neglecting the term having power greater than two, we have

$$t_{10} = \bar{Y} \left[1 + k_1 \frac{e_2}{2} - k_1 \frac{e_2^2}{4} - k_1 \frac{e_1}{2} + k_1 \frac{e_1^2}{4} + k_2 \frac{e_3}{2} - k_2 \frac{e_3^2}{4} + k_1^2 \frac{e_2^2}{8} + k_1^2 \frac{e_1^2}{8} + k_2^2 \frac{e_3^2}{8} + e_0 + k_1 \frac{e_0 e_2}{2} - k_1 \frac{e_0 e_1}{2} + k_2 \frac{e_0 e_3}{2}\right]$$

or

$$t_{10} - \bar{Y} = \bar{Y} \left[k_1 \frac{e_2}{2} - k_1 \frac{e_2^2}{4} - k_1 \frac{e_1}{2} + k_1 \frac{e_1^2}{4} + k_2 \frac{e_3}{2} - k_2 \frac{e_3^2}{4} + k_1^2 \frac{e_2^2}{8} + k_1^2 \frac{e_1^2}{8} + k_2^2 \frac{e_3^2}{8} + e_0 + k_1 \frac{e_0 e_2}{2} - k_1 \frac{e_0 e_1}{2} + k_2 \frac{e_0 e_3}{2}\right] \quad (21)$$

The bias of the estimator $t_{10}$ can be obtained by taking expectation on both sides of equation (21) and is given by

$$\boldsymbol{B(t_{10}) = \bar{Y}^2 \left[k_1 \frac{e_1^2}{4} - k_1 \frac{e_2^2}{4} - k_2 \frac{e_3^2}{4} + k_1^2 \frac{e_2^2}{8} + k_1^2 \frac{e_1^2}{8} + k_2^2 \frac{e_3^2}{8} + k_1 \frac{e_0 e_2}{2} - k_1 \frac{e_0 e_1}{2} + k_2 \frac{e_0 e_3}{2}\right]} \quad (22)$$

By squaring and taking expectation on both sides of equation (21), we get the mean square error of $t_{10}$ to the first degree of approximation.

$$E(t_{10}-\bar{Y})^2 = \bar{Y}^2 E\left[k_1^2\frac{e_2^2}{4} + k_1^2\frac{e_1^2}{4} + k_2^2\frac{e_3^2}{4} + e_0^2 - k_1^2\frac{e_1e_2}{2} + k_1k_2\frac{e_2e_3}{2} + \quad k_1e_0e_2 - k_1k_2\frac{e_1e_3}{2} - k_1e_0e_1 + \ k_2e_0e_3\right]$$

$$= \bar{Y}^2\left[\left(\frac{1-f}{n}\right)C_y^2 + \frac{k_1^2}{4}\left(\frac{1-f''}{n}\right)C_x^2 + \frac{k_2^2}{4}\left(\frac{1-f'}{n'}\right)C_z^2 - k_1\left(\frac{1-f''}{n}\right)\rho_{yx}C_yC_x + \quad k_2\left(\frac{1-f'}{n'}\right)\rho_{yz}C_yC_z\right]$$

$$MSE(t_{10}) = \bar{Y}^2\left[f_1C_y^2 + \frac{k_1^2}{4}f_3C_x^2 + \frac{k_2^2}{4}f_2C_z^2 - k_1f_3\rho_{yx}C_yC_x + k_2f_2\rho_{yz}C_yC_z\right] \tag{23}$$

Now, we have to find out the optimum values of $k_1$ and $k_2$

$$\frac{\partial MSE(t_{10})}{\partial k_1} = 0$$
$$\Rightarrow k_{1_{opt}} = \frac{2\rho_{yx}C_yC_x}{C_x^2}$$
$$\Rightarrow k_{1_{opt}} = 2\rho_{yx}\left(\frac{C_y}{C_x}\right)$$

Now, for $k_2$

$$\frac{\partial\, MSE(t_{10})}{\partial k_2} = 0$$
$$\Rightarrow k_{2_{opt}} = \frac{2\rho_{yz}C_yC_z}{C_z^2}$$
$$\Rightarrow k_{2_{opt}} = -2\rho_{yz}\left(\frac{C_y}{C_z}\right)$$

By substituting above $k_{1_{opt}}$ and $k_{2_{opt}}$ in equation (23) we obtained the minimum mean square error of the estimator $t_{10}$ as

$$MSE(t_{10})_{opt} = \bar{Y}^2\left[f_1C_y^2 + \left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right)f_3\rho_{yx}C_yC_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right)f_2\rho_{yz}C_yC_z\right] \tag{24}$$

## 5. Theoretical Comparison

The proposed estimator $t_{10}$ under its optimality condition is more efficient than the existing estimators $t_i(i = 1,2 \ldots \ldots 9)$ if and only if the following conditions hold.

(i) $MSE(t_1) - MSE(t_{10})_{opt} > 0$

$$\left[\left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < \left[f_3\left(C_x^2 - 2\rho_{yx} C_y C_x\right)\right] \quad (25)$$

(ii) $MSE(t_2) - MSE(t_{10})_{opt} > 0$

$$\left[f_1 C_y^2 + \left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < C_y^2\left[f_1\left(1 - \rho_{yx}^2\right) + f_2 \rho_{yx}^2\right] \quad (26)$$

(iii) $MSE(t_3) - MSE(t_{10})_{opt} > 0$

$$\left[\left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < \left[\frac{f_3}{4}\left(C_x^2 - 4\rho_{yx} C_y C_x\right)\right] \quad (27)$$

(iv) $MSE(t_4) - MSE(t_{10})_{opt} > 0$

$$\left[\left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < \left[\frac{f_3}{4}\left(C_x^2 + 4\rho_{yx} C_y C_x\right)\right] \quad (28)$$

(v) $MSE(t_5) - MSE(t_{10})_{opt} > 0$

$$\left[\left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < \left[f_3\left(C_x^2 - 2\rho_{yx} C_y C_x\right) + f_2\left(C_z^2 - 2\rho_{yz} C_y C_z\right)\right] \quad (29)$$

(vi) $MSE(t_6) - MSE(t_{10})_{opt} > 0$

$$\left[f_1 C_y^2 + \left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < C_y^2\left[f_1\{1 - \rho_{yx}^2\} + f_2\left\{\rho_{yx}^2 + \rho_{yx}^2 \frac{C_z^2}{C_x^2} - 2\rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right\}\right] \quad (30)$$

(vii) $MSE(t_7) - MSE(t_{10})_{opt} > 0$

$$\left[\left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < \left[f_2\left(\frac{k^2}{4} C_z^2 - k\rho_{yz} C_y C_z\right) + \frac{f_3}{4}\left(C_x^2 - 4\rho_{yx} C_y C_x\right)\right] \quad (31)$$

(viii) $MSE(t_8) - MSE(t_{10})_{opt} > 0$

$$\left[\left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < \left[\frac{1}{4}\left(f_3 C_x^2 + f_2 C_z^2\right) - \left(f_3 \rho_{yx} C_y C_x + f_2 \rho_{yz} C_y C_z\right)\right] \quad (32)$$

(ix) $MSE(t_9) - MSE(t_{10})_{opt} > 0$

$$\left[\left(\rho_{yx}\left(\frac{C_y}{C_x}\right)\right)^2 f_3 C_x^2 + \left(\rho_{yz}\left(\frac{C_y}{C_z}\right)\right)^2 f_2 C_z^2 - 2\rho_{yx}\left(\frac{C_y}{C_x}\right) f_3 \rho_{yx} C_y C_x - 2\rho_{yz}\left(\frac{C_y}{C_z}\right) f_2 \rho_{yz} C_y C_z\right] < \left[\frac{f_3}{4}\left(C_x^2 - 4\rho_{yx} C_y C_x\right) - f_2 \rho_{yz}^2 C_y^2\right] \quad (33)$$

## 6. Empirical Study

To judge the superiority of our newly proposed estimator over the competing estimators we have taken the following numerical values of different population parameters from two different population data sets.

Population-1 [Source: Singh (1967)]

The variables are

$y$: Number of females employed

$x$: Number of females in services

$z$: Number of educated females

$N = 61 n' = 20 n = 10 \bar{Y} = 7.46$

$\bar{X} = 5.31 \bar{Z} = 179.00 C_y = 0.7103 C_x = 0.7587$

$C_z = 0.2515 \rho_{yx} = 0.7737 \rho_{yz} = -0.2070 \rho_{xz} = -0.0033$

Population-2 [Source: Murthy (1967) pp. 226]

The variables are

$y$: Output

$x$: Number of workers

$z$: Fixed capital

$N = 61 n' = 20 n = 10 \bar{Y} = 7.46$

$\bar{X} = 5.31 \bar{Z} = 179.00 C_y = 0.7103 C_x = 0.7587$

$C_z = 0.2515 \rho_{yx} = 0.7737 \rho_{yz} = -0.2070 \rho_{xz} = -0.0033$

To compare the efficiency of the proposed estimator and the considered existing estimators, we have computed the percent relative efficiencies (PREs).

The formula for calculating the percent relative efficiency is given by

$$PRE(t_i) = \left[\frac{var(\overline{y})}{MSE(t_i)}\right] \times 100 \; ; \tag{34}$$

where $i = 1,2,3,4,5,6,7,8,9,10$

Findings are given in Table 1.

**Table 1:**

| Estimators | Population-1 | Population-2 |
|---|---|---|
| $t_1$ | 144.1214 | 39.1380 |
| $t_2$ | 155.8702 | 233.9572 |
| $t_3$ | 147.6438 | 180.6875 |
| $t_4$ | 60.1505 | 25.6294 |
| $t_5$ | 124.5640 | 36.8749 |
| $t_6$ | 140.9066 | 607.6388 |
| $t_7$ | 144.1214 | 285.2960 |
| $t_8$ | 139.3606 | 361.375 |
| $t_9$ | 151.3525 | 373.8362 |
| $\boldsymbol{t_{10}}$ | **159.9441** | **677.5781** |

## 7. Conclusion

On account of the percent relative efficiencies of the estimators as shown in Table 1, it has been observed that the performance of the proposed estimator is better than the other existing estimators. Hence, it is recommended for use in practice.

## Acknowledgements

We are very grateful to the referees and members of the board of editors for their useful suggestions.

## References

Anderson, T. W., (1958). An Introduction to Multivariate Statistical Analysis. *John Wileye& Sons*, Inc., New York.

Bahl, S., Tuteja, R. K., (1991). Ratio and product type exponential estimators, *Journal of information and optimization sciences*, 12(1), pp. 159–164.

Bandyopadhyay, A., Singh, G. N., (2014). Some modified classes of chain type estimators of population mean in two phase sampling, *International Journal of Statistics and Economics*, 14(2), pp. 92–103.

Chakraborty, R. P., (1968). Contribution to the theory of ratio-type Estimators, Ph.D. Thesis, Texas A and M University, U.S.A.

Chand, L., (1975). Some ratio-type estimators based on two or more Auxiliary variables, Ph.D. diss., Iowa State University, Ames, Iowa.

Cochran, W. G., (1977). Sampling Techniques, 3rd edition, John Wiley, NewYork.

Dash, P. R., Mishra, G., (2011). An improved class of estimators in two-phase sampling using two auxiliary variables, *Communications in Statistics-Theory and Methods*, 40(24), pp. 4347–4352.

Fisher, R. A., (1936). The use of multiple measurements in taxonomic problems, *Annals of eugenics*, 7(2), pp. 179–188.

Gupta, P. C., (1978). On some quadratic and higher degree ratio and product estimators, *Journal of Indian Society of Agricultural Statistics*, 30, No. 2, pp. 71–80.

Kadilar, C., Cingi, H., (2006). Ratio estimators for the population variance in simple and stratified random sampling, *Appl. Math Computation* 173(2), pp.1047–1059.

Kiregyera, B., (1980). A chain ratio-type estimator in finite population double samplingusing two auxiliary variables, *Metrika*, 27, pp.217–223.

Kiregyera, B., (1984). Regression-type estimators using two auxiliary variables andthe model of double sampling from finite populations, *Metrika*, 31, pp. 215–226.

Kumar, K., Kumar, M., (2017). Two Phase Sampling Exponential Type Estimators for Ratio and Product of two Population Means in the Presence of Non-response, *International Journal of Scientific Research in Mathematical and Statistical Sciences*, Vol. 4, Issue.6, pp. 26–34.

Mukherjee, R., Rao, T. J., and Vijayan, K., (1987). Regression type estimators using multiple auxiliary information, *Australian Journal of Statistics*, 29(3), pp. 244–254

Murthy, M. N., (1967). Sampling Theory and Methods, Statistical Publishing Society, Calcutta, India.

Panda, K. B., Sahoo, N., (2017). Estimation of Finite Population Mean Using Variable Transformation, *Int. J. Agricult. Stat. Sci*., Vol. 13, No. 1, pp. 141–144.

Prasad, B., Singh, R. S., and Singh, H. P., (1996). Some chain ratio type estimators for ratio of two population means using two auxiliary characters in two phase sampling, *Metron, LIV*, No. 1-2, pp. 95–113.

Reddy, V. N., (1973). On ratio and product methods of estimation, *Sankhyā: The Indian Journal of Statistics*, Series B, pp. 307–316.

Reddy, V. N., (1978). A study on the use of prior knowledge on certain population parameters in estimation, *Sankhya C*, 40, pp. 29–37.

Singh, B. K., Choudhury, S., (2012). Exponential chain ratio and product type estimators for finite population mean under double sampling scheme. *Journal of Science Frontier Research in Mathematics and Design Sciences*, 12(6), pp. 0975–5896.

Singh, G. N., (2001). On the use of transformed auxiliary variable in the estimation of population mean in two-phase sampling, *Statistics in Transition*, 5(3), pp. 405–416.

Singh, G. N., Khalid, M., (2015). Exponential chain dual to ratio and regression type estimators of population mean in two-phase sampling, *Statistica*, 75(4), pp. 379–389.

Singh, M.P., (1967). Ratio-cum-product method of estimation, *Metrika*, 12, pp. 34–42.

Sing, G. N., Sharma, A. K., (2014). An improved estimation procedure of population mean in two phase sampling, *Journal of International Academy of Physical Sciences*, 18(1), pp. 1–10.

Singh, G. N., Sharma, A. K., (2015). Some improved estimators for population mean in double sampling, *International Journal of Mathematics and Computation*, 26(3), pp. 36–45.

Singh, H. P., Gangele, R. K., (1999). Classes of almost unbiased ratio and product-type estimators in two phase sampling, *Statistica*, LIX, No. 1, pp. 109–124.

Singh, H. P., Tailor, R., (2005a). Estimation of finite population mean with known coefficient of variation of an auxiliary character, *Statistica*, LXV, No. 3, pp. 301–313.

Singh, H. P., Tailor, R., (2005b). Estimation of finite population mean using known correlation coefficient between auxiliary characters, *Statistica*, LXV, No. 4, pp. 407–418.

Singh, H. P., Singh, S., and Kim, J. M. (2006). General families of chain ratio type estimators of the population mean with known coefficient of variation of the second auxiliary variable in two phase sampling, *Journal of the Korean Statistical Society*, 35, No. 4, pp. 377–395.

Singh, H. P., Vishwakarma, G. K., (2007). Modified exponential ratio and product estimators for finite population mean in double sampling, *Austrian Journal of Statistics*, 36, No. 3, pp. 217–225.

Singh, H. P., Mathur, N., and Chandra, P., (2009). A chain-type estimator for population variance using two auxiliary variables in two-phase sampling. *Statistics in Transition-new series*, 10, No. 1, pp. 75–84.

Singh, H. P., Vishwakarma, G. K., (2007). Modified exponential ratio and product estimators for finite population mean in double sampling, *Austrian journal of statistics*, 36(3), pp. 217–225.

Singh, G. N., Pandey, A. K., and Jaiswal, A. K., (2019). An efficient class of chain-type exponential estimators for population mean under two phase sampling scheme, *Journal of reliability and statistical studies*, Vol. 12, Issue 1(2019), pp. 23–40.

Singh, V. K., Shukla, D., (1987). One parameter family of factor-type ratio estimators, *Metron*, 45(1-2), pp. 273–283.

Singh, V. K., Singh, G. N., (1991). Chain type regression estimators with two auxiliary variables under double sampling scheme, *Metron*, 49, pp. 279–289.

Srivenkataramana, T., Tracy, D. S., (1980). An alternative to ratio method in sample surveys, *Annals of the Institute of Statistical Mathematics*, 32(1), pp. 111–120.

Srivenkataramana, T., Tracy, D. S., (1989). Two phase sampling for selection with probability proportional to size in sample surveys, *Biometrika*, 76, pp. 818–821.

Shukla, G. K., (1966). An alternative multivariate ratio estimate for finite population, *Calcutta Statistical Association Bulletin*, 15(2–3), pp. 127–134.

Upadhyaya, L. N., Kushwaha, K. S., and Singh, H. P., (1990). A modified chain ratio type estimator in two-phase sampling using multi-auxiliary information, *Metron*, 48, pp. 381–393.

Upadhyaya, L. N., Singh, G. N., (2001). Chain type estimators using transformed auxiliary variable in two-phase sampling, *Advances in Modeling and Analysis*, 38, No. 1-2,pp. 1–10.

Upadhyaya, L. N., Singh, H. P., and Singh, S., (2004). A family of almost unbiased estimators for negatively correlated variables using Jackknife Technique, *Statistica*, LXIV, 4, pp.767–778.

Upadhyaya, L. N., Singh, H. P., and Tailor, R., (2006). Estimation of mean with known coefficient of variation of an auxiliary variable in two phase sample, *Statistics in Transition*, 7, No. 6, pp. 1327–1344.

Vos, J. W. E., (1980). Mixing of direct, ratio and product method estimators, *Statist. Neerl.*, 34, pp. 209–218.

Walsh, J. E., (1970). Generalization of ratio estimate for population total, *Sankhya*, *A*, 32, pp. 99–106.

Yadav, R., Upadhyaya L. N., Singh, H.P., and Chatterjee, S., (2013). A chain ratio exponential type estimator in two phase sampling using auxiliary information, *Statistica*, anno LXXIII, No. 2, pp. 221–234.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 77–91,* https://doi.org/10.59170/stattrans-2024-005
Received – 13.12.2021; accepted – 29.12.2023

# Improved calibration estimation of population mean in stratified sampling using two auxiliary variables

**Abimibola V. Oladugba**[1], **Oluwagbenga T. Babatunde**[2]

## Abstract

In this paper, a new improved calibration estimator for the population mean in a stratified sampling was proposed using two auxiliary variables. A simulation study was carried out to evaluate the performance and efficiency of the proposed estimator with respect to three estimators considered in the literature for estimating the population mean in a stratified sampling using two auxiliary variables. The results showed that the new estimator proved to be more efficient than the three existing estimators considered

**Key words:** calibration, estimator, stratified sampling, auxiliary variables, mean square error, bias, percentage relative efficiency.

## 1. Introduction

Calibration estimation is a popular approach in sample survey introduced by Deville and Sarndal (1992) and meant to improve the precision of the estimated population parameter. This is achieved using additional relevant information known as auxiliary information or variable. Auxiliary variable is a variable that provides some other relevant details about the study variable (Babatunde et al. (2023)). Auxiliary variables are correlated to the study variable (Babatunde et al. (2023)) and the efficiency of an estimator depends on the level of correlation between the study and auxiliary variables (Agunbiade and Ogunyinka (2013)). Agunbiade and Ogunyinka (2013) showed that using auxiliary variable that is highly correlated to the study variable produces an estimator with smaller variance compared to when the correlation level between the auxiliary variable and study variable is medium or low. This implies that the choice of auxiliary variables is restricted only to variables that are correlated to the

[1] Department of Statistics, University of Nigeria, Nsukka, Nigeria. E-mail: abimibola.oladugba@unn.edu.ng. ORCID: https://orcid.org/0000-0002-6402-8833.

[2] Corresponding Author. Department of Statistics, University of Nigeria, Nsukka, Nigeria. E-mail: oluwagbenga.babatunde@unn.edu.ng. ORCID: https://orcid.org/0000-0002-4761-9770.

study variable. This poses some limitations on the use of auxiliary variable as not all variables can be used as auxiliary variable.

In estimating the population mean of a stratified sampling using calibration approach, the calibration weights are used to replace the stratum weights in the estimator. The calibration weights are obtained by minimizing a distance function subject to well defined calibration constraints. Most often, calibration constraints restrict the sum of the selected sample statistics to be equal to the sum of the population parameters in the different strata (see Ozgul (2018), Alam et al. (2021), Adubi et al. (2022), Babatunde et al. (2023)).

Several calibration estimators for the population mean in a stratified sampling using different parameters of one auxiliary variable in the calibration constraints have been proposed in the literature (see Tracey, Singh and Arnab (2003), Rao, Tekabu and Khan (2016), Koyuncu and Kadilar (2016), Sisodia, Singh and Singh (2017), Alam, Singh and Shabbir (2019), Garg and Pachori (2019), Alam and Shabbir (2020), Babatunde et al. (2023), Oladugba et al. (2023) etc.). Several calibration estimators were arrived at by modifying existing estimators (see Kadilar and Cingi (2006), Garg and Pachori (2019)). Calibration estimators of the population mean have been shown to be more efficient than the general population estimator in a stratified sampling (see Rao, Khan and Khan (2012), Ozgul (2018) and Alam et al. (2019)). The use of two auxiliary variables have also been explored in the calibration estimation of the population mean in stratified sampling. Rao et al. (2012), Ozgul (2018) and Rai, Singh and Qasim (2021) proposed different calibration estimators for population mean in a stratified sampling using different calibration constraints based on two auxiliary variables.

In this paper, we propose a new improved calibration estimator for the population mean in a stratified sampling based on two auxiliary variables by modifying the estimator proposed in Ozgul (2018) with the aim of achieving a more efficient estimator. The standard deviation of the two auxiliary variables was used to define the calibration constraints.

The remainder of this paper is as follows: notations are presented in Section 2, some of the existing calibration estimators based on two auxiliary variables were discussed in Section 3. In Section 4, the proposed calibration estimator was presented. The simulation study carried out and conclusions are presented in Sections 5 and 6, respectively.

## 2. Notations

Consider a situation where it is desired to estimate the population mean $\overline{Y}$ in stratified sampling using additional information from two auxiliary variables. Let $M$ be a finite population consisting of $N$ units, i.e. $M = \left(M_1, M_2, ...M_N\right)$. Let $y_i$, $x_{1i}$ and

$x_{2i}$ be the $i^{th}$ value of the study variable, first auxiliary variable and second auxiliary variable respectively, i = 1, 2, … *N*. Let *M* be divided into *Z* distinct homogenous strata with each stratum containing $N_h$ units, *h* = 1, 2, …, *Z* such that $N = \sum_{h=1}^{Z} N_h$. A sample of size *n* is drawn from *M* using simple random sampling without replacement (SRSWOR) by selecting $n_h$ units from $h^{th}$ stratum such that $n = \sum_{h=1}^{Z} n_h$. The mean of the sample and population of the study variable in each stratum are given as $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$ and $\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$ respectively. The mean of the sample and population of the first auxiliary variable in each stratum are given as $\bar{x}_{1h} = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{1hi}$ and $\bar{X}_{1h} = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{1hi}$ respectively. The mean of the sample and population of the second auxiliary variable are given as $\bar{x}_{2h} = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{2hi}$ and $\bar{X}_{2h} = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{2hi}$ respectively. The sample and population standard deviation of the first auxiliary variable in each stratum are given as $s_{x_{1h}} = \sqrt{\frac{\sum_{i=1}^{n_h}\left(x_{1i} - \bar{x}_{1h}\right)^2}{n_h - 1}}$ and $S_{x_{1h}} = \sqrt{\frac{\sum_{i=1}^{N_h}\left(x_{1i} - \bar{X}_{1h}\right)^2}{N_h - 1}}$ respectively. The sample and population standard deviation of the second auxiliary variable in each stratum are given as $s_{x_{2h}} = \sqrt{\frac{\sum_{i=1}^{n_h}\left(x_{2i} - \bar{x}_{2h}\right)^2}{n_h - 1}}$ and $S_{x_{2h}} = \sqrt{\frac{\sum_{i=1}^{N_h}\left(x_{2i} - \bar{X}_{2h}\right)^2}{N_h - 1}}$ respectively. The population mean of the auxiliary variables are $\bar{X}_1 = \frac{1}{N}\sum_{i=1}^{N} x_{1i}$ and $\bar{X}_2 = \frac{1}{N}\sum_{i=1}^{N} x_{2i}$ respectively.

The population mean of a stratified sampling is estimated by:

$$\bar{y}_{st} = \sum_{h=1}^{Z} W_h \bar{y}_h \tag{2.1}$$

where $W_h = \frac{N_h}{N}$ is the $h^{th}$ stratum weights.

The precision of the estimator in (2.1) is improved upon using the calibration approach which replaces the stratum weights $W_h$ with calibrated weights obtained by optimizing the Chi-square distance function defined below:

$$D\left(\Omega_h, W_h\right) = \sum_{h=1}^{Z} \frac{\left(\Omega_h - W_h\right)^2}{W_h Q_h} \tag{2.2}$$

Subject to well defined calibration constraints.
where $\Omega_h$ are the calibrated weights and $Q_h$ are defined weights for obtaining different versions of the estimator (Alam and Shabbir (2020)).

## 3. Some Calibration Estimators in Stratified Sampling Using Two Auxiliary Variables

Different calibration estimators have been proposed for the population mean of a stratified sampling using several known parameters of the auxiliary variables. Some of the existing calibration estimators using two auxiliary variables are reviewed below.

### 3.1. Rao et al. (2012)

Rao et al. (2012) proposed a calibration estimator with two auxiliary variables using the mean of the auxiliary variables in the calibration constraints as:

$$\bar{y}_R = \sum_{h=1}^{Z} \Omega_{hR} \bar{y}_h \tag{3.1}$$

where the calibrated weights $\Omega_{hR}$ are obtained by minimizing the Chi-square distance function in (2.2) subject to the calibration constraints given by:

$$\sum_{h=1}^{Z} \Omega_{hR} \bar{x}_{1h} = \bar{X}_1 \tag{3.2}$$

$$\sum_{h=1}^{Z} \Omega_{hR} \bar{x}_{2h} = \bar{X}_2 \tag{3.3}$$

By minimizing the function in (2.2) subject to (3.1) and (3.2), the optimum weights obtained are:

$$\Omega_{hR} = W_h + W_h Q_h (\lambda_1 \bar{x}_{1h} + \lambda_2 \bar{x}_{2h}) \tag{3.4}$$

where

$$\lambda_1 = \frac{-\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2 \bar{x}_{1h} \left( \bar{X}_1 - \hat{\bar{X}}_1 \right) + \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2 \bar{x}_{2h} \left( \bar{X}_2 - \hat{\bar{X}}_2 \right)}{\left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h} \right)^2 - \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2 \right) \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2 \right)} \tag{3.5}$$

$$\lambda_2 = \frac{\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h}^2 \left( \bar{X}_1 - \hat{\bar{X}}_1 \right) - \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2 \bar{x}_{2h} \left( \bar{X}_2 - \hat{\bar{X}}_2 \right)}{\left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h} \right)^2 - \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2 \right) \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2 \right)} \tag{3.6}$$

$$\hat{\bar{X}}_1 = \sum_{h=1}^{Z} W_h \bar{x}_{1h} \text{ and } \hat{\bar{X}}_2 = \sum_{h=1}^{Z} W_h \bar{x}_{2h}$$

By substituting (3.5) and (3.6) into (3.4) and substituting (3.4) into (3.1), the obtained estimator can be expressed as:

$$\bar{y}_R = \sum_{h=1}^{Z} W_h \bar{y}_h + \hat{\gamma}_1 \left( \bar{X}_1 - \hat{\bar{X}}_1 \right) + \hat{\gamma}_2 \left( \bar{X}_2 - \hat{\bar{X}}_2 \right) \tag{3.7}$$

where

$$\hat{\gamma}_1 = \frac{\left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h} \bar{y}_h \right) \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h} \right) - \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{y}_h \right) \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2 \right)}{\left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h} \right)^2 - \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2 \right) \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2 \right)} \tag{3.8}$$

$$\hat{\gamma}_2 = \frac{\left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{y}_h \right) \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h} \right) - \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h} \bar{y}_h \right) \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2 \right)}{\left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h} \right)^2 - \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2 \right) \left( \sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2 \right)} \tag{3.9}$$

### 3.2. Ozgul (2018)

Ozgul (2018) proposed a calibration estimator with two auxiliary variables using the ratio of the mean of the auxiliary variables in the calibration constraints as:

$$\bar{y}_O = \sum_{h=1}^{Z} \Omega_{hO} \bar{y}_h \tag{3.10}$$

where the calibrated weights $\Omega_{hO}$ are obtained by minimizing the Chi-square distance function in (2.2) subject to the calibration constraints given by:

$$\sum_{h=1}^{Z} \Omega_{hO} = \sum_{h=1}^{Z} W_h \tag{3.11}$$

$$\sum_{h=1}^{Z} \Omega_{hO} \hat{R}_h = \sum_{h=1}^{Z} W_h R_h \tag{3.12}$$

where $\hat{R}_h = \sum_{h=1}^{Z} \frac{\bar{x}_{1h}}{\bar{x}_{2h}}$ and $R_h = \sum_{h=1}^{Z} \frac{\bar{X}_{1h}}{\bar{X}_{2h}}$

By minimizing the function in (2.2) subject to (3.11) and (3.12), the optimum weights obtained are:

$$\Omega_{hO} = W_h + W_h Q_h \left( \frac{\Delta_1 + \Delta_2 \hat{R}_h}{A} \right) \tag{3.13}$$

where

$$\Delta_1 = -\sum_{h=1}^{Z} W_h \left( R_h - \hat{R}_h \right) \left( \sum_{h=1}^{Z} W_h Q_h \hat{R}_h \right) \tag{3.14}$$

$$\Delta_2 = \sum_{h=1}^{Z} W_h\left(R_h - \hat{R}_h\right)\left(\sum_{h=1}^{Z} W_h Q_h\right) \tag{3.15}$$

$$A = \left(\sum_{h=1}^{Z} W_h Q_h\right)\left(\sum_{h=1}^{Z} W_h Q_h \hat{R}_h^2\right) - \left(\sum_{h=1}^{Z} W_h Q_h \hat{R}_h\right)^2 \tag{3.16}$$

By substituting (3.14) and (3.15) into (3.13) and substituting (3.13) into (3.12), the obtained estimator can be expressed as:

$$\bar{y}_O = \sum_{h=1}^{Z} W_h \bar{y}_h + \hat{\gamma}\sum_{h=1}^{Z} W_h\left(R_h - \hat{R}_h\right) \tag{3.17}$$

where

$$\hat{\gamma} = \left(\frac{\left(\sum_{h=1}^{Z} W_h Q_h \hat{R}_h \bar{y}_h\right)\left(\sum_{h=1}^{Z} W_h Q_h\right) - \left(\sum_{h=1}^{Z} W_h Q_h \bar{y}_h\right)\left(\sum_{h=1}^{Z} W_h Q_h \hat{R}_h\right)}{\left(\sum_{h=1}^{Z} W_h Q_h\right)\left(\sum_{h=1}^{Z} W_h Q_h \hat{R}_h^2\right) - \left(\sum_{h=1}^{Z} W_h Q_h \hat{R}_h\right)^2}\right) \tag{3.18}$$

### 3.3. Rai et al. (2021)

Rai et al. (2021) proposed a calibration estimator with two auxiliary variables using the sample and population mean of the auxiliary variables in the calibration constraints as:

$$\bar{y}_C = \sum_{h=1}^{Z} \Omega_{hC} \bar{y}_h \tag{3.19}$$

where the calibrated weights $\Omega_{hC}$ are obtained by minimizing the Chi-square distance function in (2.2) subject to the calibration constraints given by:

$$\sum_{h=1}^{Z} \Omega_{hC} \bar{x}_{1h} = \sum_{h=1}^{Z} W_h \bar{X}_{1h} \tag{3.20}$$

$$\sum_{h=1}^{Z} \Omega_{hC} \bar{x}_{2h} = \sum_{h=1}^{Z} W_h \bar{X}_{2h} \tag{3.21}$$

By minimizing the function in (2.2) subject to (3.20) and (3.21), the optimum weights obtained are:

$$\Omega_{hC} = W_h\left(1 + \lambda_1 Q_h \bar{x}_{1h} + \lambda_2 Q_h \bar{x}_{2h}\right) \tag{3.22}$$

where

$$\lambda_1 = \frac{\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2 \hat{\bar{X}}_1 - \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h} \hat{\bar{X}}_2}{\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2\right)\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2\right) - \left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h}\right)^2} \tag{3.23}$$

$$\lambda_2 = \frac{-\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h} \hat{\bar{X}}_1 - \sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2 \hat{\bar{X}}_2}{\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2\right)\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2\right) - \left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h}\right)^2} \tag{3.24}$$

where

$$\hat{\bar{X}}_1 = \sum_{h=1}^{Z} W_h \bar{X}_{1h} - \sum_{h=1}^{Z} W_h \bar{x}_{1h} \tag{3.25}$$

$$\hat{\bar{X}}_2 = \sum_{h=1}^{Z} W_h \bar{X}_{2h} - \sum_{h=1}^{Z} W_h \bar{x}_{2h} \tag{3.26}$$

By substituting (3.23) and (3.24) into (3.22) and substituting (3.22) into (3.19), the obtained estimator can be expressed as:

$$\bar{y}_C = \sum_{h=1}^{Z} W_h \bar{y}_h + \hat{\gamma}_1 \hat{\bar{X}}_1 + \hat{\gamma}_2 \hat{\bar{X}}_2 \tag{3.27}$$

where

$$\hat{\gamma}_1 = \frac{\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{y}_h\right)\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2\right) - \left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h} \bar{y}_h\right)\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h}\right)}{\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2\right)\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2\right) - \left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h}\right)^2} \tag{3.28}$$

$$\hat{\gamma}_2 = \frac{\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h} \bar{y}_h\right)\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2\right) - \left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{y}_h\right)\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h}\right)}{\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h}^2\right)\left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{2h}^2\right) - \left(\sum_{h=1}^{Z} W_h Q_h \bar{x}_{1h} \bar{x}_{2h}\right)^2} \tag{3.29}$$

## 4. Proposed Calibration Estimator

The calibration estimator proposed in Ozgul (2018) was improved upon in this paper by modifying the calibration constraints. Let $y_i$ be the $i^{th}$ observation of the study variable and $x_{1i}$ and $x_{2i}$, $i = 1,2,3,\ldots,N$; $h = 1,2,3,\ldots,Z$ be the $i^{th}$ observation of the two auxiliary variables. The study variable $Y$ and the auxiliary variables $X_1$ and $X_2$ contain $N$ observations divided into $h$ strata with each stratum containing $N_h$ observations such that $N = \sum_{h=1}^{Z} N_h$.

We proposed a new improved calibration estimator as:

$$\bar{y}_p = \sum_{h=1}^{Z} \Omega_{hp} \bar{y}_h \tag{4.1}$$

subject to the calibration constraints defined below:

$$\sum_{h=1}^{Z}\Omega_{hp}=\sum_{h=1}^{Z}W_h \tag{4.2}$$

$$\sum_{h=1}^{Z}\Omega_{hp}\hat{T}_h=\sum_{h=1}^{Z}W_hT_h \tag{4.3}$$

where $\Omega_{hp}$ are the calibrated weights of the proposed estimator

$$\hat{T}_h=\frac{s_{x_{1h}}+s_{x_{2h}}}{\overline{x}_{1h}+\overline{x}_{2h}} \text{ and } T_h=\frac{S_{x_{1h}}+S_{x_{2h}}}{\overline{X}_{1h}+\overline{X}_{2h}}$$

The optimum value for the proposed calibrated weight for each stratum was obtained by minimizing (2.2) subject to the constraints in (4.2) and (4.3) using Lagrange optimization method. The Lagrange function for minimizing (2.2) subject to (4.2) and (4.3) is expressed as:

$$L=\sum_{h=1}^{Z}\frac{\left(\Omega_{hp}-W_h\right)^2}{W_hQ_h}-2\lambda_1\left(\sum_{h=1}^{Z}\Omega_{hp}-\sum_{h=1}^{Z}W_h\right)-2\lambda_2\left(\sum_{h=1}^{Z}\Omega_{hp}\hat{T}_h-\sum_{h=1}^{Z}W_hT_h\right) \tag{4.4}$$

where $\lambda_1$ and $\lambda_2$ are defined as the Lagrange multipliers.

By differentiating (4.4) with respect to $\Omega_{hp}$ and equating the resultant expression to zero, we obtained the optimum calibration weight as:

$$\Omega_{hp}=W_h+W_hQ_h\left(\lambda_1+\lambda_2\hat{T}_h\right) \tag{4.5}$$

where $\lambda_1$ and $\lambda_2$ are obtained by replacing $\Omega_{hp}$ in (4.2) and (4.3) with the optimum value of $\Omega_{hp}$ in (4.5).

$$\lambda_1=\frac{-\left(\sum_{h=1}^{Z}W_h\left(T_h-\hat{T}_h\right)\right)\left(\sum_{h=1}^{Z}W_hQ_h\hat{T}_h\right)}{\left(\sum_{h=1}^{Z}W_hQ_h\right)\left(\sum_{h=1}^{Z}W_hQ_h\hat{T}_h^2\right)-\left(\sum_{h=1}^{Z}W_hQ_h\hat{T}_h\right)^2} \tag{4.6}$$

$$\lambda_2=\frac{\left(\sum_{h=1}^{Z}W_h\left(T_h-\hat{T}_h\right)\right)\left(\sum_{h=1}^{Z}W_hQ_h\right)}{\left(\sum_{h=1}^{Z}W_hQ_h\right)\left(\sum_{h=1}^{Z}W_hQ_h\hat{T}_h^2\right)-\left(\sum_{h=1}^{Z}W_hQ_h\hat{T}_h\right)^2} \tag{4.7}$$

Then by substituting (4.6) and (4.7) into (4.5), the optimum calibrated weights are expressed as:

$$\Omega_{hp}=W_h+W_hQ_h\frac{\left(\sum_{h=1}^{Z}W_h\left(T_h-\hat{T}_h\right)\right)\left(\sum_{h=1}^{Z}W_hQ_h\right)-\left(\sum_{h=1}^{Z}W_h\left(T_h-\hat{T}_h\right)\right)\left(\sum_{h=1}^{Z}W_hQ_h\hat{T}_h\right)}{\left(\sum_{h=1}^{Z}W_hQ_h\right)\left(\sum_{h=1}^{Z}W_hQ_h\hat{T}_h^2\right)-\left(\sum_{h=1}^{Z}W_hQ_h\hat{T}_h\right)^2} \tag{4.8}$$

By substituting $\Omega_{hp}$ obtained in (4.8) into (4.1), we obtained the proposed calibration estimator as:

$$\bar{y}_p = \sum_{h=1}^{Z} W_h \bar{y}_h + \hat{\gamma} \sum_{h=1}^{Z} W_h \left( T_h - \hat{T}_h \right) \tag{4.9}$$

where

$$\hat{\gamma} = \frac{\left( \sum_{h=1}^{Z} W_h Q_h \hat{T}_h \bar{y}_h \right) \left( \sum_{h=1}^{Z} W_h Q_h \right) - \left( \sum_{h=1}^{Z} W_h Q_h \bar{y}_h \right) \left( \sum_{h=1}^{Z} W_h Q_h \hat{T}_h \right)}{\left( \sum_{h=1}^{Z} W_h Q_h \right) \left( \sum_{h=1}^{Z} W_h Q_h \hat{T}_h^2 \right) - \left( \sum_{h=1}^{Z} W_h Q_h \hat{T}_h \right)^2} \tag{4.10}$$

In most situations, $Q_h$ is assumed to be equal to 1 (Ozgul, 2018). For calibration estimator involving one auxiliary variable, $Q_h$ is assumed to be equal to the reciprocal of the sample mean and any other statistic of the auxiliary variable (see Garg and Pachori (2019) and Babatunde et al. (2023)). Since the calibration estimators considered in this paper involve two auxiliary variables, we suggested $Q_h$ to be the reciprocal of the sum of the sample mean and sample standard deviation of the two auxiliary variables, i.e. $\frac{1}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)}$ and $\frac{1}{\left( s_{x_{1h}} + s_{x_{2h}} \right)}$.

The different values of $Q_h$ were used to obtain different versions of the proposed calibration estimator as:

Case I: $Q_h = 1$

$$\bar{y}_{p1} = \sum_{h=1}^{Z} W_h \bar{y}_h + \hat{\gamma}_1 \sum_{h=1}^{Z} W_h \left( T_h - \hat{T}_h \right) \tag{4.11}$$

where

$$\hat{\gamma}_1 = \frac{\left( \sum_{h=1}^{Z} W_h \hat{T}_h \bar{y}_h \right) \left( \sum_{h=1}^{Z} W_h \right) - \left( \sum_{h=1}^{Z} W_h \bar{y}_h \right) \left( \sum_{h=1}^{Z} W_h \hat{T}_h \right)}{\left( \sum_{h=1}^{Z} W_h \right) \left( \sum_{h=1}^{Z} W_h \hat{T}_h^2 \right) - \left( \sum_{h=1}^{Z} W_h \hat{T}_h \right)^2} \tag{4.12}$$

Case II: $Q_h = \frac{1}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)}$

$$\bar{y}_{p2} = \sum_{h=1}^{Z} W_h \bar{y}_h + \hat{\gamma}_2 \sum_{h=1}^{Z} W_h \left( T_h - \hat{T}_h \right) \tag{4.13}$$

where

$$\hat{\gamma}_2 = \frac{\left( \sum_{h=1}^{Z} \frac{W_h \hat{T}_h \bar{y}_h}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)} \right) \left( \sum_{h=1}^{Z} \frac{W_h}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)} \right) - \left( \sum_{h=1}^{Z} \frac{W_h \bar{y}_h}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)} \right) \left( \sum_{h=1}^{Z} \frac{W_h \hat{T}_h}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)} \right)}{\left( \sum_{h=1}^{Z} \frac{W_h}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)} \right) \left( \sum_{h=1}^{Z} \frac{W_h \hat{T}_h^2}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)} \right) - \left( \sum_{h=1}^{Z} \frac{W_h \hat{T}_h}{\left( \bar{x}_{1h} + \bar{x}_{2h} \right)} \right)^2} \tag{4.14}$$

Case III: $Q_h = \dfrac{1}{\left(s_{x_{1h}} + s_{x_{2h}}\right)}$

$$\bar{y}_{p3} = \sum_{h=1}^{Z} W_h \bar{y}_h + \hat{\gamma}_3 \sum_{h=1}^{Z} W_h \left(T_h - \hat{T}_h\right) \tag{4.15}$$

where

$$\hat{\gamma}_3 = \frac{\left(\sum_{h=1}^{Z} \frac{W_h \hat{T}_h \bar{y}_h}{\left(s_{x_{1h}} + s_{x_{2h}}\right)}\right)\left(\sum_{h=1}^{Z} \frac{W_h}{\left(s_{x_{1h}} + s_{x_{2h}}\right)}\right) - \left(\sum_{h=1}^{Z} \frac{W_h \bar{y}_h}{\left(s_{x_{1h}} + s_{x_{2h}}\right)}\right)\left(\sum_{h=1}^{Z} \frac{W_h \hat{T}_h}{\left(s_{x_{1h}} + s_{x_{2h}}\right)}\right)}{\left(\sum_{h=1}^{Z} \frac{W_h}{\left(s_{x_{1h}} + s_{x_{2h}}\right)}\right)\left(\sum_{h=1}^{Z} \frac{W_h \hat{T}_h^2}{\left(s_{x_{1h}} + s_{x_{2h}}\right)}\right) - \left(\sum_{h=1}^{Z} \frac{W_h \hat{T}_h}{\left(s_{x_{1h}} + s_{x_{2h}}\right)}\right)^2} \tag{4.16}$$

## 5. Simulation Study

We demonstrated the efficiency of the proposed calibration estimator over the estimators proposed in Rao et al. (2012), Ozgul (2018) and Rai et al. (2021) through a simulation study. A simulation study establishes the consistency of the obtained result under different scenarios. The study population is MU284 obtained from Sarndal, Swensson and Wretman (1992, pp. 652–659) consisting of 284 municipalities from Sweden partitioned by geographical region into eight strata. The study variable is the 1985 populations (in thousands) while the first and second auxiliary variables are the 1975 populations (in thousands) and total number of seats in municipal council respectively. The population parameters in each stratum are presented in Table 1. Using SRSWOR, a random sample of size $n$; $n_1 = 57$, $n_2 = 71$, $n_3 = 85$, $n_4 = 99$ and $n_5 = 114$, which correspond to 20%, 25%, 30%, 35% and 40% of the population units respectively, were drawn. The sample size for each stratum $n_h = nW_h$ was obtained using proportional allocation.

For each sample size, we simulated $K$ = 50,000 samples and computed both the proposed and existing estimators considered in this work for all the simulated samples. The performance and efficiency of the calibration estimators were assessed using the absolute relative bias (ARB), empirical mean square error (MSE) and percentage relative efficiency (PRE) expressed in (5.1), (5.2) and (5.3) respectively. The results obtained are presented in Tables 2, 3 and 4.

$$ARB\left(\bar{y}_j\right) = \left| \frac{\frac{1}{K}\sum_{k=1}^{K}\left(\bar{y}_j\right)_k - \bar{Y}}{\bar{Y}} \right| \tag{5.1}$$

$$MSE\left(\bar{y}_j\right) = \frac{1}{K}\sum_{k=1}^{K}\left(\left(\bar{y}_j\right)_k - \bar{Y}\right)^2 \tag{5.2}$$

$$PRE\left(\bar{y}_l, \bar{y}_P\right) = \frac{MSE\left(\bar{y}_l\right)}{MSE\left(\bar{y}_P\right)} \times 100 \tag{5.3}$$

where $j = R, O, C$ and $P$ and $l = R, O$ and $C$.

**Table 1:** Population parameters of the study and auxiliary variables

| Strata | Mean | | | Standard Deviation | | |
|---|---|---|---|---|---|---|
| | $Y$ | $X_1$ | $X_2$ | $Y$ | $X_1$ | $X_2$ |
| 1 | 62.4400 | 59.5200 | 51.1600 | 122.0685 | 126.1038 | 13.7860 |
| 2 | 29.6042 | 29.1667 | 47.6667 | 35.9547 | 34.6791 | 12.7628 |
| 3 | 24.0625 | 23.9375 | 50.2500 | 20.7710 | 20.5790 | 10.1704 |
| 4 | 31.0000 | 30.6316 | 48.4737 | 38.6775 | 40.9373 | 8.9406 |
| 5 | 29.4107 | 28.7143 | 46.3571 | 56.2348 | 59.1731 | 9.8060 |
| 6 | 20.8293 | 20.9756 | 46.5610 | 17.5359 | 17.1343 | 8.1272 |
| 7 | 26.6667 | 26.6000 | 54.2000 | 23.8038 | 23.2975 | 11.0224 |
| 8 | 17.5172 | 17.1379 | 40.1724 | 21.4164 | 19.7968 | 9.7912 |

**Table 2:** Absolute Relative Bias of the Calibrated Estimators Using Two Auxiliary Variables

| $Q_h = 1$ | | | | |
|---|---|---|---|---|
| Sample size | $ARB(\bar{y}_P)$ | $ARB(\bar{y}_O)$ | $ARB(\bar{y}_R)$ | $ARB(\bar{y}_C)$ |
| $n_1$ | 0.003880 | 0.031470 | 0.035461 | 0.035462 |
| $n_2$ | 0.002748 | 0.029486 | 0.032508 | 0.032509 |
| $n_3$ | 0.002305 | 0.028617 | 0.030898 | 0.030899 |
| $n_4$ | 0.001492 | 0.027078 | 0.029139 | 0.029140 |
| $n_5$ | 0.002690 | 0.024627 | 0.026226 | 0.026227 |
| $Q_h = 1/(\bar{x}_1 + \bar{x}_2)$ | | | | |
| $n_1$ | 0.008747 | 0.029921 | 0.033910 | 0.033911 |
| $n_2$ | 0.002528 | 0.027947 | 0.030941 | 0.030942 |
| $n_3$ | 0.001439 | 0.027323 | 0.029546 | 0.029547 |
| $n_4$ | 0.004879 | 0.024616 | 0.026268 | 0.026269 |
| $n_5$ | 0.000527 | 0.023686 | 0.025184 | 0.025185 |
| $Q_h = 1/(s_{x_{1h}} + s_{x_{2h}})$ | | | | |
| $n_1$ | 0.009412 | 0.033364 | 0.037549 | 0.037550 |
| $n_2$ | 0.001385 | 0.033441 | 0.036775 | 0.036776 |
| $n_3$ | 0.002847 | 0.033369 | 0.035985 | 0.035986 |
| $n_4$ | 0.000918 | 0.031900 | 0.034010 | 0.034011 |
| $n_5$ | 0.008633 | 0.030618 | 0.032596 | 0.032598 |

**Table 3:** Mean Square Error of the Calibrated Estimators Using Two Auxiliary Variables

| $Q_h = 1$ | | | | |
|---|---|---|---|---|
| Sample size | $MSE(\bar{y}_P)$ | $MSE(\bar{y}_O)$ | $MSE(\bar{y}_R)$ | $MSE(\bar{y}_C)$ |
| $n_1$ | 648.97 | 42693.33 | 54207.70 | 54211.06 |
| $n_2$ | 325.51 | 37478.80 | 45554.96 | 45558.00 |
| $n_3$ | 229.09 | 35302.94 | 41153.64 | 41156.49 |
| $n_4$ | 95.95 | 31608.20 | 36603.02 | 36605.67 |
| $n_5$ | 311.98 | 26144.38 | 29650.31 | 29652.67 |
| $Q_h = 1/(\bar{x}_1 + \bar{x}_2)$ | | | | |
| $n_1$ | 3298.22 | 38593.30 | 49568.96 | 49572.17 |
| $n_2$ | 275.58 | 33670.15 | 41268.60 | 41271.48 |
| $n_3$ | 89.27 | 32182.77 | 37631.68 | 37634.39 |
| $n_4$ | 1026.02 | 26122.38 | 29745.02 | 29747.40 |
| $n_5$ | 11.97 | 24184.35 | 27340.07 | 27342.33 |
| $Q_h = 1/(s_{x_{1h}} + s_{x_{2h}})$ | | | | |
| $n_1$ | 5002.00 | 51977.03 | 65263.01 | 65266.74 |
| $n_2$ | 85.05 | 48265.29 | 58361.29 | 58365.28 |
| $n_3$ | 334.90 | 48173.62 | 56008.55 | 56011.90 |
| $n_4$ | 36.29 | 43866.74 | 49862.30 | 49865.42 |
| $n_5$ | 3212.79 | 40411.47 | 45803.84 | 45806.80 |

**Table 4:** Percentage Relative Efficiency of the Proposed Estimator

| $Q_h = 1$ | | | |
|---|---|---|---|
| Sample size | $PRE(\bar{y}_O, \bar{y}_P)$ | $PRE(\bar{y}_R, \bar{y}_P)$ | $PRE(\bar{y}_C, \bar{y}_P)$ |
| $n_1$ | 65.79 | 83.53 | 83.53 |
| $n_2$ | 115.14 | 139.95 | 139.96 |
| $n_3$ | 154.10 | 179.64 | 179.65 |
| $n_4$ | 329.42 | 381.48 | 381.51 |
| $n_5$ | 83.80 | 95.04 | 95.05 |
| $Q_h = 1/(\bar{x}_1 + \bar{x}_2)$ | | | |
| $n_1$ | 11.70 | 15.03 | 15.03 |
| $n_2$ | 122.18 | 149.75 | 149.76 |
| $n_3$ | 360.51 | 421.55 | 421.58 |
| $n_4$ | 25.46 | 28.99 | 28.99 |
| $n_5$ | 2020.41 | 2284.05 | 2284.24 |
| $Q_h = 1/(s_{x_{1h}} + s_{x_{2h}})$ | | | |
| $n_1$ | 10.39 | 13.05 | 13.05 |
| $n_2$ | 567.49 | 686.20 | 686.25 |
| $n_3$ | 143.84 | 167.24 | 167.25 |
| $n_4$ | 1208.78 | 1374.00 | 1374.08 |
| $n_5$ | 12.58 | 14.26 | 14.26 |

Table 2 presents the absolute relative bias for all the calibration estimators. The proposed estimator has the least absolute relative bias followed by the estimators proposed by Ozgul (2018), Rao et al. (2012) and Rai et al. (2021) for all the cases of $Q_h$ considered. This implies that the estimates of the population mean obtained from the proposed estimator are closer to the population mean compared to the estimates obtained from the estimators proposed in Ozgul (2018), Rao et al. (2012) and Rai et al. (2021). For all the different sample sizes considered, the ARB values of the proposed estimators are smaller compared to the ARB values of the estimators proposed in Ozgul (2018), Rao et al. (2012) and Rai et al. (2021). For $Q_h$ =1, the ARB values of the proposed estimator reduced as the sample size increases except for $n_5$ = 114 where the ARB value increased but for $Q_h = 1/(\bar{x}_1 + \bar{x}_2)$ and $Q_h = 1/(s_{x_{1h}} + s_{x_{2h}})$, the ARB values are not consistent. For $Q_h$ =1 and $Q_h = 1/(s_{x_{1h}} + s_{x_{2h}})$, the least ARB value is observed when the sample size $n_4$ = 99 while for $Q_h = 1/(\bar{x}_1 + \bar{x}_2)$ the least ARB value is observed when the sample size $n_5$ = 114. This suggest that the proposed estimator performed better with a large sample size.

The results in Table 3 are the mean square errors for all the calibration estimators. The proposed estimator has the least mean square error followed by the estimators proposed by Ozgul (2018), Rao et al. (2012) and Rai et al. (2021) for all the cases of $Q_h$ considered. For all the different sample sizes considered, the MSE values of the proposed estimators are smaller compared to the MSE values of the estimators proposed in Ozgul (2018), Rao et al. (2012) and Rai et al. (2021). For $Q_h$ =1, the MSE values of the proposed estimator reduced as the sample size increases except for $n_5$ = 114 where the MSE value increased but for $Q_h = 1/(\bar{x}_1 + \bar{x}_2)$ and $Q_h = 1/(s_{x_{1h}} + s_{x_{2h}})$, the MSE values are not consistent. For $Q_h$ =1 and $Q_h = 1/(s_{x_{1h}} + s_{x_{2h}})$ the least MSE value is observed when the sample size $n_4$ = 99 while for $Q_h = 1/(\bar{x}_1 + \bar{x}_2)$ the least MSE value is observed when the sample size $n_5$ = 114.

From Table 4, all the percentage relative efficiencies obtained are greater than 100% implying that the proposed estimator is more efficient when compared to the estimators proposed by Ozgul (2018), Rao et al. (2012) and Rai et al. (2021) for all the cases of $Q_h$ considered. For all the different sample sizes considered, the proposed estimator is more efficient compared to the estimators proposed in Ozgul (2018), Rao et al. (2012) and Rai et al. (2021). However, the result of this study shows that the proposed estimator is more efficient when compared to the estimators proposed in Ozgul (2018), Rao et al. (2012) and Rai et al. (2021) for large sample sizes. For example, for $Q_h$ =1 and $Q_h = 1/(s_{x_{1h}} + s_{x_{2h}})$, the efficiency of the proposed estimator compared to the estimators proposed in Ozgul (2018), Rao et al. (2012) and Rai et al. (2021) was higher for sample

size $n_4 = 99$ while for $Q_h = 1/(\bar{x}_1 + \bar{x}_2)$ the efficiency of the proposed estimator compared to the estimators proposed in Ozgul (2018), Rao et al. (2012) and Rai et al. (2021) was higher for sample size $n_5 = 114$.

## 6. Conclusion

The standard estimator of the population mean in a stratified sampling was improved in this paper through calibration estimation approach using two auxiliary variables. The calibration estimator proposed in Ozgul (2018) was modified by defining a new set of calibration constraints. Through a simulation study, the efficiency of the proposed calibration estimator was assessed and compared to the estimators proposed in Rao et al. (2012), Ozgul (2018) and Rai et al. (2021). The proposed estimator has the least absolute relative bias and mean square error for all the cases of $Q_h$ considered. These results are consistent with the results obtained in Ozgul (2018), where the absolute relative bias and mean square error of the estimators proposed by Ozgul (2018) and Rao et al. (2018) were compared for $Q_h = 1$. Also, the proposed estimator is more efficient when compared to the estimators proposed in Rao et al. (2012), Ozgul (2018) and Rai et al. (2021) for all the cases of $Q_h$ considered. Furthermore, the efficiency of the proposed estimator was found to be higher for large sample sizes.

## References

Adubi, A. S., Babatunde, O. T., and Ude, I. O., (2022). Modified ratio estimators for population means with two auxiliary parameters using calibration weights. *Asian Journal of Probability and Statistics,* 20(4), pp. 169–183.

Agunbiade, D., Ogunyinka, P., (2013). Effect of correlation level on the use of auxiliary variable in double sampling for regression estimation. *Open Journal of Statistics,* 3(5), pp. 312–318.

Alam, S., Singh, S., and Shabbir, J., (2019). New methodology of calibration in stratified sampling. Proceedings of JSM-Survey Research Methods Section.

Alam, S., Shabbir, J., (2020). Calibration estimation of mean by using double use of auxiliary information. *Communications in Statistics – Simulations and Computation,* https://doi.org/10.1080/03610918.2020.1749660.

Babatunde, O. T., Oladugba, A. V., Ude, I. O., and Adubi, A. S., (2023) Calibration estimation of population mean in stratified sampling using standard deviation. *Quality & Quantity,* https://doi.org/10.1007/s11135-023-01737-1.

Deville, J. C., Sarndal, C., -E., (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87(418), pp. 376–382.

Garg, N., Pachori, M., (2019). Use of coefficient of variation in calibration estimation of population mean in stratified sampling. *Communications in Statistics - Theory and Methods,* https://doi.org/10.1080/03610926.2019.1622729.

Kadilar, C., Cingi, H., (2006). Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters,* 19, pp. 75–79.

Koyuncu, N., Kadilar, C., (2016). Calibration weighting in stratified random sampling. *Communications in Statistics: Simulation and Computation,* 45(7), pp. 2267–2275.

Oladugba, A. V., Adubi, A. S., Okafor, F. C., Babatunde, O. T. and Adubi, P. C., (2023). Calibrated estimators for population means using standard deviation of auxiliary variable. *Journal of the Indian Society for Probability and Statistics,* https://doi.org/10.1007/s41096-023-00167-4.

Ozgul, N., (2018). New calibration estimator based on two auxiliary variables in stratified sampling. *Communication in Statistics – Theory and Methods,* https://doi.org/10.1080/03610926.2018.1433852.

Rao, D., Khan, M., and Khan, S., (2012). Mathematical programming on multivariate calibration estimation in stratified sampling. *World Academy of Science, Engineering and Technology,* 72, pp. 12–27.

Rao, D., Tekabu, T., and Khan, M., (2016). New calibration estimators in stratified sampling. In Computer Science and Engineering (APWC on CSE), 3rd Asia-Pacific World Congress Nadi, Fijion, December 5–6, pp. 66–70, IEEE.

Rai, P. K., Singh, A., and Qasim, M., (2021). Calibration-based estimators using different distance measures under two auxiliary variables: a comparative study. *Journal of Modern Applied Statisticsl Methods,* 19(1), pp. 1–20.

Sarndal, C. -E., Swensson, B., and Wretman, J., (1992). *Model assisted survey sampling*. New York: Springer-Verlag Publishing, https://doi.org/10.1007/978-1-4612-4378-6.

Sisodia, B. V. S., Singh, S., and Singh, S. K., (2017). Calibration approach estimation of the mean in stratified sampling and stratified double sampling. *Communications in Statistics – Theory and Methods,* 46(10), pp. 4932–4942.

Tracey, D. S., Singh, S., and Arnab, R., (2003). Note on calibration in stratified and double Sampling. *Survey Methodology,* 29(1), pp. 99–104.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 93–107,* https://doi.org/10.59170/stattrans-2024-006
Received: 03.04.2022; accepted: 15.05.2023

# Inter-voivodship migration in Poland in the 2000–2020 period based on Markov chain analysis

**Agnieszka Palma**[1], **Dorota Kałuża-Kopias**[2]

## Abstract

The paper presents the scale and directions of inter-voivodship migration in Poland in selected years of the 2000–2020 period. The study focused on permanent residence migration and aimed to identify areas of migration attractiveness and migration catchment voivodships. To study the stochastic nature of these migrations, a Markov chain model was used, in which the states were voivodships. An important aspect of the study involved determining the properties of the transition probability matrix as well as stationary probability in order to characterise the mechanism of inter-voivodship migrations in the years 2000, 2010 and 2020. Data obtained from Statistics Poland were used in the analysis. The transition probability matrix showed that the states were connected and irreducible to each other, while the stationary probability of migration to Dolnośląskie, Małopolskie, Pomorskie, and Wielkopolskie voivodships increased in 2020 compared to 2000. The analysis of the mechanism of migration in the years 2000, 2010, 2020 indicated that Mazowieckie Voivodship was still the main destination for migrants, with the highest stationary probability reaching 0.18 in 2010.

**Key words:** inter-voivodship migration, Markov chain, random transition count, transition probability matrix, stationary probability, mechanism of migration.

## 1. Introduction

Migration, after births and deaths, is the third basic factor influencing the population dynamics of an area. Its importance in influencing population growth and decline and in modifying the demographic characteristics of areas of origin and destination has long been obvious and recognised. The measurement and analysis of migration are important in preparing population estimates and projections for a nation or part of a nation.

The issue of the migration phenomenon is very broad (foreign migration, internal migration) and concerns every spatial scale, starting from the national scale, through the regional (provincial) level, other selected functional administrative units, up to complex settlement systems (Śleszyński et al. 2018). In addition, it deals with explaining the reasons for the variation in migration volumes. Certainly, many factors influence the intensity and direction of migration flows. One of them is the pace of socio-economic development of a given region, employment opportunities, housing or land prices, cost of living, availability of infrastructure, to a large extent family ties, and the age of migrants.

[1]Institute of Statistics and Demography, Faculty of Economics and Sociology, University of Lodz, Poland. E-mail: agnieszka.palma@uni.lodz.pl. ORCID: https://orcid.org/0000-0002-3558-1568.
[2]Institute of Statistics and Demography, Faculty of Economics and Sociology, University of Lodz, Poland. E-mail: dorota.kaluza@uni.lodz.pl. ORCID: https://orcid.org/ 0000-0001-5023-2596.

The topic of internal migration for permanent residence in Poland has been addressed in several works. Pietrzak (2013) described inter-regional migration using a gravity model. Roszko (2018) found a strong relationship between the standard of living of residents expressed by Gross Domestic Product per capita and the number of incoming new residents and the balance of interprovincial migration. Several works have addressed internal migration in a specific province: Rosner (2014), Ilnicki (2020), Józefowicz (2020), Kałuża-Kopias (2021), Ilnicki (2021).

To investigate the stochastic nature of inter-regional migration in Poland in the period 2000–2020, a Markov chain model was used. Markov chains describe many real-world processes, and they are used in many different fields such as physics, geography, chemistry, biology, medicine, music, economics and finance, game theory, sports, and more. Many examples of their application can be found in the literature, see for example: Clark (1960), Marble (1967), Iosifescu (2007), Privault (2018). Collins (1972) used Markov chains in forecasting industrial migration, while Berry (1971) in outlined a short-term model of neighbourhood turnover, Bourne (1976) suggested this method to monitor changes in Toronto's spatial structure, Azizah (2019) applied Markov chain to forecast rainfall data and Chu (2020) used a Markov chain model to forecast future land use change, Barra (2020) to count and model migraine attacks and Romeu (2020) to analyse covid-19 survival.

Markov chain models are useful and convenient tools for describing and analysing the nature of dynamic changes of a phenomenon or process of interest. They are an important tool for geographers who deal with mobility problems. These can be movements from one place to another, as well as movements from one state to another. State can be defined in different ways, it can refer to a class of provinces, municipalities, city size or income or type of land use or some other variable. They can also be used to forecast future changes: Sempewo (2016), Rahimipour (2018).

One of the applications of Markov chains is their use in migration studies, mainly to determine the dominant direction or rate of change, as well as the development of a system, for example, an urban system. Thanks to Markov chains we can determine which cities, provinces, or other territorial units have a tendency to increase in population and which to decrease. The study takes into account permanent migration movements between provinces to analyse the changes taking place.

In order to analyse the changes taking place in the years 2000–2020, the migration movements for permanent residence taking place between provinces were taken into account, and the probability matrices of the transition between states and the corresponding stationary distributions were determined. The empirical material used in the study was data from the Local Data Bank of Statistics Poland concerning the inflow and outflow of the population in individual voivodships. They allowed measuring the migration volume with the Markov chain model. It should be emphasised that data from public statistics do not allow for a reliable assessment of the migration situation in Poland. It results mainly from the fact that they are based on registration data, thus they do not register real permanent migration, which is not connected with checking out or registering. Despite the continuous improvement of current migration statistics both in terms of data collection methods and compilation techniques, some authors are critical of their quality (e.g. Jończy 2014; Śleszyński 2005, 2011). Theoretically, the number of previous and current places of regis-

tration for migrants should be equal, but in practice, the differences can be significant. The largest underestimation of internal migration concerns large cities and their suburban zones, as a large part of the inflow remains unregistered, Korcelii (1997). Although the source of information on migration is not perfect, it should be emphasised that data from registration registers, due to their continuity and updating, form the basis of migration reporting.

## 2. Migration between voivodships in Poland in dynamic terms

The last two decades have seen a decrease in the size of internal migration in Poland. Between 2000 and 2020, on average, the number of inter-regional migrations decreased by about 0.5% from year to year, (the calculated geometric mean value for 2000–2020 was 0.995). In spatial terms, the effect of these movements is the migration balance.

Table 1: Internal migration by voivodships

| | Net migration | | | Net migration per 1000 people | | | Index of migration attractiveness | | |
|---|---|---|---|---|---|---|---|---|---|
| **voivodships** | **2000** | **2010** | **2020** | **2000** | **2010** | **2020** | **2000** | **2010** | **2020** |
| **Dolnośląskie** | -573 | 1579 | 3295 | -0,2 | 0,54 | 1,14 | -0,04 | 0,1 | 0,22 |
| **Kujawsko-pomorskie** | -407 | -1443 | -2112 | -0,2 | -0,69 | -1,02 | -0,04 | -0,14 | -0,21 |
| **Lubelskie** | -2969 | -4867 | -4685 | -1,35 | -2,23 | -2,23 | -0,27 | -0,41 | -0,43 |
| **Lubuskie** | -440 | -474 | -842 | -0,44 | -0,46 | -0,83 | -0,06 | -0,07 | -0,14 |
| **Łódzkie** | -1107 | -1757 | -1812 | -0,42 | -0,69 | -0,74 | -0,09 | -0,15 | -0,19 |
| **Małopolskie** | 2376 | 3673 | 3412 | 0,74 | 1,1 | 1 | 0,16 | 0,24 | 0,23 |
| **Mazowieckie** | 8825 | 12687 | 10448 | 1,73 | 2,41 | 1,92 | 0,32 | 0,4 | 0,38 |
| **Opolskie** | -88 | -671 | -747 | -0,08 | -0,66 | -0,76 | -0,01 | -0,1 | -0,13 |
| **Podkarpackie** | -1730 | -1973 | -2226 | -0,82 | -0,93 | -1,05 | -0,2 | -0,22 | -0,25 |
| **Podlaskie** | -1255 | -1616 | -1702 | -1,04 | -1,34 | -1,45 | -0,19 | -0,26 | -0,31 |
| **Pomorskie** | 1651 | 2749 | 3825 | 0,76 | 1,21 | 1,63 | 0,14 | 0,22 | 0,3 |
| **Śląskie** | -1652 | -3194 | -3342 | -0,35 | -0,69 | -0,74 | -0,07 | -0,16 | -0,2 |
| **Świętokrzyskie** | -2061 | -2567 | -2143 | -1,58 | -2 | -1,74 | -0,23 | -0,31 | -0,31 |
| **Warmińsko-Mazurskie** | -2002 | -2721 | -2113 | -1,4 | -1,87 | -1,49 | -0,18 | -0,26 | -0,23 |
| **Wielkopolskie** | 1595 | 1706 | 1491 | 0,48 | 0,5 | 0,43 | 0,11 | 0,12 | 0,11 |
| **Zachodniopomorskie** | -163 | -1111 | -747 | -0,1 | -0,64 | -0,44 | -0,01 | -0,11 | -0,09 |

Source: Authors' own calculations.

Most of the voivodships in Poland were characterized by a negative balance of inter-voivodship movements. In 2000, 12 voivodships had negative migration balances. The regions with the lowest urbanisation level - Świętokrzyskie and Lubelskie - had relatively the largest population migration losses exceeding 2,000 people (Table 1). The largest positive net migration volumes occurred in the Mazowieckie, Pomorskie, Wielkopolskie, and Małopolskie voivodships. In subsequent years, Dolnośląskie joined the group of voivodships with a positive migration balance. The regions characterized by permanent migration losses of the population were traditionally the areas of eastern and north-eastern Poland, which are mostly poorly urbanised agricultural areas. Relating migration balances to the population in individual voivodships, it turns out that in 2000, only in one region out of four with positive inter-voivodship migration balances, there was a migration increase exceeding 1 person per 1000 population (Table 1).

As regards the intensity of the migration loss of inhabitants, the northeastern voivodships had relatively the largest negative balances: Warmińsko-Mazurskie (-1.4 persons per 1,000 inhabitants) and Świętokrzyskie (-1.58). Over the years, there was a clear advantage in the size of positive migration balances (both in absolute and relative terms) of the Mazowieckie

voivodship over other regions with higher inflows than outflows. Similarly, in the case of voivodships characterized by negative migration balances, in the subsequent years the migration loss of inhabitants assumed (in most regions) greater dimensions than in 2000. It should be noted, however, that - when assessing the size of migration - it is worth relating the size of migration balances to the overall migration turnover (inflow and outflow), as the same values of migration balances may be the result of a large inflow and outflow on the one hand or a much smaller size of these components (Obraniak 2007). As a measure of migration attractiveness (in demographic literature, the name "migration efficiency index" is also used), an index representing the ratio of migration balance to migration turnover was adopted (Table 1). Due to its construction and ease of interpretation, it is a quite commonly used measure (Potrykowska, Śleszyński 1999; Obraniak 2007; Kałuża-Kopias 2021). The calculated values of the index for 2000 indicate that the Mazowieckie voivodship was the most attractive in terms of migration (Table 1). The following places were occupied by agglomeration voivodships: Małopolskie, Pomorskie, Wielkopolskie. The list was closed by the voivodships of north-eastern, eastern, and central Poland - Lubelskie, Podlaskie, Warmińsko-Mazurskie, Podkarpackie, and Świętokrzyskie.

## 3. Mathematical model description and calculations

### 3.1. Markov chain model.

Let $n$ and $k$ be elements of $\mathbf{N}$, such that $n \geq 1$ and $k \geq 1$. Define $S = \{1, \ldots, k\}$. Consider a sequence of random variables $\{X_1, X_2, \ldots, X_n\}$ such that

$$p_{ij} = P(X_{d+1} = j | X_d = i) = P(X_{d+1} = j | X_1 = i_1,\ X_2 = i_2,\ \ldots,\ X_d = i)$$

is independent of $d$ for all $i, j \in S$. Then, the sequence $\{X_1, X_2, \ldots, X_n\}$ is a first-order Markov chain with state space $S$ and transition probabilities $p_{ij}$ for $i, j \in S$.

The process starts in one of the states and moves successively from one state to another. Each move is called a step. If the chain is currently in state $i$, then it moves to state $j$ at the next step with a probability denoted by $p_{ij}$, and this probability does not depend upon which state the chain was in before the current state.
The transition probability matrix $P = [p_{ij}]_{i,j \in S}$ of finite Markov chains, is defined as follows:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}. \tag{1}$$

Since the elements of row $i$ of this matrix represent the conditional probabilities for all possible state changes from state $i$, they must satisfy

$$\forall_{i,j \in S} \quad p_{ij} \geq 0, \quad \sum_{j=1}^{k} p_{ij} = 1 \text{ for each } i \in S. \tag{2}$$

A square matrix for which all elements are non-negative and the sum of the elements in each row is 1 is called a stochastic matrix. It follows from this definition that a Markov chain with known probability distribution of the initial state is completely characterized by a $k \times k$ matrix containing the transition probabilities $p_{ij}$.

Transition probability matrix $P$ shows transition probabilities from one state to another in one step of Markov chain. Transition probabilities from state $i$ to state $j$ in $n$ steps is noted as

$$p_{ij}^{(n)} = P(X_{d+n} = j | X_d = i) \text{ for } i, j \in S, \text{ and } d, n \in \mathbf{N}. \tag{3}$$

Transition probability matrix in $n$ steps is $P^{(n)}$.

Throughout the paper we are dealing with a random transition count matrix $M$, which is defined in the usual way as

$$M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1k} \\ m_{21} & m_{22} & \dots & m_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ m_{k1} & m_{k2} & \dots & m_{kk} \end{bmatrix}, \tag{4}$$

where $m_{ij}$ denotes number of transitions from state $i$ to state $j$, where $i, j \in S$.

Let $\sum_{j=1}^{k} m_{ij} = m_{i.}$, $\sum_{i=1}^{k} m_{ij} = m_{.j}$ and $\sum_{i,j=1}^{k} m_{ij} = m_{..}$.

Then,

Table 2: Random transition count

| States | 1 | 2 | … | $k$ | Total |
|---|---|---|---|---|---|
| 1 | $m_{11}$ | $m_{12}$ | … | $m_{1k}$ | $m_{1.}$ |
| 2 | $m_{21}$ | $m_{22}$ | … | $m_{2k}$ | $m_{2.}$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| $k$ | $m_{k1}$ | $m_{k2}$ | … | $m_{kk}$ | $m_{k.}$ |
| | $m_{.1}$ | $m_{.2}$ | … | $m_{.k}$ | $m_{..}$ |

We are interested in the estimation of the elements $p_{ij}$ of matrix $P$; we denote them by $\hat{p}_{ij}$.

Using the maximum likelihood estimation method (Bhat and Miller, 2002; Billingsley, 1961) we obtain an estimate of the matrix P and denote it as

$$P = [\hat{p}_{ij}], \text{ where } \hat{p}_{ij} = \frac{m_{ij}}{m_{i.}} \quad \text{for each} \quad i, j \in \{1, \dots, k\}. \tag{5}$$

### 3.2. Markov chain's state space and calculations of migration

In our Markov chain model the states are voivodships. For the convenience of the reader we denote voivodships respectively $1, 2, \dots, 16$ and we give this below.

Table 3: States and corresponding voivodships

| States | voivodships |
|---|---|
| 1 | Dolnośląskie |
| 2 | Kujawsko-Pomorskie |
| 3 | Lubelskie |
| 4 | Lubuskie |
| 5 | Łódzkie |
| 6 | Małopolskie |
| 7 | Mazowieckie |
| 8 | Opolskie |
| 9 | Podkarpackie |
| 10 | Podlaskie |
| 11 | Pomorskie |
| 12 | Śląskie |
| 13 | Świętokrzyskie |
| 14 | Warmińsko-Mazurskie |
| 15 | Wielkopolskie |
| 16 | Zachodniopomorskie |

In this way we obtain the state space $S = \{1, 2, \ldots, 16\}$.

Input data were obtained from Statistics Poland. They are available in the Local Data Bank (BDL) and the Demography database. Data on migration volumes are presented in a matrix system. The matrix shows the internal migration of the population for permanent residence by voivodship of previous and current place of residence. The elements $m_{ij},\ i, j \in S$, of the matrix $M_i,\ i = 1, 2, 3$, denote the number of migrants who emigrated and stayed in another voivodship in Poland, respectively in the years 2000, 2010, 2020. Migrants are people who changed their place of permanent residence and moved to another province.

Based on the input data $M_i,\ i = 1, 2, 3$, after applying the maximum likelihood estimation defined by formula (5), we obtain an estimate of the probability of migration from $i$- th province to $j$-th province. In this way, the transition probability matrices $P_1,\ P_2,\ P_3$ for migration in Poland were determined for the years 2000, 2010, and 2020, respectively.

### 3.3. Test for first-order Markov chain

In this section, we will check whether the transition probability matrix for migration in 2000, 2010, and 2020 satisfies the assumption of a first-order Markov chain. To check the validity of the Markov chain model, the chi-square test of goodness of fit is used. The hypothesis to be tested is the null hypothesis that the collected observations are independent of the alternative hypothesis that the observed process is a first-order Markov chain. The hypothesis is $Ho : P = P_0$, where $P_0$ has identical rows under the assumption of independence. The $\chi^2$ statistic to test independence against the first order Markov Chain (Bhat and Miller, 2002; Billingsley, 1961) is calculated from the following relationship:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{(m_{ij} - \frac{m_{i\cdot}m_{\cdot j}}{m_{\cdot\cdot}})^2}{\frac{m_{i\cdot}m_{\cdot j}}{m_{\cdot\cdot}}}$$

with the degrees of freedom $(k-1)^2 - d$, where $d$ is number of zero cells. For the input data, $\chi^2$ statistics were determined as well as Cramer's association coefficient. The results are given below:

Table 4: $\chi^2$ test of independence for data migration in 2000, 2010, 2020

| Migration data | $\chi^2$ statistic | p-value | Cramer's V |
|---|---|---|---|
| 2000 | 82103 | <0.0000 | 0.234 |
| 2010 | 85064 | <0.0000 | 0.238 |
| 2020 | 85425 | <0.0000 | 0.251 |

Source: Authors' own calculations.

It follows from Table 4, that the assumption of independence can be rejected (p-value< 0.0000). Based on these results, it can be concluded that modelling the data as a first-order Markov chain is reasonable.

### 3.4. Stationary distribution

As we progress through time, the probability of being in certain states more likely than others. Over the long run, the distribution will reach equilibrium with an associated probability of being in each state. This is known as the stationary distribution. A stationary distribution $\pi$ is a vector whose entries are non-negative and sum to 1, is unchanged by the operation of transition matrix P on it, so it satisfied (Bhat 2002),

$$\pi_j = \sum_{i \in S} \pi_i \cdot p_{ij} \quad \text{and} \quad \forall_{j \in S}\, \pi_j \geq 0, \quad \sum_j \pi_j = 1.$$

In the matrix notation, it can be written as

$$\pi = \pi \cdot P_i \text{ for } i = 1,2,3,$$

where $\pi$ is some distribution, which is a row vector with the number of columns equal to the states in the state space $S$ and $P_i,\ i = 1,2,3$, is the transition probability matrix.

The stationary probability can also be found from limiting transition probability matrix

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j.$$

Comparison of stationary distribution between the years 2000, 2010, 2020 for individual voivodships is presented in Figure 1.

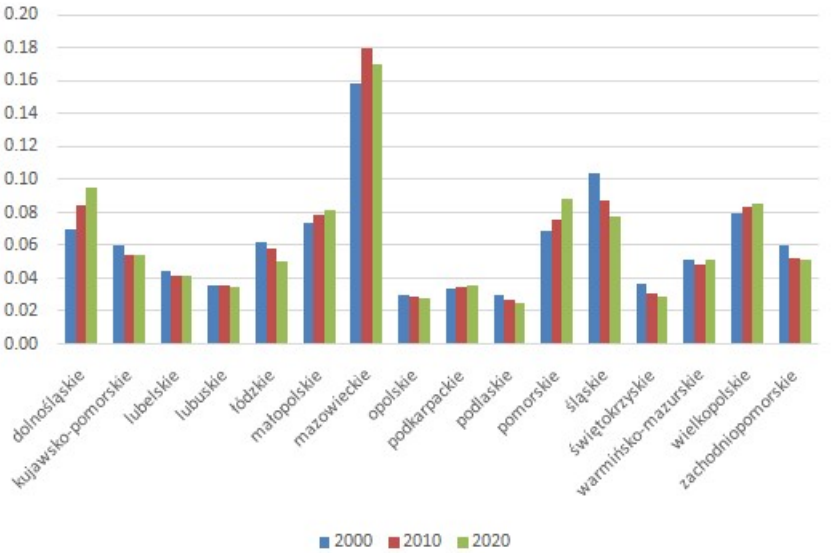


Figure 1: Comparison of stationary inter-voivodship migration distributions in Poland in 2000, 2010, 2020

Table 5: Stationary probability $\pi$

| Transition matrix | $\pi_1$ | $\pi_2$ | $\pi_3$ | $\pi_4$ | $\pi_5$ | $\pi_6$ | $\pi_7$ | $\pi_8$ | $\pi_9$ | $\pi_{10}$ | $\pi_{11}$ | $\pi_{12}$ | $\pi_{13}$ | $\pi_{14}$ | $\pi_{15}$ | $\pi_{16}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_1$ | 0.07 | 0.06 | 0.04 | 0.04 | 0.06 | 0.07 | 0.16 | 0.03 | 0.03 | 0.03 | 0.07 | 0.1 | 0.04 | 0.05 | 0.08 | 0.06 |
| $P_2$ | 0.08 | 0.05 | 0.04 | 0.04 | 0.06 | 0.08 | 0.18 | 0.03 | 0.04 | 0.03 | 0.08 | 0.09 | 0.03 | 0.05 | 0.08 | 0.05 |
| $P_3$ | 0.1 | 0.05 | 0.04 | 0.04 | 0.05 | 0.08 | 0.17 | 0.03 | 0.04 | 0.03 | 0.09 | 0.08 | 0.03 | 0.05 | 0.09 | 0.05 |

Source: Authors' own calculations.

Figure 1 and Table 5 - the results of the analysis of stationary probability show that the variation for the following provinces: Dolnośląskie, Mazowieckie, Pomorskie, Śląskie is greater compared to the remaining provinces, for which the stationary distributions are differentiated, but the changes are relatively small. The values of stationary probabilities indicate that the probability of migration to Łódzkie decreased from 0.062 in 2000 to 0.05 in 2020, while for Śląskie from 0.103 to 0.077. On the other hand, the probability of migration to Dolnośląskie increased from 0.07 to 0.095, for the Pomorskie voivodship from 0.069 to 0.089, analogically, for the Podkarpackie and Wielkopolskie voivodships, the probabilities increased, although to a smaller extent. Based on these results, one may conclude that the Mazowieckie voivodship will continue to be the main direction of migration in the future. However, it is worth noting that in 2020 the probability of migration to the Mazowieckie voivodship decreased in comparison to 2010.

### 3.5. Mechanism of the inter-voivodship migration in Poland

In order to characterize the mechanism of interprovincial migration in Poland, one should first estimate the transition probability matrix using the procedure given in Bhat and Miller (2002) and Miall (1973) as follows:

$$P^0 = \begin{bmatrix} 0 & \frac{m_{.2}}{m_{..}-m_{.1}} & \frac{m_{.3}}{m_{..}-m_{.1}} & \cdots & \frac{m_{.k}}{m_{..}-m_{.1}} \\ \frac{m_{.1}}{m_{..}-m_{.2}} & 0 & \frac{m_{.3}}{m_{..}-m_{.2}} & \cdots & \frac{m_{.k}}{m_{..}-m_{.2}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{m_{.1}}{m_{..}-m_{.k}} & \frac{m_{.2}}{m_{..}-m_{.k}} & \frac{m_{.3}}{m_{..}-m_{.k}} & \cdots & 0 \end{bmatrix}. \tag{6}$$

To analyze the mechanism of the inter-voivodship migration in Poland, we find the estimated transition probability matrix $P^0$ by formula (6) and using $M_1$, $M_2$, $M_3$ - data from Statistics Poland for the 2000, 2010, 2020, respectively. The results of these calculations are presented, as an example, only for the matrix $P_1^0$. The matrices $P_2^0$, $P_3^0$ are determined analogically.

$$P_1^0 = \begin{bmatrix} 0 & 0.06 & 0.04 & 0.04 & 0.06 & 0.09 & 0.2 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.07 & 0 & 0.04 & 0.04 & 0.06 & 0.09 & 0.19 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.07 & 0.06 & 0 & 0.04 & 0.06 & 0.09 & 0.19 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.07 & 0.05 & 0.04 & 0 & 0.06 & 0.09 & 0.19 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.07 & 0.06 & 0.04 & 0.04 & 0 & 0.09 & 0.19 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.08 & 0.06 & 0.04 & 0.04 & 0.06 & 0 & 0.2 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.09 & 0.06 \\ 0.09 & 0.06 & 0.05 & 0.04 & 0.07 & 0.1 & 0 & 0.04 & 0.04 & 0.03 & 0.08 & 0.13 & 0.04 & 0.05 & 0.1 & 0.07 \\ 0.07 & 0.05 & 0.04 & 0.04 & 0.06 & 0.09 & 0.19 & 0 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.07 & 0.05 & 0.04 & 0.04 & 0.06 & 0.09 & 0.19 & 0.03 & 0 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.07 & 0.05 & 0.04 & 0.04 & 0.06 & 0.09 & 0.19 & 0.03 & 0.04 & 0 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.07 & 0.06 & 0.04 & 0.04 & 0.06 & 0.09 & 0.2 & 0.03 & 0.04 & 0.03 & 0 & 0.11 & 0.04 & 0.05 & 0.08 & 0.06 \\ 0.08 & 0.06 & 0.04 & 0.04 & 0.06 & 0.09 & 0.2 & 0.04 & 0.04 & 0.03 & 0.08 & 0 & 0.04 & 0.05 & 0.09 & 0.06 \\ 0.07 & 0.05 & 0.04 & 0.04 & 0.06 & 0.09 & 0.19 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0 & 0.05 & 0.08 & 0.06 \\ 0.07 & 0.06 & 0.04 & 0.04 & 0.06 & 0.09 & 0.19 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0 & 0.08 & 0.06 \\ 0.08 & 0.06 & 0.04 & 0.04 & 0.06 & 0.09 & 0.2 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0 & 0.06 \\ 0.07 & 0.06 & 0.04 & 0.04 & 0.06 & 0.09 & 0.19 & 0.03 & 0.04 & 0.03 & 0.07 & 0.11 & 0.04 & 0.05 & 0.08 & 0 \end{bmatrix},$$

In the next step, transition probability difference matrices is created, assuming that the transition probabilities are given by (5) and (6), so that the properties of the migration mechanism can be inferred. In this case, we compute the difference matrices

$$D_i = P_i - P_i^0 \quad \text{for each} \quad i \in \{1,2,3\}.$$

Below, we present only the matrix $D_1$, to illustrate the method used.

$$D_1 = \begin{bmatrix}
0 & -0.03 & -0.01 & 0.08 & 0 & -0.02 & -0.08 & 0.09 & -0.01 & -0.01 & -0.04 & -0.01 & -0.01 & -0.03 & 0.07 & 0 \\
-0.04 & 0 & -0.03 & -0.01 & -0.02 & -0.07 & -0.01 & -0.03 & -0.03 & -0.02 & 0.13 & -0.05 & -0.03 & 0.03 & 0.13 & 0.03 \\
-0.03 & -0.04 & 0 & -0.02 & -0.03 & -0.02 & 0.27 & -0.02 & 0.05 & -0.01 & -0.02 & -0.03 & 0 & -0.03 & -0.06 & -0.02 \\
0.17 & -0.02 & -0.03 & 0 & -0.02 & -0.06 & -0.12 & -0.01 & -0.02 & -0.02 & -0.04 & -0.06 & -0.03 & -0.03 & 0.18 & 0.1 \\
-0.01 & -0.01 & -0.02 & -0.02 & 0 & -0.05 & 0.13 & 0 & -0.02 & -0.01 & -0.03 & 0.05 & 0 & -0.03 & 0.03 & -0.01 \\
-0.02 & -0.04 & -0.01 & -0.02 & -0.03 & 0 & -0.07 & -0.01 & 0.09 & -0.02 & -0.05 & 0.26 & 0.05 & -0.04 & -0.06 & -0.04 \\
-0.04 & 0.01 & 0.07 & -0.03 & 0.06 & -0.05 & 0 & -0.02 & -0.01 & 0.05 & 0 & -0.05 & 0.03 & 0.06 & -0.05 & -0.02 \\
0.2 & -0.04 & -0.03 & -0.01 & 0.01 & -0.03 & -0.1 & 0 & -0.01 & -0.02 & -0.06 & 0.2 & -0.02 & -0.03 & -0.03 & -0.03 \\
-0.02 & -0.04 & 0.08 & -0.02 & -0.03 & 0.22 & -0.03 & -0.02 & 0 & -0.02 & -0.05 & 0.02 & 0.03 & -0.03 & -0.05 & -0.03 \\
-0.05 & -0.03 & -0.01 & -0.03 & -0.03 & -0.06 & 0.26 & -0.03 & -0.02 & 0 & 0.01 & -0.06 & -0.03 & 0.15 & -0.06 & -0.03 \\
-0.04 & 0.11 & -0.02 & -0.01 & -0.02 & -0.06 & -0.02 & -0.02 & -0.02 & 0 & 0 & -0.07 & -0.02 & 0.09 & 0.01 & 0.11 \\
0.01 & -0.02 & 0 & -0.02 & 0.02 & 0.14 & -0.09 & 0.05 & 0.01 & -0.01 & -0.04 & 0 & 0.03 & -0.01 & -0.04 & -0.02 \\
-0.03 & -0.05 & 0 & -0.03 & 0.01 & 0.09 & 0.08 & -0.02 & 0.04 & -0.02 & -0.04 & 0.11 & 0 & -0.04 & -0.06 & -0.04 \\
-0.05 & 0.04 & -0.02 & -0.02 & -0.03 & -0.07 & 0.09 & -0.02 & -0.03 & 0.11 & 0.14 & -0.05 & -0.03 & 0 & -0.04 & -0.02 \\
0.07 & 0.11 & -0.02 & 0.08 & 0.03 & -0.06 & -0.09 & -0.01 & -0.03 & -0.02 & -0.01 & -0.05 & -0.03 & -0.03 & 0 & 0.06 \\
0 & 0.04 & -0.01 & 0.06 & -0.01 & -0.06 & -0.05 & -0.02 & -0.02 & -0.02 & 0.09 & -0.05 & -0.02 & -0.02 & 0.1 & 0
\end{bmatrix},$$

The positive elements of the matrices $D_1$, $D_2$, $D_3$ represent those transitions which have higher probability of occurrence. This makes it possible to characterize the mechanism of the process of interregional migration in the years 2000, 2010, 2020.

The stationary distribution allowed finding voivodships that are migration catchment areas, while the values of the matrix $D_i$, $i = 1,2,3$, indicate areas attractive to migration for individual voivodships.

Table 6: Positive values of matrix $D_i$, $i = 1,2,3$, i.e. mechanism of migration between voivodships in the years 2000, 2010, 2020

| state | Dolnośląskie | Kujawsko-Pomorskie | Lubelskie | Lubuskie | Łódzkie | Małopolskie | Mazowieckie | Opolskie | Podkarpackie | Podlaskie | Pomorskie | Śląskie | Świętokrzyskie | Warmińsko-Mazurskie | Wielkopolskie | Zachodnio-pomorskie | year |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dolnośląskie** | | | | 0.08 | | | | 0.09 | | | | 0.01 | | | 0.07 | | 2000 |
| | | | | 0.10 | | | | 0.08 | | | | 0.01 | | | 0.08 | | 2010 |
| | | | | 0.08 | 0.01 | | | 0.09 | | | | 0.01 | | | 0.08 | 0.01 | 2020 |
| **Kujawsko-Pomorskie** | | | | | | | | | | | 0.13 | | | 0.03 | 0.13 | 0.03 | 2000 |
| | | | | | | | | | | | 0.16 | | | 0.01 | 0.16 | 0.01 | 2010 |
| | | | | | | | | | | | 0.14 | | | 0.02 | 0.15 | 0.01 | 2020 |
| **Lubelskie** | | | | | | | 0.27 | | 0.05 | | | | | | | | 2000 |
| | | | | | | | 0.32 | | 0.04 | | | | | | | | 2010 |
| | | | | | | | 0.33 | | 0.05 | | | | | | | | 2020 |
| **Lubuskie** | 0.17 | | | | | | | | | | | | | | 0.18 | 0.10 | 2000 |
| | 0.15 | | | | | | | | | | | | | | 0.20 | 0.08 | 2010 |
| | 0.20 | | | | | | | | | | | | | | 0.21 | 0.10 | 2020 |
| **Łódzkie** | 0.01 | | | | | | 0.13 | | | | | 0.05 | | | 0.03 | | 2000 |
| | 0.04 | | | | | | 0.15 | | | | | 0.03 | | | 0.02 | | 2010 |
| | | | | | | | 0.12 | | | | | 0.03 | | | 0.03 | | 2020 |
| **Małopolskie** | | | | | | | | | 0.09 | | | 0.26 | 0.05 | | | | 2000 |
| | | | | | | | | | 0.10 | | | 0.26 | 0.03 | | | | 2010 |
| | | | | | | | | | 0.11 | | | 0.24 | 0.04 | | | | 2020 |
| **Mazowieckie** | | 0.01 | 0.07 | | 0.06 | | | | | 0.05 | | | 0.03 | 0.06 | | | 2000 |
| | | 0.01 | 0.08 | | 0.07 | | | | | 0.05 | | | 0.02 | 0.05 | | | 2010 |
| | | 0.01 | 0.09 | | 0.05 | | | | | 0.05 | | | 0.02 | 0.06 | | | 2020 |
| **Opolskie** | 0.20 | | | | 0.01 | | | | | | | 0.20 | | | | | 2000 |
| | 0.27 | | | | | | | | | | | 0.17 | | | | | 2010 |
| | 0.31 | | | | | | | | | | | 0.15 | | | | | 2020 |
| **Podkarpackie** | | | 0.08 | | | 0.22 | | | | | | 0.02 | 0.03 | | | | 2000 |
| | | | 0.05 | | | 0.24 | | | | | | 0.01 | 0.03 | | | | 2010 |
| | | | 0.05 | | | 0.29 | | | | | | | 0.02 | | | | 2020 |
| **Podlaskie** | | | | | | | 0.26 | | | | 0.01 | | | 0.15 | | | 2000 |
| | | | | | | | 0.28 | | | | | | | 0.13 | | | 2010 |
| | | | | | | | 0.29 | | | | | | | 0.12 | | | 2020 |
| **Pomorskie** | | 0.11 | | | | | | | | | | | | 0.09 | 0.01 | 0.11 | 2000 |
| | | 0.11 | | | | | | | | | | | | 0.01 | | 0.10 | 2010 |
| | | 0.11 | | | | | | | | | | | | 0.12 | | 0.10 | 2020 |
| **Śląskie** | 0.01 | | | | 0.02 | 0.14 | | 0.05 | 0.01 | | | | 0.03 | | | | 2000 |
| | 0.01 | | | | 0.01 | 0.16 | | 0.05 | 0.01 | | | | 0.03 | | | | 2010 |
| | 0.02 | | | | 0.02 | 0.17 | | 0.06 | | | | | 0.02 | | | | 2020 |
| **Świętokrzyskie** | | | 0.01 | | 0.01 | 0.09 | 0.08 | | 0.04 | | | 0.11 | | | | | 2000 |
| | | | 0.01 | | 0.01 | 0.11 | 0.09 | | 0.04 | | | 0.04 | | | | | 2010 |
| | | | | | 0.01 | 0.14 | 0.10 | | 0.04 | | | 0.02 | | | | | 2020 |
| **Warmińsko-Mazurskie** | | 0.04 | | | | | 0.09 | | | 0.11 | 0.14 | | | | | | 2000 |
| | | 0.03 | | | | | 0.07 | | | 0.07 | 0.20 | | | | | | 2010 |
| | | 0.03 | | | | | 0.06 | | | 0.07 | 0.25 | | | | | | 2020 |
| **Wielkopolskie** | 0.07 | 0.11 | | 0.08 | 0.03 | | | | | | | | | | | 0.06 | 2000 |
| | 0.09 | 0.09 | | 0.06 | 0.03 | | | | | | | | | | | 0.07 | 2010 |
| | 0.11 | 0.08 | | 0.06 | 0.03 | | | | | | | | | | | 0.06 | 2020 |
| **Zachodnio-pomorskie** | | 0.04 | | 0.06 | | | | | | | 0.09 | | | | 0.10 | | 2000 |
| | | 0.02 | | 0.08 | | | | | | | 0.08 | | | | 0.11 | | 2010 |
| | | 0.01 | | 0.08 | | | | | | | 0.08 | | | | 0.14 | | 2020 |

Source: Authors' own calculations.

Figure 2 presents graphically the mechanism of inter-voivodship migration in 2020. Individual voivodships have been marked with appropriate colours, depending on the values

of probabilities $\pi_i,\ i \in \{1,2,\cdots,16\}$ of the stationary distribution $P_3$ in Table 5. The arrows between provinces are of different thickness, corresponding to the values of the matrix $D_3$ and showing the intensity of migration between provinces. For the convenience of the reader, the arrows coming out of different provinces are marked with different colours.

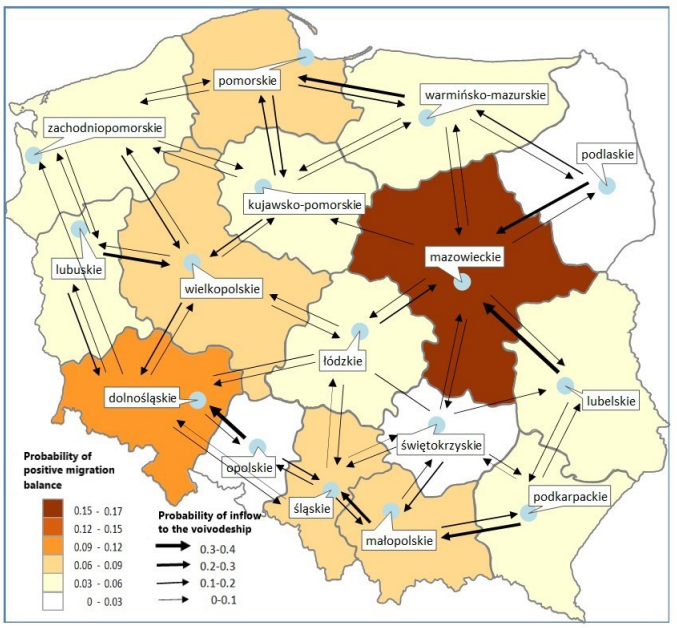


Figure 2: The mechanism of the inter-voivodship migration in Poland in 2020

## 4. Conclusions

The aim of this study was to analyse the scale and spatial range of inter-voivodship migrations of Poland's population in the years 2000, 2010, 2020. The application of the Markov chain model allows for a meticulous evaluation of the population flow between individual voivodships. The results of the study indicate that the most favourable situation remains in the Mazowieckie voivodship, which is the most attractive area of settlement for people from other regions of the country, mainly from the Lubelskie, Podlaskie, and Łódzkie voivodships, and to a lesser extent from Świętokrzyskie and Warmińsko-Mazurskie voivodships. However, during the analysed period, the inflow from the Łódzkie and Warmińsko-Mazurskie voivodships decreased, while for the remaining voivodships the inflow increased. The Mazowieckie voivodship is the most populous region in Poland and the largest in terms of area, and also very spatially diversified in terms of socio-economic development (Struzik 2007). The Dolnośląskie voivodship is ranked next. It is an attractive area for migrants from the Opolskie, Lubuskie, and Wielkopolskie voivodships. The inflow to the Dolnośląskie

voivodship from those voivodships increased in the analysed period. In Figure 2, Wielkopolskie, Śląskie, Małopolskie, and Pomorskie voivodships are marked with the same colour and their probability of stationary distribution is at the level of 0.09-0.12. Migrants come to Wielkopolskie voivodship from Lubuskie, Zachodniopomorskie, Kujawsko-Pomorskie, Dolnośląskie, and to a small extent from Łódzkie. The weakest regions are Opolskie, Świętokrzyskie, and Podlaskie, which are not attractive in terms of migration. Migratory movement of people between voivodships is characterized by an inflow of people mainly from the areas of the neighbouring voivodships.

In reality, the inflow to the most attractive voivodships may be much higher. Due to imperfections in the analysed statistical data, the probabilities of inter-voivodship migration calculated on their basis reflect only the main trends in migration movement in the analysed period.

Certainly, many factors influence the intensity and direction of interprovincial movements. One of them is the pace of social and economic development of a given voivodship, more precisely, the possibility of employment, housing or land prices, cost of living, accessibility to infrastructure, to a large extent family ties, as well as the age of the migrants. According to Kałuża-Kopias (2021), we find that the most mobile group is made up of people aged 25-29, who have the largest spatial extent of migration. The least mobile group includes those aged 35-39 (people who have achieved stabilisation in the labour market and in their family life, and are presumably moving in order to improve their quality of life and rather over short distances) and 65-69 year-olds (retired people).

Another impacting factor for the scale of inter-voivodship flows is migration policy, which is the responsibility of the central government. Local governments have little influence on the government's migration policy instruments (Gońda 2021, Leśniewska, Matuszczyk 2018). However, they can impact the migration decisions of residents by introducing appropriate measures to encourage arrivals and settlement. In the case of all provinces, references to the issue of migration can be found in strategic documents on regional development [3]. In these strategies, the problem of migration is discussed only in the context of the demographic situation in the region as a factor accelerating population ageing and depopulation in the voivodship. Undoubtedly, conducting an active and conscious population policy taking into account the problem of population migration should become an important element of the regional development policy.

Rakowska (2009) points out that since the subject of the study is inter-voivodship migrations, the scale of population inflow will also depend on the size of the voivodship area, the number of its inhabitants, the number of rural communes and towns, its geographical location in relation to other voivodships, as well as socio-economic ties with the nearest surroundings and the level of development and the rate of economic growth.

These subjects are another challenge in migration research.

[3] https://strateg.stat.gov.pl//strategie/wojewodzkie

## References

Azizah, A., Welastika, R., Nur Falah, A., Ruchjana, B., Abdullah, A., (2019). An Application of Markov Chain for Predicting Rainfall Data at West Java using Data Mining Approach, IOP Conf. Ser.: Earth Environ. Sci. 303 012026.

Barra, M., Dahl, F. A., Vetvik, K.G. et al., (2020). A Markov chain method for counting and modelling migraine attacks. Sci Rep 10, 3631.

Bilingsley P., (1961). Statistical Methods in Markov Chain , *Ann. Math. Stat*, 32, pp. 12–40.

Bhat U. N., Miller, K., (2002). Elements of Applied Stochastic Processes.*New York: John Wiley and Sons.*

Constant A., Zimmermann K., (2012). The Dynamics of Repeat Migration: A Markov Chain Analysis. *International Migration Review*, Vol. 46, No. 2, The Center for Migration Studies of New York, Inc., pp. 362-88.

Collins L., (1975). An introduction to Markov chain analysis. *Geo Abstracts,* Norwich, CT.

Gońda M., (2021). Czy imigracja to remedium dla wyludniających się regionów? Cudzoziemcy w dokumentach strategicznych województw łódzkiego i śląskiego, *Polityka Społeczna*, t. XLVIII, No. 9, pp. 15–24.

Ilnicki, D., (2020). Wielkość i kierunki migracji na pobyt stały w województwie wielkopolskim w latach 2002–2017. *Rozwój Regionalny i Polityka Regionalna*, (52), 141–160. https://doi.org/10.14746/rrpr.2020.52.09

Ilnicki, D., Szczyrba, Z., (2021). Migracje wewnętrzne na pobyt stały w makroregionie południowo-zachodnim w latach 2002–2017. *Studia Miejskie*, 36, pp. 25–44. https://doi.org/10.25167/sm.1534

Iosifescu, M., (2007). Finite Markov Processes and Applications. *Dover, New York.*

Jończy R., (2014). Problem nierejestrowanej emigracji definitywnej (emigracji zawieszonej) w badaniu procesów społeczno-gospodarczych na obszarach wiejskich, *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, 360, pp. 11–18.

Józefowicz, K., (2020). Atrakcyjność migracyjna miast i obszarów wiejskich województwa wielkopolskiego. *Space – Society – Economy*, pp. 213–227.

Kałuża-Kopias, D., (2021). Ruch wędrówkowy ludności w województwie łódzkim po 2002 r., *Space – Society – Economy*, (32), pp. 61–81. https://doi.org/10.18778/1733-3180.32.03

Korcelli P., (1997). Alternatywne projekcje zmian demograficznych i migracji w aglomeracjach miejskich, [w:] Korcelli P,.(red.), Aglomeracje miejskie w procesie transformacji, *Zeszyty IGiPZ PAN 45*, Warszawa, pp. 5–22.

Lesińska, M., Matuszczyk, K., (2018). Działania samorządów wobec migracji w kontekście zmian demograficznych. Przykład trzech polskich województw. *Studia regionalne i lokalne*, No. 3(77), pp. 64–82

Privault N.,(2018). Understanding Markov chains: Examples and applications. *Springer.*

Rahimipour Sheikhani Nejad, M.A., Nasiri Jan Agha, F. and Khatami, S. S., (2018). Monitoring and predicting land cover changes in the coastal areas for optimal land allocation (Case study: Chaf and Chamkhaleh, Guilan, Iran), *International Journal of Development and Sustainability*, Vol. 7 No. 3, pp. 973–985.

Rakowska, B., Rakowski W., (2009). Województwo mazowieckie jako obszar napływu i odpływu ludności. *Rocznik Żyrardowski*, 7, pp. 343–369.

Rogers, A., (1968). Matrix Analysis of Interregional Population Growth and Distribution. *University of Kalifornia.*

Romeu, J., (2020). A Markov Chain Model for Covid-19 Survival Analysis. Syracuse University. https://www. researchgate. net/profile/Jorge–Romeu.

Rosner, A., (2014). Migracje wewnętrzne i ich związek z przestrzennym zróżnicowaniem rozwoju społeczno-gospodarczego wsi, *Wieś i Rolnictwo*, 1 (162), pp. 63–79.

Roszko-Wójtowicz, E., (2018). Migracje międzywojewódzkie w Polsce w latach 2010–2016 a jakość życia, *Ekspertyzy i Opracowania*, 75, pp. 1–13.

Sempewo, J., Kyokaali, L., (2016). Prediction of the Future Condition of a Water Distribution Network Using a Markov Based Approach: A Case Study of Kampala Water, *Procedia Engineering*, Vol.154, pp. 374–383.

Sojka, E., (2018). Odległość geograficzna i miernik rozwoju społeczno-gospodarczego a wielkość migracji w województwie śląskim, *Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach*, 353, pp. 73–88.

Struzik, A., (2007). Współczesne problemy rozwoju województwa mazowieckiego, *Rocznik Żyrardowski*, tom V, WSRL, Żyrardów.

Śleszyński, P., (2005). Różnice liczby ludności ujawnione w Narodowym Spisie Powszechnym 2002, *Przegląd Geograficzny*, 77(2), pp. 193–212.

Śleszyński, P., (2011). Oszacowanie rzeczywistej liczby ludności gmin województwa mazowieckiego z wykorzystaniem danych ZUS, *Studia demograficzne*, 2, pp. 35–58.

Śleszyński, P., Heffner, K., Solga, B., Wiśniewski, R. (2018). Perspektywa geografii i studiów regionalnych w badaniach nad migracjami. [W:] A. Horolets, M. Lesińska, M. Okólski (red.), Raport o stanie badań nad migracjami w Polsce po 1989 roku. *Komitet Badań nad Migracjami PAN*, pp. 174–210.

Usman, M. E., Faiz, M.B., (2015). Markov chains analysis and mechanism of migration in Indonesia in the period 1980-2010, *Journal of Engineering and Applied Sciences*, 10 (22). pp. 17256–17264.

Willekens, F., (1999). Probability models of migration: complete and incomplete data. *Journal of Demography* 7(1).

Venkatesan, G., Sasikala, V., (2018). Statistical analysis on migrants using Markov chain model. *The International Journal of Creative Research Thoughts,* Vol. 6(1).

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 109–124,* https://doi.org/10.59170/stattrans-2024-007
Received – 13.02.2023; accepted – 03.10.2023

# Implementation of K-Nearest Neighbor using the oversampling technique on mixed data for the classification of household welfare status

**Nur Mutmainnah Djafar**[1], **Achmad Fauzan**[2]

## Abstract

Welfare is closely related to poverty and the socio-economic disparities in a society. Based on data from the Central Bureau of Statistics, Kulon Progo in Indonesia had the highest poverty rate in the province of the Special Region of Yogyakarta; an increasing trend was observed every year from 2019 to 2021; Kulon Progo also had a low poverty line (after Gunung Kidul) compared to other regencies/cities in this province. This study aimed to classify the household welfare status in Kulon Progo in March 2021 using the K-Nearest Neighbor (KNN) method. Since imbalance was found between the poor and non-poor classes, an oversampling technique was employed. Imbalanced data affect classification, particularly when predicting the results of the classification. The following oversampling techniques were employed in this study: Random Oversampling (RO), the Adaptive Synthetic (ADASYN) and the Synthetic Minority Oversampling Technique (SMOTE). It was found that, of the three techniques, RO was the most efficient with $k = 5$, which yielded the best performance in terms of sensitivity, specificity, the G-mean, and accuracy reaching 0.643, 0.805, 0.719, and 78.873%, respectively. Therefore, it can be concluded that the classification model performed well enough to classify household welfare status, especially among the poor (minority class).

**Key words:** ADASYN, KNN, random oversampling, SMOTE, welfare.

## 1. Introduction

The development and progress made these days have become one of the challenges for countries to also develop and make progress over time, so these countries must make efforts to improve the welfare for their community. Welfare is a benchmark in social life, in which the society members are prosperous. It can be measured from the

[1] Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Indonesia. E-mail: nur.djafar@students.uii.ac.id.

[2] Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Indonesia. E-mail: achmadfauzan@uii.ac.id. ORCID: https://orcid.org/0000-0002-0533-5518.

society's health, economic conditions, happiness, and quality of life (Suud & Harsono, 2006). Welfare, however, has a close association with poverty and socio-economic disparities in society. Poverty is an economic, material, and physical inability to meet basic needs, including both food and non-food items, as measured by expenditure.

Poverty alleviation is the main responsibility for all central and regional governments throughout the world. Therefore, this is also the responsibility of the government of Indonesia, including the government of Kulon Progo Regency and the Province of the Special Region of Yogyakarta (DIY), Indonesia. Kulon Progo is one of the regencies in DIY which has 12 sub-districts and 88 villages. This regency has an area of 586.28 km2 and a population in the first half of 2021 of 442,838 people. In the last three years, the poverty rate in Kulon Progo had an increasing trend every year from 2019 to 2021, reaching 18.38%, which means that there was a total of 81,140 poor people in this regency. In addition, in 2021 Kulon Progo had the highest poverty rate compared to other regencies/city in DIY (BPS-Statistics of DI Yogyakarta Province, 2021).

In addition to using the monthly per capita consumption, according to research by Suryadarma et al. (2005), poverty can also be classified using other important indicators, namely information on asset ownership such as housing, employment status, family health, livestock ownership, food consumption, etc. In addition, it can also be identified from the Head of the Household (KRT), including age, gender, marital status, education level, the number of dependents, and income of the head of the household.

Based on the above-mentioned description, welfare-related issues, particularly poverty in Kulon Progo, is an interesting topic of discussion. The classification of household welfare status uses many indicators as influencing factors. A total of 17 variables were used in this study. The classification method was K-Nearest Neighbor (KNN), i.e. a non-parametric classification based on the closest neighbor according to the distance-based k value (Haseela H A, 2022). The KNN method was used in this study because KNN has several benefits, including a fast, simple, and effective training process although involving large training data. Such simplicity was used in this study to determine the results of classification using the research data. In addition, KNN can also classify data which contain categorical and numerical data.

The data on the welfare status of Kulon Progo showed that very few households were classified as poor compared to those classified as non-poor. The resulting classification, however, tends to classify the majority class well, but it has a poor performance in predicting the minority class, thus causing (Jian et al., 2016). One of the solutions to such imbalance is oversampling. Oversampling is a method to oversample the minority class data to be close to or equal to the majority class (Chawla, 2005). There are various oversampling methods, some of which were used in this research, including Random Oversampling (RO), Adaptive Synthetic Sampling (ADASYN), and Synthetic

Minority Oversampling Technique (SMOTE). The three oversampling techniques have different procedures and all of them were applied to the imbalanced data to determine their effectiveness. RO duplicates existing data, ADASYN uses weighted distribution, and SMOTE generates replicated data.

The study aimed to determine the general description of household welfare status data and to determine the comparison of the results of the KNN classification on data without oversampling and data with different oversampling techniques, namely RO, ADASYN, and SMOTE.

## 2. Method

### 2.1. Data

The study used data from the National Socioeconomic Survey (SUSENAS) by the Central Bureau of Statistics in Kulon Progo Regency, Indonesia in March 2021 accessed from http://silastik.bps.go.id. SUSENAS was carried out by direct interviews or self-administered questionnaires.

There were 18 variables used in this study, consisting of 17 independent variables defined as X and 1 dependent variable defined as Y, i.e. household welfare status which was classified into two categories, namely poor (0) and non-poor (1). The variables used are presented in Table 1.

| No. | Variables | Type of data |
|---|---|---|
| 1. | Welfare status | Nominal |
| 2. | Age of head of household | Numerical |
| 3. | Family size | Numerical |
| 4. | Area of house | Numerical |
| 5. | Gender of head of household | Categorical |
| 6. | Marital status of head of household | Categorical |
| 7. | Main source of household income | Categorical |
| 8. | Type of home ownership | Categorical |
| 9. | Latest education of head of household | Categorical |
| 10. | Bank account ownership of head of household | Categorical |
| 11. | Head of household works in the last one week | Categorical |
| 12. | Head of household's health issues in the last one month | Categorical |
| 13. | Main light source | Categorical |
| 14. | Main source of energy used for cooking | Categorical |
| 15. | Main source of drinking water | Categorical |
| 16. | Main source of water for washing | Categorical |
| 17. | Mobile phone ownership of head of household | Categorical |
| 18. | Laptop/notebook ownership of head of household | Categorical |

### 2.2. K-Nearest Neighbors (KNN)

KNN was developed by Evelyn Fix and Joseph Hodges in 1951. KNN is a non-parametric classification method based on the closest neighbor according to a distance-based k value (Haseela H A, 2022). KNN classification requires a distance that is in line with the type of research data (Wu et al., 2008). Several papers on the KNN technique and its development included Alsammak et al. (2020) conducting research to improve the performance of the K-Nearest Neighbor (KNN) classifier to satisfy emerging big data requirements. Kirtania et al. (2020) is working on a new adaptive KNN classifier for addressing imbalances in Magnetic Resonance Imaging (MRI) brain. Awotunde et al. (2022) investigated the feature choice based KNN Model for Rapid Software Defect Prediction. Hoque et al. (2021) created a KNN-DK classifier, which is a modified KNN classifier with dynamic *k* nearest neighbors.

Wilson & Martinez (1997a) explained that one of the distances for numerical and categorical data types is the Heterogeneous Euclidean-Overlap Metric (HEOM). HEOM handles both continuous and nominal attributes with overlap metric for nominal attributes and normalize Euclidean distance for linear attributes (ChitraDevi et al., 2012; Tusyakdiah, 2021). A heterogeneous distance function that uses different attribute distance functions on diverse categories of features has been used to address the issues of applications with continuous and nominal attributes. The unique technique is the overlap metric for combined nominal attributes and normalized Euclidean distance for linear features or numerical data (Dalatu & Midi, 2020). Numerical data are calculated using normalized Euclidean distance (Randall et al., 2000; Wilson & Martinez, 1997), written in equation (1).

$$D(x_{it}, z_{jt}) = \frac{|x_{it} - w_{jt}|}{\max(x_{it}) - \min(x_{it})}, \qquad t = 1,2,3, \dots, m_n \tag{1}$$

Categorical data are calculated using overlap metrics, written in equation (2).

$$D^2(x_{it}, z_{jt}) = \begin{cases} 0 & x_{it} = z_{jt} \\ 1 & x_{it} \neq z_{jt} \end{cases}, \quad t = 1,2,3, \dots, m_c \tag{2}$$

After the distances for both numerical and categorical data have been obtained, the square root of the sum of the two distances was calculated to obtain the HEOM distance in equation (3).

$$D_{HEOM}(x_i, z_j) = \sqrt{\sum_{t=1}^{m_n} D(x_{it}, z_{jt}) + \sum_{t=1}^{m_c} D^2(x_{it}, z_{jt})} \tag{3}$$

with D: distance, $\boldsymbol{x_{it}}$: training data value, $\boldsymbol{w_{jt}}$ : data testing value, a: data variable to-$\boldsymbol{i}$, max: the maximum value of each numeric variable, min: the minimum value of each numeric variable, $\boldsymbol{m_n}$: numeric data type, and $\boldsymbol{m_c}$: categorical data type. The use of

HEOM distance will remove the impacts of arbitrary nominal value ordering, but it is an overly simplistic approach to dealing with nominal attributes that fails to make use of extra information offered by nominal attribute values that can aid in generalization (Wilson & Martinez, 1997). The usage of HEOM distance fits this study extremely well, in this case, since the data in this study are mixed data, as seen in Table 1.

### 2.3. Class Imbalance

Class imbalance is one of the problems in data mining. It is a condition where the minority class is very small compared to the majority class (Ren et al., 2017). In a classification with imbalanced data, the accuracy of the minority class tends to be low due to the dominance of the majority class, thus causing biases (Jian et al., 2016). In addition, class imbalance and noise can affect the quality of data in classification performance (Gao et al., 2014).

One of the solutions to class imbalance is resampling. Resampling is a preprocessing technique to balance data distribution to reduce the effect of class imbalance. There are three types of resampling, namely oversampling, undersampling, and hybrid, which combines both over and undersampling (Jian et al., 2016). Oversampling is used because of its benefits, i.e. adding data to the minority class to prevent the loss of data information.

### 2.4. Oversampling

Oversampling is a method to oversample the minority class data to be close to or equal to the majority class (Chawla, 2005). There are various oversampling methods, some of which were used in this research, including Random Oversampling (RO), Adaptive Synthetic Sampling (ADASYN), and Synthetic Minority Oversampling Technique (SMOTE).

#### 2.4.1. Random Oversampling (RO)

Random Oversampling is a technique to randomly add data from the minority class to the training data, in which the addition process is repeated until the data in the minority class are equal in number to those in the majority class. The difference between the majority and minority classes is first calculated. Then, the repetition is randomly done as many times as the difference resulting from the calculation and added to the dataset.

#### 2.4.2. Adaptive Synthetic Sampling (ADASYN)

ADASYN uses the weighted distribution of the data in the minority class, in which synthetic data are generated from the minority class (Rahayu et al., 2017). The steps for performing ADASYN are as follows (He et al., 2008).

i. Calculating the degree of data imbalance.

$$d_c = \frac{m_s}{m_l} \tag{4}$$

with $d \in [0.1]$. $m_s$ is the number of minority class data and $m_l$ is the number of majority class data.

ii. If $d_c < d_{th}$, where $d_{th}$ is the threshold for the maximum degree of class imbalance, so:

1. Calculating the amount of synthetic data to be generalized for the minority class.

$$G = (m_l - m_s) \times \beta \tag{5}$$

where $\beta \in [0,1]$ is the parameter used to set the level, balance expected after the synthetic data has been generalized. $\beta = 1$ means completely balanced data after the generalization.

2. For each $x_i \in$ minority class, determining the k-nearest neighbor based on *HEOM* distance in equation (3) in n dimensional space and calculating ratio ($r_i$).

$$r_i = \frac{\Delta_i}{K}, i = 1, \dots, m_s \tag{6}$$

with $r_i$: ration, $\Delta_i$: the number of samples in KNN but from data that include all classes except the minority and $K$: the number $k$ in KNN, and interval of $r_i \in [0,1]$,

3. Normalizing $r_i$

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m} r_i} \tag{7}$$

4. Calculating the amount of synthetic data to be generated for each $x_i$ in the minority class.

$$g_i = \hat{r}_i \times G \tag{8}$$

$G$ is the total amount of synthetic data to be generated for the minority class in equation (5).

5. For each $x_i$ in the minority class, generating synthetic data as many times as $g_i$ by making repetition from 1 to $g_i$ with the steps as follows:

iii. Randomly selecting one of the data in the minority class from $x_{knn}$ in data $x_i$

iv. Generating synthetic data by equation (9).

$$x_{adasyn} = x_i + (x_{knn} - x_i) \times \lambda \tag{9}$$

where $\lambda$ is random [0,1]

### 2.4.3. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is done by adding more data in the minority class by generating synthetic or artificial data. The synthetic data are generated based on the attributes from the k-nearest neighbor. Several studies related to oversampling with the SMOTE technique

include Elreedy & Atiya (2019), Li et al. (2021), Noorhalim et al. (2019), and Srinilta & Kanharattanachai (2021). The steps for performing SMOTE are as follows (Chawla, 2005):

i. Determining the data to be replicated ($x_i$) from a randomly selected minority class.

ii. Determining the value of $k$ (the number of the nearest neighbors), then calculating the distance from data $x_i$ to the nearest neighbor data ($x_{knn}$) in the same minority class.

iii. For each $x_{knn}$ that is selected, calculating the difference between $x_i$ and $x_{knn}$, then multiplying the difference by a random number [0,1], and adding it to the features under study.

$$x_{sintesis} = x_i + (x_{knn} - x_i) \times \delta \tag{10}$$

where $\delta$ is random [0,1]

## 2.5. Validation Technique

In data mining, there are many techniques to measure the performance of an algorithm, one of which is confusion matrix, which is a binary table, where classes are divided into 2 categories (Pramana et al., 2018). Confusion matrix is a table that states the amount of testing data correctly classified, and the amount of testing data misclassified (Indriani, 2014).

**Table 2:** Confusion Matrix

| Real Classes | Predicted Classes | |
|---|---|---|
| | Poor | Non-Poor |
| Poor | TP | FN |
| Non-Poor | FP | TN |

True Positive (TP) is the number of poor classes predicted to be poor; False Positive (FP) is the number of non-poor classes predicted to be poor; False Negative (FN) is the number of poor classes predicted to be non-poor; True Negative (TN) is the number of non-poor classes predicted to be poor.

Based on the confusion matrix, several evaluation metrics including accuracy, sensitivity, specificity, and G-mean can be derived.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \tag{10}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \tag{13}$$

Accuracy is a comparison value between correctly classified data and real data. This measurement is used to measure the correctness of the classification (Hamel, 2009). The higher the accuracy value, the better the resulting classification (Widayati et al., 2021). The value of classification accuracy needs to be increased as a measure of standard criteria, especially in cases of class imbalance. Because it will produce good accuracy only for the majority class, while the resulting predictions will be dire for the minority class (Pangastuti, 2018), evaluation of the performance of the method as a whole can be done using the geometric mean (Kubát & Matwin, 1997).

Sensitivity is a value that shows the results of the actual data classification, which is a positive class, and the predicted value is also a positive class (Hamel, 2009). The higher the sensitivity value, the less likely the results of the positive class classification are wrong (Zhu et al., 2010). Specificity is a value that shows the results of the actual data classification, which is labeled negative, and the predicted value is also labeled negative (Jahangiri et al., 2020). The higher the specificity value, the better the classification performance for making predictions because it has low false positives (Maxim et al., 2014). A good classification is a classification that has high sensitivity and specificity values (Zhu et al., 2010). The G-mean is the geometric average value used to measure overall performance. Poor classification results will produce a small G-mean value (Bekkar et al., 2013).

## 3. Results and Discussion

Based on BPS-Statistics of DI Yogyakarta Province data, there are 71 households in the poor category and 638 households in the non-poor category, or it can be said that there is an imbalance in the data. KNN classification was performed by dividing 80% for the training data and 20% for the testing data. Oversampling was performed using 567 training data.

### 3.1. RO, ADASYN, and SMOTE Oversampling

RO was done by making 447 repetitions by randomly taking data from the minority class. The ADASYN was performed using $k = 11$ and $\beta = 1$ (with the expected balance of 100%) using the HEOM distance to determine the nearest neighbor. Synthesized data were generated by taking from the poor class, resulting in an addition of 450 data. SMOTE was performed on the data from the minority class using $k = 11$ and generated 8 times from the total observations of the poor class (8×57) using the HEOM distance to determine the nearest neighbor. The replicated data were equivalent to the non-poor class, namely 399 data. Illustration of data results after oversampling and without oversampling is presented in Figure 1.

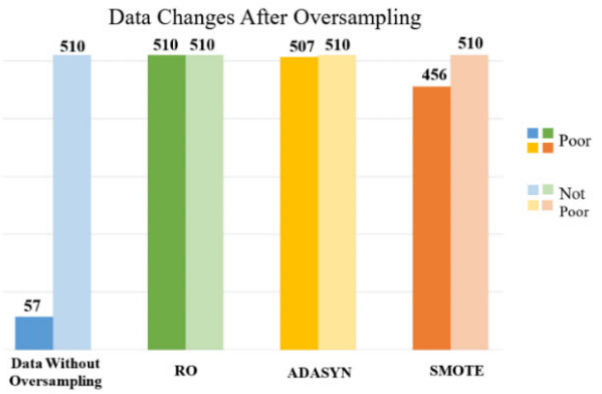


**Figure 1:** Illustration of data after oversampling.

### 3.2. KNN with HEOM Distance

The classification using *KNN* with *HEOM* distance was simulated using the data as shown in Table 3.

**Table 3:** Simulation Data

| Data | $x_{i1}$ | $x_{i2}$ | $x_{i3}$ | $x_{i4}$ | … | $x_{i15}$ | $x_{i16}$ | $x_{i17}$ | Y |
|---|---|---|---|---|---|---|---|---|---|
| Training data 1 ($x_{1t}$) | 65 | 1 | 56 | 0 | … | 2 | 0 | 1 | 0 |
| Training data 2 ($x_{2t}$) | 57 | 2 | 96 | 0 | … | 4 | 0 | 1 | 1 |
| Training data 3 ($x_{3t}$) | 35 | 1 | 87 | 0 | … | 3 | 0 | 1 | 1 |
| Training data 4 ($x_{4t}$) | 62 | 0 | 54 | 0 | … | 4 | 0 | 1 | 0 |
| Testing data 1 ($z_{1t}$) | 45 | 1 | 184 | 1 | … | 4 | 0 | 1 | ? (predicted) |

Based on Table 3, the household welfare status in testing data 1 ($z_{1t}$) was predicted, whether it was classified as either poor or non-poor class. The distances for the numerical data, namely $X_1$, $X_2$, and $X_3$ were firstly calculated. The distance calculation used the normalized Euclidean distance according to equation (1).

The distance between testing data ($z_{1t}$) and training data 1 ($x_{1t}$)

$$\sum_{t=1}^{3} D^2(x_{1t}, z_{1t}) = \left(\frac{|65-45|}{97-21}\right)^2 + \left(\frac{|1-1|}{5-0}\right)^2 + \left(\frac{|56-184|}{240-9}\right)^2 = 0.57$$

The distance between testing data ($z_{1t}$) and training data 2 ($x_{2t}$)

$$\sum_{t=1}^{3} D^2(x_{2t}, z_{1t}) = \left(\frac{|57-45|}{97-21}\right)^2 + \left(\frac{|2-1|}{5-0}\right)^2 + \left(\frac{|96-184|}{240-9}\right)^2 = 0.21$$

The distance between testing data $(z_{1t})$ and training data 3 $(x_{3t})$

$$\sum_{t=1}^{3} D^2(x_{3t}, z_{1t}) = \left(\frac{|35-45|}{97-21}\right)^2 + \left(\frac{|1-1|}{5-0}\right)^2 + \left(\frac{|87-184|}{240-9}\right)^2 = 0.193$$

The distance between testing data $(z_{1t})$ and training data 4 $(x_{4t})$

$$\sum_{t=1}^{3} D^2(x_{4t}, z_{1t}) = \left(\frac{|62-45|}{97-21}\right)^2 + \left(\frac{|0-1|}{5-0}\right)^2 + \left(\frac{|54-184|}{240-9}\right)^2 = 0.407$$

Once the distance for the numerical data had been obtained, the distance for the categorical data on variables $X_4$ to $X_{17}$ had to be calculated using the overlap metric method by observing the incompatibility of the two vectors, i.e. if vector $x_i$ was different from vector $z_j$, then the value = 1.

The *HEOM* distance was calculated by calculating distances $(z_{1t})$ and $(x_{it})$ in each categorical variable based on equation (2).

The distance between testing data $(z_{1t})$ and training data 1 $(x_{1t})$

$$\sum_{t=4}^{17} D^2\,(x_{1t}, z_{1t}) = [D^2(x_{14}, z_{14}) + D^2(x_{15}, z_{15}) + \cdots + D^2(x_{1,17}, z_{1,17})$$
$$= [D^2(0,1) + D^2(1,3) + D^2(0,0) + \cdots + D^2(1,0)] = [1^2 + 1^2 + 0^2 + \cdots + 1^2] = 9$$

The distance between testing data $(z_{1t})$ and training data 2 $(x_{2t})$

$$\sum_{t=4}^{17} D^2(x_{2t}, z_{1t}) = [D^2(x_{24}, z_{14}) + D^2(x_{25}, z_{15}) + D^2(x_{26}, z_{16}) + \cdots + D^2(x_{2,17}, z_{1,17})$$
$$= [D^2(0,1) + D^2(3,3) + D^2(0,0) + \cdots + D^2(1,0)] = [1^2 + 0^2 + 0^2 + \cdots + 1^2] = 6$$

The distance between testing data $(z_{1t})$ and training data 3 $(x_{3t})$

$$\sum_{t=4}^{17} D^2(x_{3t}, z_{1t}) = [D^2(x_{34}, z_{14}) + D^2(x_{35}, z_{15}) + D^2(x_{36}, z_{16}) + \cdots + D^2(x_{3,17}, z_{1,17})$$
$$= [D^2(0,1) + D^2(1,3) + D^2(0,0) + \cdots + D^2(1,0)] = [1^2 + 1^2 + 0^2 + \cdots + 1^2] = 7$$

The distance between testing data $(z_{1t})$ and training data 4 $(x_{4t})$

$$\sum_{t=4}^{17} D^2(x_{4t}, z_{1t}) = [D^2(x_{44}, z_{14}) + D^2(x_{45}, z_{15}) + D^2(x_{46}, z_{16}) + \cdots + D^2(x_{4,17}, z_{1,17})$$
$$= [D^2(0,1) + D^2(0,3) + D^2(0,0) + \cdots + D^2(1,0)] = [1^2 + 1^2 + 0^2 + \cdots + 1^2] = 7$$

After the distance for the numerical data had been obtained using the normalized difference and for the categorical data using the overlap metric, the two distances were combined to obtain the overall distance $(x_i, z_j)$ according to equation (3).

$$\text{HEOM Dist } (x_1, z_1) = \sqrt{0.57 + 9} = 3.094$$
$$\text{HEOM Dist } (x_2, z_1) = \sqrt{0.21 + 6} = 2.492$$
$$\text{HEOM Dist } (x_3, z_1) = \sqrt{0.193 + 7} = 2.682$$
$$\text{HEOM Dist } (x_4, z_1) = \sqrt{0.407 + 7} = 2.722$$

After the distance of each object had been obtained, the classification using *KNN* was done with the specified $k$. The following is an illustration for $k = 1$ and $k = 3$ using R Programming.

**Table 4:** Classification Results of Simulation

| Data | Y (Welfare Status) | Distance | $k = 1$ | $k = 3$ |
|---|---|---|---|---|
| $(x_1, z_1)$ | Poor | 3.094 | | |
| $(x_2, z_1)$ | Non-poor | 2.492 | Non-poor | Non-poor |
| $(x_3, z_1)$ | Non-poor | 2.682 | | Non-poor |
| $(x_4, z_1)$ | Poor | 2.722 | | Poor |

To classify using KNN, the classification results were obtained from the training data with the nearest distance to the testing data. Using $k = 1$, the classification results were non-poor because the nearest distance was training data 2 with category 1 (non-poor). Using $k = 3$, the classification results were still non-poor because the nearest distance was training data 2 and 3 with category 1 (non-poor), while training data 4 with category 0 (poor) only had 1 data. In other words, the dominant class was used as the classification result.

The classification using KNN was carried out four times with four data, namely data without any oversampling treatment (imbalanced data), data with RO, ADASYN, and SMOTE treatment. Validation was carried out with various values of $k$ (3,5,7,9) based on the confusion matrix obtained, by using the R program the results are presented in Table 5.

**Table 5:** Classification Results

| $k$ | Data | Sensitivity | Specificity | G-mean | Accuracy |
|---|---|---|---|---|---|
| 3 | Data without oversampling | 0.071 | 1 | 0.267 | 90.845% |
| | RO | 0.500 | 0.867 | 0.658 | 83.099% |
| | ADASYN | 0.500 | 0.836 | 0.647 | 80.281% |
| | SMOTE | 0.429 | 0.820 | 0.593 | 78.170% |
| 5 | Data without oversampling | 0 | 1 | 0 | 90.141% |
| | RO | 0.643 | 0.805 | 0.719 | 78.873% |
| | ADASYN | 0.500 | 0.852 | 0.653 | 81.691% |
| | SMOTE | 0.500 | 0.773 | 0.622 | 74.648% |
| 7 | Data without oversampling | 0 | 1 | 0 | 90.141% |
| | RO | 0.714 | 0.734 | 0.724 | 73.240% |
| | ADASYN | 0.500 | 0.836 | 0.647 | 80.282% |
| | SMOTE | 0.500 | 0.758 | 0.616 | 73.240% |
| 9 | Data without oversampling | 0 | 1 | 0 | 90.141% |
| | RO | 0.786 | 0.664 | 0.722 | 67.606% |
| | ADASYN | 0.500 | 0.820 | 0.640 | 78.873% |
| | SMOTE | 0.643 | 0.688 | 0.665 | 68.310% |

It was insufficient to validate the results of the classification through the accuracy due to the imbalanced data between the poor and non-poor classes. Based on Table 5, the data without any oversampling treatment obtained high accuracy and sensitivity, but very low specificity. This means that the model was very bad for the non-poor class classification, yet very good for the poor class classification. Thus, the data without any oversampling treatment were not good in this classification because the results of the classification were dominated by the accuracy of the minority class.

The G-mean is one of the best measurements for evaluating classification, especially in class imbalances in data (Pristyanto et al., 2018). Based on Table 5, all G-mean values in the data that were not oversampled resulted in low values with $k$ values of 3, 5, 7, and 9. Meanwhile, if using data that had been treated RO, ADASYN, and SMOTE with a value of $k$=5 produces a G-mean of 0.719, 0.653, and 0.622, which means that with balanced data, the resulting classification is good enough for the poor and non-poor classes.

Based on Table 5, it can also be said that with oversampling, the RO oversampling technique with a value of $k$=5 gives the best results when viewed from the G-mean, accuracy, and ease of forming nearest neighbors. With a G-mean value and accuracy of 0.719 and 78.873%. After calculating the accuracy, sensitivity, specificity, and G-mean, the best classification was generated with the RO-treated data. The model can classify object classes well, especially in the welfare status classification. The findings of this study are also consistent with the findings of several other studies, including Akbar et al. (2019), Hussain et al. (2022), and Xin & Rashid (2021) which specify the performance of k-NN sensitivity values more precisely. Furthermore, research from Islam et al. (2022) and Shi ( 2020) dealing with imbalance data with oversampling approaches raises the value of precision.

## 4. Conclusions

This research begins by performing oversampling techniques with three methods, including Random Oversampling (RO), Adaptive Synthetic Sampling (ADASYN), and Synthetic Minority Oversampling Technique (SMOTE), and comparing with data without oversampling techniques. Since the data used are mixed (numeric and categorical), the Heterogeneous Euclidean-Overlap Metric (HEOM) distance is more appropriate for calculating the distance. Then, from the oversampling results, classification is carried out using the k-nearest neighbors (KNN) method with various simulated $k$ values. With the division of 80% training data and 20% testing data, the classification using KNN with $k$ = 5 and the HEOM distance produced the best results on data with a RO treatment. This is evident from the sensitivity, specificity, G-mean, and accuracy i.e. 0.643, 0.805, 0.719, and 78.873% respectively. This means that the classification model was quite good in classifying welfare status, especially in the minority class (poor class).

## Acknowledgement

We thank all the parties who have provided support and funding for this research.

## References

Akbar, S., Hayat, M., Kabir, M., and Iqbal, M., (2019). iAFP-gap-SMOTE: An Efficient Feature Extraction Scheme Gapped Dipeptide Composition is Coupled with an Oversampling Technique for Identification of Antifreeze Proteins. *Letters in Organic Chemistry*, *16*(4), pp. 294–302. https://doi.org/10.2174/1570178615666180816101653

Alsammak, I. L. H., Sahib, H. M. A., and Itwee, W. H., (2020). An Enhanced Performance of K-Nearest Neighbor (K-NN) Classifier to Meet New Big Data Necessities. *IOP Conference Series: Materials Science and Engineering*, *928*(3). https://doi.org/10.1088/1757-899X/928/3/032013

Awotunde, J. B., Misra, S., Adeniyi, A. E., Abiodun, M. K., Kaushik, M., and Lawrence, M. O., (2022). A Feature Selection-Based K-NN Model for Fast Software Defect Prediction. In O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, & C. Garau (Eds.), *Computational Science and Its Applications – ICCSA 2022 Workshops,* pp. 49–61. Springer International Publishing.

Bekkar, M., Djemaa, H. K., and Alitouche, T. A., (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, *3*, pp. 27–38.

BPS-Statistics of DI Yogyakarta Province, (2021). *Persentase Penduduk Miskin menurut Kabupaten/Kota di Provinsi DI Yogyakarta (Persen), 2009-2021*.

Chawla, N. V., (2005). Data Mining for Imbalanced Datasets: An Overview. In L. Maimon Oded and Rokach (Ed.), *Data Mining and Knowledge Discovery Handbook* (pp. 853–867). Springer US. https://doi.org/10.1007/0-387-25465-X_40

ChitraDevi, N., Palanisamy, V., Baskaran, K., and Prabeela, S., (2012). A Novel Distance for Clustering to Support Mixed Data Attributes and Promote Data Reliability and Network Lifetime in Large Scale Wireless Sensor Networks. *Procedia Engineering*, *30*, pp. 669–677. https://doi.org/10.1016/j.proeng.2012.01.913

Dalatu, P. I., Midi, (2020). Modified Statistical Approach for Data Preprocessing to Improve Heterogeneous Distance Functions. In *Malaysian Journal of Mathematical Sciences* (Vol. 14, Issue 2).

Elreedy, D., Atiya, A. F., (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, pp. 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Gao, K., Khoshgoftaar, T. M., and Wald, R., (2014). Combining Feature Selection and Ensemble Learning for Software Quality Estimation. *The Florida AI Research Society*.

Hamel, L., (2009). Model Assessment with ROC Curves. In *Encyclopedia of Data Warehousing and Mining, Second Edition,* pp. 1316–1323. IGI Global. https://doi.org/10.4018/978-1-60566-010-3.ch204

Haseela H A., (2022). Hybrid Method for Image Classification. *EPRA International Journal of Research and Development (IJRD)*, *7*(2), pp. 59–61. https://doi.org/10.36713/epra2016

He, H., Bai, Y., Garcia, E. A., and Li, S., (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hoque, N., Bhattacharyya, D. K., and Kalita, J. K., (2021). KNN-DK: A Modified K-NN Classifier with Dynamic k Nearest Neighbors. In J. C. Bansal, L. C. C. Fung, M. Simic, & A. Ghosh (Eds.), *Advances in Applications of Data-Driven Computing,* pp. 21–34. Springer Singapore. https://doi.org/10.1007/978-981-33-6919-1_2

Hussain, L., Lone, K. J., Awan, I. A., Abbasi, A. A., and Pirzada, J.-R., (2022). Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. *Waves in Random and Complex Media*, *32*(3), pp. 1079–1102. https://doi.org/10.1080/17455030.2020.1810364

Indriani, A., (2014). Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) Yogyakarta*. www.bluefame.com,

Islam, A., Belhaouari, S. B., Rehman, A. U., and Bensmail, H., (2022). KNNOR: An oversampling technique for imbalanced datasets. *Applied Soft Computing*, *115*, 108288. https://doi.org/10.1016/j.asoc.2021.108288

Jahangiri, M., Jahangiri, M., and Najafgholipour, M., (2020). The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (COVID-19) in different provinces of Iran. *Science of The Total Environment*, *728*, 138872. https://doi.org/10.1016/j.scitotenv.2020.138872

Jian, C., Gao, J., and Ao, Y., (2016). A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble. *Neurocomput.*, *193*(C), pp. 115–122. https://doi.org/10.1016/j.neucom.2016.02.006

Kirtania, R., Mitra, S., and Shankar, B. U., (2020). A novel adaptive k-NN classifier for handling imbalance: Application to brain MRI. *Intelligent Data Analysis*, *24*, pp. 909–924. https://doi.org/10.3233/IDA-194647

Kubát, M., Matwin, S., (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *International Conference on Machine Learning*.

Li, J., Zhu, Q., Wu, Q., and Fan, Z., (2021). A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences*, *565*, pp. 438–455. https://doi.org/10.1016/j.ins.2021.03.041

Maxim, L. D., Niebo, R., and Utell, M. J., (2014). Screening tests: a review with examples. *Inhalation Toxicology*, *26*(13), pp. 811–828. https://doi.org/10.3109/08958378.2014.955932

Noorhalim, N., Ali, A., and Shamsuddin, S. M., (2019). Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE. In L.-K. Kor, A.-R. Ahmad, Z. Idrus, & K. A. Mansor (Eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017),* pp. 19–30. Springer Singapore.

Pangastuti, S. S., (2018). *Perbandingan Metode Ensemble Random Forest dengan Smote-Boosting dan Smote-Bagging pada Klasifikasi Data Mining untuk Kelas Imbalance (Studi Kasus: Data Beasiswa Bidikmisi Tahun 2017 di Jawa Timur)*. Institut Teknologi Sepuluh Nopember.

Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., and Nooraeni, R., (2018). *Data Mining dengan R: Konsep Serta Implementasi*. IN MEDIA.

Pristyanto, Y., Pratama, I., and Nugraha, A. F., (2018). Data level approach for imbalanced class handling on educational data mining multiclass classification. *2018 International Conference on Information and Communications Technology (ICOIACT)*, pp. 310–314. https://doi.org/10.1109/ICOIACT.2018.8350792

Rahayu, S., Bharata Adji, T., Akhmad Setiawan, N., and Teknik Elektro dan Teknologi Informasi, D., (2017). Penghitungan k-NN pada Adaptive Synthetic-Nominal (ADASYN-N) dan Adaptive Synthetic-kNN (ADASYN-kNN) untuk Data Nominal-Multi Kategori. *Ktrl.Inst (J.Auto.Ctrl.Inst)*, *9*(2).

Randall, D., And, W., and Martinez, T. R., (2000). An Integrated Instance-Based Learning Algorithm. *Computational Intelligence*, *16*(1).

Ren, F., Cao, P., Li, W., Zhao, D., and Zaiane, O., (2017). Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm. *Computerized Medical Imaging and Graphics*, *55*, pp. 54–67. https://doi.org/https://doi.org/10.1016/j.compmedimag.2016.07.011

Shi, Z., (2020). Improving k-Nearest Neighbors Algorithm for Imbalanced Data Classification. *IOP Conference Series: Materials Science and Engineering*, *719*(1), 012072. https://doi.org/10.1088/1757-899X/719/1/012072

Srinilta, C., Kanharattanachai, S., (2021). Application of Natural Neighbor-based Algorithm on Oversampling SMOTE Algorithms. *2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST)*, pp. 217–220. https://doi.org/10.1109/ICEAST52143.2021.9426310

Suryadarma, D., Akhmadi, Hastuti, and Toyamah, N., (2005). *Objective measures of family welfare for individual targeting: results from pilot project on community based monitoring system in Indonesia*. SMERU Research Institute.

Suud, M., Harsono, (2006). *3 Orientasi Kesejahteraan Sosial*. Prestasi Pustaka.

Tusyakdiah, H., (2021). *Implementasi K Nearest Neighbor (KNN) dalam Klasifikasi Status Kerja Lulusan Sekolah Menengah Kejuruan (SMK) dengan Oversampling Synthetic Minority Oversampling Technique (SMOTE) dan Adaptive Synthetic (ADASYN)*. Universitas Islam Indonesia.

Widayati, Y. T., Prihati, Y., and Widjaja, S., (2021). Analisis dan Komparasi Algoritma Naïve Bayes dan C4.5 untuk Klasifikasi Loyalitas Pelanggan MNC Play Kota Semarang. *TRANSFORMTIKA*, *18*(2), pp. 161–172.

Wilson, D. R., Martinez, T. R., (1997). Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*, *6*, pp. 1–34.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D., (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), pp. 1–37. https://doi.org/10.1007/s10115-007-0114-2

Xin, L. K., and Rashid, N. binti A., (2021). Prediction of Depression among Women Using Random Oversampling and Random Forest. *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, pp. 1–5. https://doi.org/10.1109/WiDSTaif52235.2021.9430215

Zhu, W., Zeng, N. F., and Wang, N., (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS. *Northeast SAS Users Group 2010: Health Care and Life Sciences*.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 125–143*, https://doi.org/10.59170/stattrans-2024-008
Received – 02.11.2022; accepted – 08.01.2024

# On Bayesian inference of reliability parameter in Burr-type XII model based on imprecise data: a survey on fuzzy modelling

**Iman Makhdoom**[1], **Abbas Pak**[2]

## Abstract

There are always two major sources of uncertainty in measurements related to lifetime surveys: *variation among the observations* and *imprecision of individual observation* called *fuzziness*. The typical statistical analysis is based on variation among the observations and does not consider the imprecision due to individual observation. However, ignoring the imprecision of individual observations may cause losing information and getting misleading results. It is mandatory to analyse such data, to extend the real numbers classically and Bayesian estimation methods to fuzzy numbers. Inference on the Burr-type (BT) XII model, based on precise measurements, is carried out by researchers, yet the problem of estimating parameters, in the presence of fuzzy data, remains unresolved. We are estimating the BT XII distribution parameters and their corresponding reliability when the available data are in the fuzzy numbers. The maximum likelihood estimation (MLE), the Bayesian method and the method of moments are used for estimating parameters. Finally, these estimators are compared via a Monte-Carlo simulation study.

**Key words:** Bayesian estimation, Burr-type XII distribution, Maximum likelihood estimates, Markov chain Monte Carlo, EM algorithm, Fuzzy data analysis.

## 1. Introduction

One of the important application of statistics is to analyse the lifetime data. Various distributions are suggested to model the existing real lifetime data. In this regard, Burr (1942), in his original paper, presented a system of distributions that contains twelve different types of distribution functions useful in lifetime studies, which yield a variety of density shapes. The two-parameter Burr-type (BT) XII model has the cumulative distribution function (CDF) and the probability density function (PDF) as follows:

$$F(x;c,k) = 1-(x^c+1)^{-k}, \quad f(x;c,k) = \frac{kcx^{c-1}}{(1+x^c)^{k+1}}, \quad x>0 \tag{1}$$

respectively, where $c>0$ and $k>0$ are the shape parameters. The corresponding reliability function (RF) and failure rate function (FR) are also given by

$$R(t) = (1+t^c)^{-k}, \qquad \gamma(t) = \frac{ckt^{c-1}}{1+t^c} \quad t>0. \tag{2}$$

[1]Department of Statistics, Payame Noor University(PNU), P.O.Box,19395-4697 Tehran, Iran.
E-mail: makhdoom@pnu.ac.ir. ORCID: https://orcid.org/0000-0002-2768-1024.
[2]Department of Computer Sciences, Faculty of Mathematical Sciences, Shahrekord University, P.O.Box 115, Shahrekord, Iran. E-mail: abbas.pak1982@gmail.com. ORCID: https://orcid.org/0000-0003-2388-3523.

respectively. Figure 1 represents the PDFs of BT XII distribution for various quantities of $c$ and $k$.

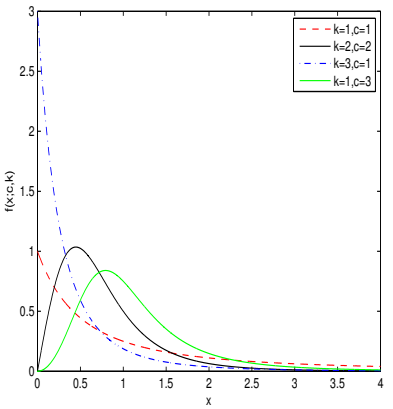


Figure 1: Probability density function for different values of $c$ and $k$.

This family of distributions, especially types III, X, and XII, have been considered and frequently studied in recent years. Investigations on the BT XII distribution, from point of view its flexibilities, have been carried out by many authors (Hate (1949), Burr (1973), Rodriguez (1977) and Singh, Singh and Kumar (2016)). Wingo (1983) has presented the maximum likelihood methods for fitting of the BT XII model to life test. Wingo (1993) also extended his work and obtained the estimating parameters of this distribution for the progressively censored scheme in life test data. See also Wu and Yu (2005), Xiuchun, Yimin, Jieqiong and Jian (2007), Soliman (2005), Moore and Papadopoulos (2000), Mousa, and Jaheen (2002), for a nice account of it.

The above inference methods, to estimate the parameters of the BT XII distribution, are limited to crisp data. But a matter of concern for the statisticians has been the exact measurement of continuous real variables. For this, numerous methods are considered to measure continuous variables precisely, yet the problem of precise measurement is unresolved, and the numbers solely are approximated.

So, there are always situations that data sometimes cannot be measured and recorded precisely due to machine errors, human error, or unexpected situations and it always remained a problem for the researchers. Note that the problem here is different from censoring and our interest is not the imprecision arising from inspection times, but it is the result of random experiment reported from the observer and its limited perception or recollection of the precise numerical value.

Here, we mention a few examples for fuzzy data. The measurement of the depth of a river because of its water level fluctuation is an imprecise quantity. It may be said that its depth is approximately 40 meters. It cannot be measured precisely at blood pressure, differentiation between a high or low blood pressure. The measurement of temperature is fuzziness. High or low temperature is imprecise quantity due to lack of differentiated be-

tween high and low temperature. The effective or ineffective teacher is also one example of fuzziness. Now, we present one case for fuzzy lifetime data as follows. The lifetime of some shafts may be stated as vague values such as: “about 1000 hours",“approximately 1400 hours", “almost between 1000 and 1200 hours", “essentially less than 1200 hours" and so on. Hence, we can conclude that there are always two types of uncertainty in measurements: *variation among the observations* and *imprecision of individual observations* ( see Viertl (2011) ).

The conventional statistical analysis is only related to variation among the observations and does not consider the vagueness of individual observations. Ignoring this imprecision may be the reason why we lose some information and get false results. The vagueness of such data can be characterized using fuzzy sets that were first introduced by Zadeh (1965). Realizing the importance of fuzziness in recent years, several authors get deep concentration on the fuzzy sets to estimation theory; but still, in most of the publications fuzziness is ignored. Gertner and Zhu (1996) considered Bayesian approximation in the forest studies when samples or prior knowledges are fuzzy. Wu (2004) obtained the Bayesian estimates on lifetime data for fuzzy environments. Gil, López-Diaz and Ralescu (2006) indicated a backward analysis on interpretation, modelling, and impact of the meaning of the fuzzy random variable. Huang, Zuo and Sun (2006) proposed a new method to determine the membership function of the estimates of the parameters and the reliability function of multiparameter lifetime distributions. Coppi, Gilb and Kiersc (2006) presented some applications of fuzzy techniques in statistical analysis. Viertl (2006) discussed a generalization of classical statistical inference methods for univariate fuzzy numbers. Akbari and Rezaei (2007) proposed a new method for uniformly minimum variance unbiased fuzzy point estimation. Zarei, Amini, Taheri and Rezaei (2012) considered the Bayesian estimation of failure rate and the mean time to failure based on vague set theory in the case of complete and censored data sets. Pak, Parham and Saraj (2013, 2014) carried out a series of studies to develop the inferential procedures for the lifetime distributions based on vague information and Shafiq and Atif (2015) obtained the survival models that deal with imprecise lifetime measurements. Very recently, Pak (2016) has investigated some inferences for the Lindley distribution based on fuzzy data.

To our knowledge, there exist no interpretation on estimating the parameters of BT XII distribution from fuzzy data. Since the classical statistical estimation procedures are not suitable for the fuzzy sets; we have to extend the conventional methods to estimate the parameters of BT XII distribution in the new situations. Therefore our main object is to develop the inferential procedures for BT XII parameters when the available data are fuzzy numbers.

The rest of the paper is set up as follows. In Section 2, we consider a review on the original understandings and basic definition of fuzzy set theory. A generalized likelihood function based on fuzzy data is introduced in Section 3. We also present the common method of maximum likelihood for estimating the parameters $c$ and $k$ by taking advantage of the Newton-Raphson (NR) and Expectation Maximization (EM) algorithms in this section. In Section 4, we carry out the estimating parameters of $c$ and $k$ using the moment method. In Section 5, we apply a Bayesian approach for estimating of the unknown parameters using the approximation forms of Tirney and Kadane (1986) and Markov Chain Monte

Carlo (MCMC) technique. Extensive numerical experiments are performed to compare the accuracy of the various proposed methods in Section 6. Finally, Section 7 concludes this research.

## 2. Preliminary concepts of fuzzy sets theory

Let us first review the fundamental notation and basic definitions of fuzzy set theory used in the paper. In the following, we explaine some special concepts of fuzzy sets theory from Viertl, (2011) and Zadeh (1965, 1968).

Notice an experiment determined by a probability space $X = (X, \mathscr{B}_X, P_\theta)$, where a measurable space is as $(X, \mathscr{B}_X)$ and $P_\theta$ belongs to a certain family of probability measures $\{P_\theta, \theta \in \Theta\}$ on $(X, \mathscr{B}_X)$. Consider that the observer cannot distinguish or transmit with exactness the outcome in the performance of $X$, but that rather the available observation may be described in terms of fuzzy information, which is defined as follows:

**Definition 1** A fuzzy event $\tilde{x}$ on $X$, determined by a Borel measurable membership function $\mu_{\tilde{x}}(x)$ from X to $[0,1]$, where $\mu_{\tilde{x}}(x)$ represents the "grade of membership" of $x$ to $\tilde{x}$, is called *fuzzy information* associated with the experiment $X$.

The set consisting of all observable events from the experiment $X$ determines a fuzzy information system associated with it, which is defined as follows.

**Definition 2** (see Tanaka, Okuda and Asai (1979)). *A fuzzy information system* (f.i.s.) $\tilde{X}$ associated with the experiment $X$ is a fuzzy partition $\{\tilde{x}_1, ..., \tilde{x}_k\}$, i.e., a set of fuzzy events on $X$ satisfying the orthogonality condition $\sum_{i=1}^{k} \mu_{\tilde{x}_i}(x) = 1 \quad$ for all $x \in X$.

On the other hand, according to Zadeh (1968) given the experiment $X = (X, \mathscr{B}_X, P_\theta)$, $\theta \in \Theta$, and a f.i.s. $\tilde{X}$ associated with it, each probability measure $P_\theta$ on $(X, \mathscr{B}_X)$ induces a probability measure on $\tilde{X}$ defined as follows:

**Definition 3** The probability distribution on $\tilde{X}$ induced by $P_\theta$ is the mapping P from $\tilde{X}$ to $[0,1]$ such that

$$\mathrm{P}(\tilde{x}) = \int_X \mu_{\tilde{x}}(x) dP_\theta(x), \quad \tilde{x} \in \tilde{X}. \tag{3}$$

In particular, the conditional density of a continuous random variable $\mathbf{Y}$ with PDF $g(\mathbf{y})$ given the fuzzy event $\tilde{A}$ can be defined as

$$g(\mathbf{y}|\tilde{A}) = \frac{\mu_{\tilde{A}}(\mathbf{y}) g(\mathbf{y})}{\int \mu_{\tilde{A}}(\mathbf{u}) g(\mathbf{u}) d\mathbf{u}}. \tag{4}$$

**Definition 4** (see Shafiq and Viertl (2014)): A subset $\tilde{x}$ of the set of real numbers (denoted by $\mathbb{R}$) is named *fuzzy number* and is characterized by the so-called membership function $\mu_{\tilde{x}}(.)$. A fuzzy number must fulfill $\mu_{\tilde{x}} : \mathbb{R} \longrightarrow [0,1]$ is Borel-measurable; $\exists x_0 \in \mathbb{R} : \mu_{\tilde{x}}(x_0) = 1$; and the so-called $\lambda-$cuts $(0 < \lambda \leq 1)$, defined as $B_\lambda(\tilde{x}) = \{x \in \mathbb{R} : \mu_{\tilde{x}}(x) \geq \lambda\}$, are all closed interval, i.e., $B_\lambda(\tilde{x}) = [a_\lambda, b_\lambda]$.

The conventional membership functions for analysing of fuzzy lifetime data are called as *triangular and trapezoidal* fuzzy numbers. A triangular fuzzy number is described as $\tilde{x} = (\nu, \omega, \delta)$ and the trapezoidal fuzzy number can also be characterized as $\tilde{x} = (\delta, \nu, \omega, \eta)$

with the corresponding membership functions

$$\mu_{\tilde{x}}(x)=\begin{cases}\frac{x-\nu}{\omega-\nu} & \nu\le x\le\omega,\\ \frac{\delta-x}{\delta-\omega} & \omega\le x\le\delta,\\ 0 & otherwise.\end{cases},\quad \mu_{\tilde{x}}(x)=\begin{cases}\frac{x-\delta}{\nu-\delta} & \delta\le x\le\nu,\\ 1 & \nu\le x\le\omega,\\ \frac{\eta-x}{\eta-\omega} & \omega\le x\le\eta,\\ 0 & otherwise.\end{cases}$$

respectively. For a detailed study on the fuzzy sets, membership functions and triangular and trapezoidal fuzzy numbers one can refer to Singpurwalla and Booker (2004) and Pak, Parham and Saraj (2013).

## 3. Maximum likelihood estimation

Let $X_1,...,X_n$ be a random sample of size $n$ from the BT XII distribution with PDF given by (1). Let $\mathbf{X}=(X_1,...,X_n)$ denotes the corresponding random vector. If a realization $\mathbf{x}=(x_1,...,x_n)$ of $\mathbf{X}$ is known exactly, we can obtain the complete data likelihood function as

$$L(c,k;\mathbf{x})=(kc)^n\prod_{i=1}^{n}\frac{x_i^{c-1}}{(1+x_i^c)^{k+1}} \tag{5}$$

Now, consider the problem where the results of an experimental performance are not recorded or measured precisely, but that rather the available data are identified by means of fuzzy observation $\tilde{\mathbf{x}}=(\tilde{x}_1,...,\tilde{x}_n)$ with the Borel measurable membership function $\mu_{\tilde{\mathbf{x}}}(\mathbf{x})$. In practice, the grade of membership $\mu_{\tilde{\mathbf{x}}}(\mathbf{x})$ is often regarded as a kind of probability with which the observer gets the information $\tilde{\mathbf{x}}$ when he really has obtained the exact outcome $\mathbf{x}$ . Once $\tilde{\mathbf{x}}$ is given, and assuming the joint membership function $\mu_{\tilde{\mathbf{x}}}(\mathbf{x})$ to be decomposable as $\mu_{\tilde{\mathbf{x}}}(\mathbf{x})=\mu_{\tilde{x}_1}(x_1)\times...\times\mu_{\tilde{x}_n}(x_n)$, its probability can be computed based on Zadeh's definition (see Zadeh (1968)) of the probability of a fuzzy event. By using the expression (3), the observed-data likelihood function based on the fuzzy sample $\tilde{\mathbf{x}}$ can then be obtained as

$$Lo(c,k;\tilde{\mathbf{x}})=P(\tilde{\mathbf{x}};c,k)=\int f(\mathbf{x};c,k)\mu_{\tilde{\mathbf{x}}}(\mathbf{x})d\mathbf{x}. \tag{6}$$

Since the data vector $\mathbf{x}$ is a realization of an independent identically distributed (i.i.d.) random vector $\mathbf{X}$, the likelihood function (6) can be written as:

$$Lo^*(c,k;\tilde{\mathbf{x}})=(kc)^n\prod_{i=1}^{n}\int\frac{x^{c-1}}{(1+x^c)^{k+1}}\mu_{\tilde{x}_i}(x)dx \tag{7}$$

Then, the observed data log-likelihood function is as follows:

$$L^{**}(c,k,\tilde{\mathbf{x}})=n\ln(kc)+\sum_{i=1}^{n}\ln\left(\int\frac{x^{c-1}}{(1+x^c)^{k+1}}\mu_{\tilde{x}_i}(x)dx\right). \tag{8}$$

The maximum likelihood estimate (MLE) of parameters $c$ and $k$ can be obtained by maximizing the log-likelihood $L^{**}(c,k,\tilde{x})$. Equating the partial derivatives of the log-likelihood

(8) with respect to $c$ and $k$ to zero, the resulting two equations are:

$$\frac{\partial}{\partial k}L^{**}(c,k,\tilde{x}) = \frac{n}{k} - \sum_{i=1}^{n} \frac{\int x^{c-1}(1+x^c)^{-k-1}\ln(1+x^c)\mu_{\tilde{x}_i(x)}dx}{\int x^{c-1}(1+x^c)^{-k-1}\mu_{\tilde{x}_i(x)}dx} = 0 \tag{9}$$

and

$$\frac{\partial}{\partial c}L^{**}(c,k,\tilde{x}) = \frac{n}{c} + \sum_{i=1}^{n} \frac{\int x^{c-1}(1+x^c)^{-k-1}(1-(k+1)(1+x^c)^{-1}x^c)\ln(x)\mu_{\tilde{x}_i(x)}dx}{\int x^{c-1}(1+x^c)^{-k-1}\mu_{\tilde{x}_i(x)}dx} = 0. \tag{10}$$

There is no closed form solution for the likelihood equation, therefore an iterative numerical search is used to obtain the MLEs. In next section the Newton-Raphson method and EM algorithm are used to obtain the MLE of the $c$ and $k$ parameters.

### 3.1. Newton-Raphson Algorithm

In this method, the solution of the likelihood equations is obtained through an iterative procedure. Let $\theta = (k,c)^T$ be the parameter vector. Then, at $(h+1)$th step of iteration process, the updated parameter is computed as

$$\theta^{(h+1)} = \theta^{(h)} - \left[\frac{\partial^2 L^{**}(\theta;\tilde{\mathbf{x}})}{\partial\theta\partial\theta^T}|_{\theta=\theta^{(h)}}\right]^{-1}\left[\frac{\partial L^{**}(\theta;\tilde{\mathbf{x}})}{\partial\theta}|_{\theta=\theta^{(h)}}\right] \tag{11}$$

in which

$$\frac{\partial L^{**}(\theta;\tilde{\mathbf{x}})}{\partial\theta} = \begin{pmatrix} \frac{\partial L^{**}(c,k;\tilde{\mathbf{x}})}{\partial c} \\ \frac{\partial L^{**}(c,k;\tilde{\mathbf{x}})}{\partial k} \end{pmatrix}, \quad \frac{\partial^2 L^{**}(\theta;\tilde{\mathbf{x}})}{\partial\theta\partial\theta^T} = \begin{pmatrix} \frac{\partial L^{**}(c,k;\tilde{\mathbf{x}})}{\partial c^2} & \frac{\partial L^{**}(c,k;\tilde{\mathbf{x}})}{\partial c\partial k} \\ \frac{\partial L^{**}(c,k;\tilde{\mathbf{x}})}{\partial k\partial c} & \frac{\partial L^{**}(c,k;\tilde{\mathbf{x}})}{\partial k^2} \end{pmatrix}.$$

For proceeding with the NR method, we need the second-order derivatives of the log-likelihood with respect to the parameters that are obtained as follows.

$$\frac{\partial^2}{\partial c^2}L^{**}(c,k,\tilde{\mathbf{x}}) = \frac{-n}{c^2} \tag{12}$$
$$+\sum_{i=1}^{n}\frac{\left\{\int x^{c-1}(1+x^c)^{-k-1}mdx - \int(k+1)(1+x^c)^{-k-2}x^{2c-1}mdx\right\}B}{B^2}$$
$$+\sum_{i=1}^{n}\frac{\left\{\int(k+1)(1+x^c)^{-k-3}x^{2c-1}\ln^2 x\mu_{\tilde{x}_i}(x)dx - \int x^{2c-1}(k+1)(1+x^c)^{-k-2}\ln^2 x\mu_{\tilde{x}_i}(x)dx\right\}B}{B^2}$$
$$-\sum_{i=1}^{n}\frac{\left\{\int x^{c-1}(1+x^c)^{-k-1}\ln x\mu_{\tilde{x}_i}(x)dx - \int(k+1)(1+x^c)^{-k-2}x^{2c-1}\ln x\mu_{\tilde{x}_i}(x)dx\right\}A}{B^2},$$

$$\frac{\partial^2}{\partial k^2}L^{**}(c,k,\tilde{\mathbf{x}}) = \frac{-n}{k^2} \tag{13}$$
$$+\sum_{i=1}^{n}\frac{\left(\int x^{c-1}(1+x^c)^{-k-1}\ln^2(1+x^c)\mu_{\tilde{x}_i(x)}dx\right)B}{B^2} - \sum_{i=1}^{n}\frac{\left(\int x^{c-1}(1+x^c)^{-k-1}\ln(1+x^c)\mu_{\tilde{x}_i(x)}dx\right)^2}{B^2},$$

$$\frac{\partial^2}{\partial c\partial k}L^{**}(c,k,\tilde{\mathbf{x}}) = \tag{14}$$
$$\sum_{i=1}^{n}\frac{\left(-\int x^{c-1}(1+x^c)^{-k-1}(1-(k+1)(1+x^c)^{-1}x^c)\ln(1+x^c)\ln x\mu_{\tilde{x}_i}(x)dx - \int(1+x^c)^{-k-2}x^{2c-1}\ln x\mu_{\tilde{x}_i}(x)dx\right)B}{B^2}$$
$$+\sum_{i=1}^{n}\frac{\left(\int x^{c-1}\ln(1+x^c)(1+x^c)^{-k-1}\mu_{\tilde{x}_i}(x)dx\right)\left(\int x^{c-1}(1+x^c)^{-k-1}(1-(k+1)(1+x^c)^{-1}x^c)\ln(x)\mu_{\tilde{x}_i(x)}dx\right)}{B^2},$$

in which, $m = (1-(k+1)(1+x^c)^{-1}x^c)\ln^2(x)\mu_{\tilde{x}_i}(x)$, $A = \int x^{c-1}(1+x^c)^{-k-1}(1-(k+1)(1+x^c)^{-1}x^c)\ln(x)\mu_{\tilde{x}_i(x)}dx$, $B = \int x^{c-1}(1+x^c)^{-k-1}\mu_{\tilde{x}_i(x)}dx$. The iteration process continues until convergence, i.e., until $\|\theta^{(h+1)} - \theta^{(h)}\| < \varepsilon$ for some pre-fixed $\varepsilon > 0$.

Note that the second-order derivatives of the log-likelihood are needed at every iteration in this method. The calculation of the derivatives based on fuzzy data can be some tedious in most of the time. To solve this problem, an EM algorithm will be present in the following section.

### 3.2. EM Algorithm

The EM algorithm is a convenient method for incomplete data problems. Since the observed fuzzy data $\tilde{\mathbf{x}}$ can be considered an incomplete data vector $\mathbf{x}$, therefore an EM algorithm is used to obtain the MLE of the unknown parameters $c$ and $k$, (see Denoeux (2011)).
From the Eq. (5), the log-likelihood function for the complete data vector $\mathbf{x}$ is given by,

$$\ln L(c,k,\mathbf{x}) = n\ln k + n\ln c + (c-1)\sum_{i=1}^{n}\ln x_i - (k+1)\sum_{i=1}^{n}\ln(1+x_i^c) \tag{15}$$

Taking the derivative with respect to $c$ and $k$, respectively, on (15), the following likelihood equations are obtained:

$$\frac{n}{c} = (k+1)\sum_{i=1}^{n}\frac{x_i^c\ln x_i}{1+x_i^c} - \sum_{i=1}^{n}\ln x_i, \qquad \frac{n}{k} = \sum_{i=1}^{n}\ln(1+x_i^c). \tag{16}$$

So, the EM algorithm is given by the following iterative process:

- Given starting values of $c$ and $k$ say $c^{(0)}$ and $k^{(0)}$ and set $h = 0$. In the $(h+1)$-th iteration, the E-step requires to compute the following conditional expectations using the expression (4):

$$E_{1i} = E_{c^{(h)},k^{(h)}}(\ln X|\tilde{x}_i) = \frac{\int x^{c^{(h)}-1}(\ln x)(1+x^{c^{(h)}})^{-k^{(h)}-1}\mu_{\tilde{x}_i}(x)dx}{\int x^{c^{(h)}-1}(1+x^{c^{(h)}})^{-k^{(h)}-1}\mu_{\tilde{x}_i}(x)dx} \tag{17}$$

$$E_{2i} = E_{c^{(h)},k^{(h)}}\left(\frac{X^c\ln X}{1+X^c}|\tilde{x}_i\right) = \frac{\int x^{2c^{(h)}-1}(\ln x)(1+x^{c^{(h)}})^{-k^{(h)}-2}\mu_{\tilde{x}_i}(x)dx}{\int x^{c^{(h)}-1}(1+x^{c^{(h)}})^{-k^{(h)}-1}\mu_{\tilde{x}_i}(x)dx} \tag{18}$$

$$E_{3i} = E_{c^{(h)},k^{(h)}}(\ln(1+X^c)|\tilde{x}_i) = \frac{\int x^{c^{(h)}-1}(\ln(1+x^{c^{(h)}})(1+x^{c^{(h)}})^{-k^{(h)}-1}\mu_{\tilde{x}_i}(x)dx}{\int x^{c^{(h)}-1}(1+x^{c^{(h)}})^{-k^{(h)}-1}\mu_{\tilde{x}_i}(x)dx}. \tag{19}$$

The likelihood equations (16) are replaced by

$$\frac{n}{c} = (k+1)\sum_{i=1}^{n}E_{2i} - \sum_{i=1}^{n}E_{1i}, \qquad \frac{n}{k} = \sum_{i=1}^{n}E_{3i}. \tag{20}$$

The M-step requires to solve the Eqs. in (20), and obtain the next values, $c^{(h+1)}$ and

$k^{(h+1)}$, of $c$ and $k$, respectively, as follows:

$$c^{(h+1)} = \frac{n}{(k^{(h+1)}+1)\sum_{i=1}^{n} E_{2i} - \sum_{i=1}^{n} E_{1i}}, \qquad k^{(h+1)} = \frac{n}{\sum_{i=1}^{n} E_{3i}}. \tag{21}$$

- Checking convergence, if the convergence occurs then the current $c^{(h+1)}$ and $k^{(h+1)}$ are the maximum likelihood estimates of $c$ and $k$ using the EM algorithm; otherwise, set $h = h+1$ and go to previous step.

The maximum likelihood estimate of $(c,k)$ by applying the EM algorithm is thereafter refereed to as "$(\hat{c}_{EM}, \hat{k}_{EM})$" in this article.

## 4. Method of moment

The $r$th moment of the BT XII distribution (see Rodriguez (1977)) is given by

$$E(X^r) = kB\left(\frac{r}{c}+1, k-\frac{r}{c}\right) = \frac{\Gamma(\frac{r}{c}+1)\Gamma(k-\frac{r}{c})}{\Gamma(k)} \tag{22}$$

in which, $B(.)$ and $\Gamma(.)$ are the beta and the complete gamma functions, respectively.

By equating the first and the second sample moments to the corresponding population moments, to obtain the estimates of moments approach, the following equations are used:

$$\frac{\Gamma(\frac{1}{c}+1)\Gamma(k-\frac{1}{c})}{\Gamma(k)} = \frac{1}{n}\sum_{i=1}^{n} E_{c,k}(X|\tilde{x}_i), \qquad \frac{\Gamma(\frac{2}{c}+1)\Gamma(k-\frac{2}{c})}{\Gamma(k)} = \frac{1}{n}\sum_{i=1}^{n} E_{c,k}(X^2|\tilde{x}_i). \tag{23}$$

Since the closed form of the solutions to Eqs. in (23) could not be obtained, to achieve the parameter estimates we use the following iterative numerical process. Let the initial estimates of $c$ and $k$, say $c^{(0)}$ and $k^{(0)}$ with $h = 0$. In the $(h+1)$ th iteration, we first compute

$$E_{c^{(h)},k^{(h)}}(X^r|\tilde{x}_i) = \frac{\int x^{c^{(h)}+r-1}(1+x^{c^{(h)}})^{-k^{(h)}-1}\mu_{\tilde{x}_i}(x)dx}{\int x^{c^{(h)}-1}(1+x^{c^{(h)}})^{-k^{(h)}-1}\mu_{\tilde{x}_i}(x)dx}, \quad r = 1,2.$$

We have to solve the system of two equations and two unknowns in the Eqs. (23). These equations are the complex nonlinear equations. Consequently, we may need to use an iterative numerical method to handle the finding of the roots $c$ and $k$ of equations in (23).

## 5. Bayesian approach

A robust and valid alternative to traditional statistical perspectives has been called Bayesian inference in recent decades. It received frequent attention for statistical inference. In this section the Bayesian estimates under the assumptions that $c$ and $k$ have independent gamma

priors are obtained with the pdfs respectively,

$$\pi_1(c) = \frac{b_1^{a_1}}{\Gamma(a_1)} c^{a_1-1} \exp(-cb_1), \quad c > 0 \tag{24}$$

and

$$\pi_2(k) = \frac{b_2^{a_2}}{\Gamma(a_2)} k^{a_2-1} \exp(-kb_2), \quad k > 0 \tag{25}$$

with the parameters $c \sim Gamma(a_1, b_1)$ and $k \sim Gamma(a_2, b_2)$ (see Singh, Singh and Kumar (2016)). Due to the likelihood function in Eq. (7) and the prior distributions mentioned in (24) and (25) the complete form of the posterior function is as follows:

$$\pi(\theta|\tilde{\mathbf{x}}) \propto L^{**}(\theta, \tilde{\mathbf{x}})(c^{a_1-1} \exp(-cb_1))(k^{a_2-1} \exp(-kb_2)). \tag{26}$$

Hence, the joint posterior density function of $c$ and $k$ given the data can be written as follows:

$$\pi(c,k|\tilde{\mathbf{x}}) = \frac{\pi_1(c)\pi_2(k)Lo^*(c,k;\tilde{\mathbf{x}})}{\int_0^\infty \int_0^\infty \pi_1(c)\pi_2(k)Lo^*(c,k;\tilde{\mathbf{x}})dcdk}. \tag{27}$$

Therefore, the Bayes estimate of any function of $c$ and $k$, say $g(c,k)$, under a squared error loss function is

$$E(g(c,k)|\tilde{\mathbf{x}}) = \frac{\int_0^\infty \int_0^\infty g(c,k)\pi_1(c)\pi_2(k)Lo^*(c,k;\tilde{\mathbf{x}})dcdk}{\int_0^\infty \int_0^\infty \pi_1(c)\pi_2(k)Lo^*(c,k;\tilde{\mathbf{x}})dcdk}. \tag{28}$$

But, we cannot evaluate these estimates explicitly. Hence, we suggest Tierney and Kadane's procedure and MCMC method to approximate them.

### 5.1. Tierney and Kadane's method

The Eq. (28) can be re-written as follows:

$$E(g(c,k)|\tilde{\mathbf{x}}) = \frac{\int_0^\infty \int_0^\infty g(c,k)e^{Q(c,k)}dcdk}{\int_0^\infty \int_0^\infty e^{Q(c,k)}dcdk} \tag{29}$$

in which, $Q(c,k) = \ln[\pi_1(c)\pi_2(k)] + \ln Lo^*(c,k;\tilde{\mathbf{x}}) \equiv \rho(c,k) + L^{**}(c,k)$. Note that Eq. (29) cannot be obtained analytically. Using this approximation can be useful to solve this issue.

Setting $H(c,k) = \frac{Q(c,k)}{n}$ and $H^*(c,k) = \frac{[\ln g(c,k)+Q(c,k)]}{n}$, the expression in (29) can be reexpressed as

$$E(g(c,k)|\tilde{\mathbf{x}}) = \frac{\int_0^\infty \int_0^\infty e^{nH^*(c,k)}dcdk}{\int_0^\infty \int_0^\infty e^{nH(c,k)}dcdk}. \tag{30}$$

Following the Tierney & Kadane method, which is based on Laplace's method ( Tierney and Kadane (1986)), Eq. (30) can be computed as follows:

$$\hat{g}_{Bayes}(c,k) = \left[\frac{det\,\Sigma^*}{det\,\Sigma}\right]^{\frac{1}{2}} \exp\left\{n\left[H^*(\overline{c}^*, \overline{k}^*) - H(\overline{c}, \overline{k})\right]\right\} \tag{31}$$

in which$(\overline{c}^*,\overline{k}^*)$ and $(\overline{c},\overline{k})$ maximize $H^*(c,k)$ and $H(c,k)$, respectively. Also, $\sum^*$ and $\sum$ are the negatives of the inverse Hessians of $H^*(c,k)$ and $H(c,k)$ at $(\overline{c}^*,\overline{k}^*)$ and $(\overline{c},\overline{k})$, respectively.

For our case, we have

$$H(c,k)=\frac{1}{n}\{K^*+(a_1-1+n)\ln c+(a_2-1+n)\ln k-b_1c-b_2k\} \tag{32}$$
$$+\frac{1}{n}\left\{\sum_{i=1}^{n}\ln\left(\int\frac{x^{c-1}}{(1+x^c)^{k+1}}\mu_{\tilde{x}_i}(x)dx\right)\right\}$$

in which $K^*$ is a constant. So, we can obtain $(\overline{c},\overline{k})$ by solving the following two equations:

$$\frac{\partial}{\partial c}H(c,k)=\frac{1}{n}\left\{\frac{a_1-1+n}{c}-b_1\right\} \tag{33}$$
$$+\frac{1}{n}\left\{\sum_{i=1}^{n}\frac{\int x^{c-1}(1+x^c)^{-k-1}(1-(k+1)(1+x^c)^{-1}x^c)\ln(x)\mu_{\tilde{x}_i(x)}dx}{\int x^{c-1}(1+x^c)^{-k-1}\mu_{\tilde{x}_i(x)}dx}\right\}$$
$$\frac{\partial}{\partial k}H(c,k)=\frac{1}{n}\left\{\frac{a_2-1+n}{k}-b_2\right\} \tag{34}$$
$$-\frac{1}{n}\left\{\sum_{i=1}^{n}\frac{\int x^{c-1}(1+x^c)^{-k-1}\ln(1+x^c)\mu_{\tilde{x}_i(x)}dx}{\int x^{c-1}(1+x^c)^{-k-1}\mu_{\tilde{x}_i(x)}dx}\right\}.$$

The determinant of the negative of the inverse Hessian of $H(c,k)$ at $(\overline{c},\overline{k})$ is as follows:

$$det\sum=\left(H_{11}H_{22}-H_{12}^2\right)^{-1}, \tag{35}$$

in which,

$$H_{11}=\tfrac{1}{n}\tfrac{-(a_1-1+n)}{\overline{c}^2} \tag{36}$$
$$+\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\left\{\int x^{\overline{c}-1}(1+x^{\overline{c}})^{-\overline{k}-1}\overline{m}dx-\int(\overline{k}+1)(1+x^{\overline{c}})^{-\overline{k}-2}x^{2\overline{c}-1}\overline{m}dx\right\}\overline{B}}{\overline{B}^2}$$
$$+\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\left\{\int(\overline{k}+1)(1+x^{\overline{c}})^{-\overline{k}-3}x^{2\overline{c}-1}\ln^2 x\mu_{\tilde{x}_i}(x)dx-\int x^{2\overline{c}-1}(\overline{k}+1)(1+x^{\overline{c}})^{-\overline{k}-2}\ln^2 x\mu_{\tilde{x}_i}(x)dx\right\}\overline{B}}{\overline{B}^2}$$
$$-\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\left\{\int x^{\overline{c}-1}(1+x^{\overline{c}})^{-\overline{k}-1}\ln x\mu_{\tilde{x}_i}(x)dx-\int(k+1)(1+x^{\overline{c}})^{-\overline{k}-2}x^{2\overline{c}-1}\ln x\mu_{\tilde{x}_i}(x)dx\right\}\overline{A}}{\overline{B}^2},$$
$$H_{22}=\tfrac{1}{n}\tfrac{-(a_2-1+n)}{\overline{k}^2} \tag{37}$$
$$-\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\left(\int x^{\overline{c}-1}(1+x^{\overline{c}})^{-\overline{k}-1}\ln^2(1+x^{\overline{c}})\mu_{\tilde{x}_i(x)}dx\right)B}{B^2}+\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\left(\int x^{\overline{c}-1}(1+x^{\overline{c}})^{-\overline{k}-1}\ln(1+x^{\overline{c}})\mu_{\tilde{x}_i(x)}dx\right)^2}{\overline{B}^2},$$
$$H_{12}= \tag{38}$$
$$-\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\left(\int x^{\overline{c}-1}(1+x^{\overline{c}})^{-\overline{k}-1}(1-(\overline{k}+1)(1+x^{\overline{c}})^{-1}x^{\overline{c}})\ln(1+x^{\overline{c}})\ln x\mu_{\tilde{x}_i}(x)dx+\int(1+x^{\overline{c}})^{-\overline{k}-2}x^{2\overline{c}-1}\ln x\mu_{\tilde{x}_i}(x)dx\right)\overline{B}}{\overline{B}^2}$$
$$+\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{\left(\int x^{\overline{c}-1}\ln(1+x^{\overline{c}})(1+x^{\overline{c}})^{-\overline{k}-1}\mu_{\tilde{x}_i}(x)dx\right)\left(\int x^{\overline{c}-1}(1+x^{\overline{c}})^{-\overline{k}-1}(1-(\overline{k}+1)(1+x^{\overline{c}})^{-1}x^{\overline{c}})\ln(x)\mu_{\tilde{x}_i(x)}dx\right)}{\overline{B}^2},$$

in which $\overline{m} = (1-(\overline{k}+1)(1+x^{\overline{c}})^{-1}x^{\overline{c}})\ln^2(x)\mu_{\tilde{x}_i}(x)$, $\overline{A} = \int x^{\overline{c}-1}(1+x^{\overline{c}})^{-\overline{k}-1}(1-(\overline{k}+1)(1+x^{\overline{c}})^{-1}x^{\overline{c}})\ln(x)\mu_{\tilde{x}_i(x)}dx$, $\overline{B} = \int x^{\overline{c}-1}(1+x^{\overline{c}})^{-\overline{k}-1}\mu_{\tilde{x}_i(x)}dx$. Now, following the same arguments with $g(c,k) = c$ and $k$, respectively, in $H^*(c,k)$, $\hat{c}_{Bayes}$ and $\hat{k}_{Bayes}$ in Eq. (31) can then be obtained in a straightforward manner.

### 5.2. MCMC Method

MCMC methods use the computer simulation procedure to get a Markov sequence with ergodic properties in such a way that they have a limiting distribution. We know if the loss function is squared error, then the Bayes estimates of the parameters $\theta = (c,k)$ are their respective posteriors mean. But due to the complexity of the extraction of samples from the posterior function, we have to apply the well-known "**Gibbs sampling** " technique. Gibbs sampling defines a broad class of MCMC methods that is used in Bayesian analysis. It is also a special example of a general approach referred to as Metropolis-Hasting (MH) algorithm (Hanagl and Ahmadi (2009)).Thus, a multivariate version of MH algorithm is Gibbs sampling. In the following, the Gibbs sampler method is appropriated to compute the Bayes estimates numerically.

The posterior PDFs of $c$ and $k$ are given by,

$$\pi_1^*(c|k,\tilde{\mathbf{x}}) \propto c^{a_1-1+n}\exp(-cb_1)\prod_{i=1}^{n}\int \frac{x^{c-1}}{(1+x^c)^{k+1}}\mu_{\tilde{x}_i}(x)dx \tag{39}$$

and

$$\pi_2^*(k|c,\tilde{\mathbf{x}}) \propto k^{a_2-1+n}\exp(-kb_2)\prod_{i=1}^{n}\int \frac{x^{c-1}}{(1+x^c)^{k+1}}\mu_{\tilde{x}_i}(x)dx \tag{40}$$

Note that the posterior PDFs of $c$ and $k$ in (39) and (40) respectively, are unknown. Hence, we use the MH method to generate a random sample from these distributions. We use the normal distribution as the proposal distribution for the method. So, the Gibbs sampling algorithm is as follows:

Step 1: Start with an initial value $(c^{(0)},k^{(0)})$ and fix $t = 1$.

Step 2: Generate $c^{(0)}$ from (39) by using of the MH with the $N(c^{(t-1)},1)$ proposal distribution and generate $k^{(0)}$ from (40) by using of the MH with the $N(k^{(t-1)},1)$ proposal distribution. Fix $t = t+1$.

Step 3: Repeat Step 2, $T$ times.

For running the algorithm above, we need to perform the MH algorithm in Step 2. These algorithms are given by,

I : Fix t=1.

II : Let $v_1 = c_i^{(t-1)}$ and $v_2 = k_i^{(t-1)}$. Generate $w_1$ and $w_2$ from the proposal distributions $q \sim N(c_i^{(t-1)},1)$ and $q \sim N(k_i^{(t-1)},1)$, respectively. Let $p_1^*(v_1,w_1) = \min\left\{1, \frac{\pi_1^*(w_1|\tilde{x})q(v_1)}{\pi_1^*(v_1|\tilde{x})q(w_1)}\right\}$ and $p_2(v_2,w_2) = \min\left\{1, \frac{\pi_2^*(w_2|\tilde{x})q(v_2)}{\pi_2^*(v_2|\tilde{x})q(w_2)}\right\}$. Generate $u$ from $Uniform(0,1)$. If $u <$

$p_1(v_1,w_1)$ we accept $w_1$ and else accept $v_1$ and also if $u < p_2(v_2,w_2)$ we accept $w_2$ and else accept $v_2$. Fix $t = t+1$.

III : Repeat Step II , $T$ times.

So, the retained sample values, say $c_1,\ldots,c_{T-M_1}$, and $k_1,\ldots,k_{T-M_2}$ are random samples from the posterior densities in the equations of (39) and (40 ), respectively. Now, by using Monte Carlo integration technique (Rodriguez (1977)), the Bayes estimates of $c$ and $k$ under squared error loss function are given by,

$$\hat{c}_{Bayes} = \frac{1}{T-M_1}\sum_{i=M_1+1}^{T} c_i^{(i)}, \hat{k}_{Bayes} = \frac{1}{T-M_2}\sum_{i=M_2+1}^{T} k_i^{(i)},$$

where $M_1$ and $M_2$ are the burn-in periods in generating $c_i^{(i)}$ and $k_i^{(i)}$ ,$(i = 1,\ldots,n)$ respectively. We can also conduct the highest posterior density (HPD) confidence interval of parameter $\theta = (c,k)$. First order $c_1^{(i)},\ldots,c_{M_1}^{(i)}$ as $c_{(1)}^{(i)} < \ldots < c_{(M_1)}^{(i)}$, then construct all the $(100(1-\eta)\%)$ confidence intervals of $c$ are given by

$$\left(c_{(1)}^{(i)}, c_{([M_1(1-\eta)])}^{(i)}\right),\ldots,\left(c_{([M_1\eta])}^{(i)}, c_{([M_1])}^{(i)}\right), \tag{41}$$

where $[M]$ is symbolized as the largest integer less than or equal to $M$. So, the HPD confidence interval of $c$ is the shortest length interval. Similarly, we can make a $100(1-\eta)\%$ HPD confidence interval of $k$ as follows:

$$\left(k_{(1)}^{(i)}, k_{([M_2(1-\eta)])}^{(i)}\right),\ldots,\left(k_{([M_2\eta])}^{(i)}, k_{([M_2])}^{(i)}\right). \tag{42}$$

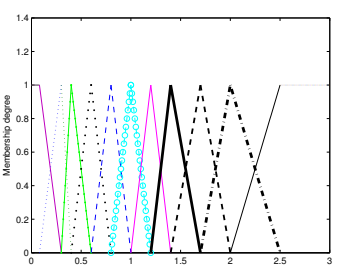


Figure 2: Fuzzy information system used to encode the simulated data.

## 6. Numerical Study

In the present section, some of the simulation studies are done to compare the performance of Bayesian estimation methods. All numerical computations are made using *MATLAB* software.

### 6.1. Monte Carlo simulations

In this section, some numerical results via Monte Carlo simulations are used to see how the different methods behave for various sample sizes. We evaluate the estimate of unknown parameters $c$ and $k$ using the methods provided in the preceding sections. For this reason the *i.i.d.* random samples, say $\mathbf{x}$ of the BT XII distribution for parameter values, namely, $(c,k) = (1,1),(2,3),(3,1)$ and various choices of $n = 10,20,30,40,50,70,100$ are generated. Each realization of $\mathbf{x}$ was produced using the method proposed by Pak, Parham and Saraj (2014). By employing fuzzy information system (f.i.s.) $\{\tilde{x}_1,\ldots,\tilde{x}_{11}\}$ shown in Fig. 2, the corresponding membership functions are given by

$$\mu_{\tilde{x}_1}(x) = \begin{cases} 1 & x \le 0.08, \\ \frac{0.3-x}{0.22} & 0.08 \le x \le 0.3, \\ 0 & otherwise, \end{cases} \qquad \mu_{\tilde{x}_2}(x) = \begin{cases} \frac{x-0.08}{0.22} & 0.08 \le x \le 0.3, \\ \frac{0.4-x}{0.1} & 0.3 \le x \le 0.4, \\ 0 & otherwise, \end{cases}$$

$$\mu_{\tilde{x}_3}(x) = \begin{cases} \frac{x-0.3}{0.1} & 0.3 \le x \le 0.4, \\ \frac{0.6-x}{0.2} & 0.4 \le x \le 0.6, \\ 0 & otherwise, \end{cases} \qquad \mu_{\tilde{x}_4}(x) = \begin{cases} \frac{x-0.4}{0.2} & 0.4 \le x \le 0.6, \\ \frac{0.8-x}{0.2} & 0.6 \le x \le 0.8, \\ 0 & otherwise, \end{cases}$$

$$\mu_{\tilde{x}_5}(x) = \begin{cases} \frac{x-0.6}{0.2} & 0.6 \le x \le 0.8, \\ \frac{1-x}{0.2} & 0.8 \le x \le 1, \\ 0 & otherwise, \end{cases} \qquad \mu_{\tilde{x}_6}(x) = \begin{cases} \frac{x-0.8}{0.2} & 0.8 \le x \le 1, \\ \frac{1.2-x}{0.2} & 1 \le x \le 1.2, \\ 0 & otherwise, \end{cases}$$

$$\mu_{\tilde{x}_7}(x) = \begin{cases} \frac{x-1}{0.2} & 1 \le x \le 1.2, \\ \frac{1.4-x}{0.2} & 1.2 \le x \le 1.4, \\ 0 & otherwise, \end{cases} \qquad \mu_{\tilde{x}_8}(x) = \begin{cases} \frac{x-1.2}{0.2} & 1.2 \le x \le 1.4, \\ \frac{1.7-x}{0.3} & 1.4 \le x \le 1.7, \\ 0 & otherwise, \end{cases}$$

$$\mu_{\tilde{x}_9}(x) = \begin{cases} \frac{x-1.4}{0.3} & 1.4 \le x \le 1.7, \\ \frac{2-x}{0.3} & 1.7 \le x \le 2, \\ 0 & otherwise, \end{cases} \qquad \mu_{\tilde{x}_{10}}(x) = \begin{cases} \frac{x-1.7}{0.3} & 1.7 \le x \le 2, \\ \frac{2.5-x}{0.5} & 2 \le x \le 2.5, \\ 0 & otherwise, \end{cases}$$

$$\mu_{\tilde{x}_{11}}(x) = \begin{cases} \frac{x-2}{0.5} & 2 \le x \le 2.5 \\ 1 & x \ge 2.5, \\ 0 & otherwise, \end{cases}$$

### 6.2. Implementation

Since the MCMC method will stabilize asymptotically, it needs to examine the reliability of the chain outcome. *Burn-in* is a significant problem that is necessary to be considered. It means that discarding the number of iterations is essential. Some diagnostic tests that indicate the convergence problem can be found in the literature. One of them is trace plot in which the history of the chain is exhibited (see Fig. 3). Plots in Fig. 3, after discarding the initial 3000 iterates, show that the sequences have a stationary pattern. The estimates of the parameters $c$, $k$, and $R$ for the fuzzy sample were obtained using the Bayesian approach. For simulation purpose, we have assumed that $c$, $k$ have gamma priors, including the *noninformative* prior (**Prior I**), i.e. $a_1 = b_1 = a_2 = b_2 = 0$, *less informative* prior (**Prior II**), i.e.

$a_1 = b_1 = a_2 = b_2 = 0.01$, and *most informative* prior (**Prior III**), i.e. $a_1 = b_1 = a_2 = b_2 = 4$. We replicate the process 15000 times and use 12000 iterates after discarding the initial 3000 iterates as *Burn-in* to make the inference. We have also reported the average values (AV) and mean squared errors (MSE) of the estimates through Tables 1-3.

Table 1: The average values (AV) and the mean squared errors (MSE) of the estimate of parameters $\theta = (c,k) = (3,2)$, and $R = 0.7901$.

| n | | | | *priorI* | | |
|---|---|---|---|---|---|---|
| | AV(c) | MSE(c) | AV(k) | MSE(k) | AV(R) | MSE(R) |
| 10 | 3.6086 | 0.3704 | 3.5707 | 2.4673 | 0.7417 | 0.0031 |
| 20 | 3.3724 | 0.1387 | 2.4171 | 0.3905 | 0.7909 | 0.0023 |
| 30 | 3.0210 | 0.0864 | 2.6249 | 0.2004 | 0.7337 | 0.0017 |
| 40 | 3.0133 | 0.0166 | 2.4477 | 0.1739 | 0.7484 | 0.0014 |
| 50 | 2.7060 | 0.0102 | 1.9781 | 0.0007 | 0.7518 | 0.0004 |
| 70 | 2.8709 | 0.0004 | 2.0273 | 0.0004 | 0.7692 | 0.0002 |
| 100 | 2.8987 | 0.0001 | 2.0015 | 0.0000 | 0.7759 | 0.0001 |
| n | | | | *priorII* | | |
| 10 | 3.6193 | 0.3836 | 3.5889 | 2.5247 | 0.7427 | 0.0032 |
| 20 | 3.3789 | 0.1435 | 2.4162 | 0.3839 | 0.7920 | 0.0022 |
| 30 | 3.0121 | 0.0836 | 2.6196 | 0.1732 | 0.7331 | 0.0014 |
| 40 | 2.7108 | 0.0836 | 1.9768 | 0.0009 | 0.7526 | 0.0014 |
| 50 | 2.7108 | 0.0159 | 1.9768 | 0.0005 | 0.7526 | 0.0004 |
| 70 | 2.8737 | 0.0114 | 2.0306 | 0.0005 | 0.7693 | 0.0002 |
| 100 | 2.8930 | 0.0001 | 2.0027 | 0.0001 | 0.7750 | 0.0001 |
| n | | | | *priorIII* | | |
| 10 | 2.3191 | 0.4635 | 1.8712 | 0.0270 | 0.7080 | 0.0067 |
| 20 | 2.6611 | 0.2570 | 1.8932 | 0.0169 | 0.7526 | 0.0057 |
| 30 | 2.5557 | 0.1973 | 2.1302 | 0.0165 | 0.7140 | 0.0034 |
| 40 | 2.6613 | 0.1147 | 2.1151 | 0.0132 | 0.7310 | 0.0026 |
| 50 | 2.4930 | 0.1146 | 1.8355 | 0.0113 | 0.7389 | 0.0014 |
| 70 | 2.6819 | 0.1011 | 1.9200 | 0.0063 | 0.7557 | 0.0011 |
| 100 | 2.7654 | 0.0550 | 1.9227 | 0.0059 | 0.7669 | 0.0005 |

## 7. Conclusion

In this paper, we have examined the classical and Bayesian inference procedures for the BT XII distribution parameters, as well as the corresponding reliability parameter when the available data are described regarding fuzzy numbers. In this context, we considered three priors as **noninformative prior**, i.e. $a_1 = b_1 = a_2 = b_2 = 0$, **less informative prior**, i.e. $a_1 = b_1 = a_2 = b_2 = 0.01$, and **informative prior**, i.e. $a_1 = b_1 = a_2 = b_2 = 4$. The general results can be made from Tables 1-3 as follows. Considering the criterion MSE for all methods, with increasing $n$, the estimates are improved. The performance of the Bayes estimates with assumptions of noninformative prior and less informative prior regarding

AVs and MSEs, are almost identical. So, we prefer the prior *X*II since it will make the priors proper. The simulation study for all methods shows that the estimate of $R$ is satisfactory, even for samples with sizes small and moderate. Using the NR or EM algorithms for the computation of MLEs gives similar estimation results. Because these two procedures have different features in the complexity of the iterative numerical search, we let users choose which to be used based on their preferences. The Bayes estimates obtained by Tierney and Kadane's approximation and the MCMC method behave in a very similar manner. However, from the computational point of view, Tierney and Kadane's procedure is easier to obtain. Note that these estimation results cannot be attributed to the assumed fuzzy numbers in Fig. 2. We have implemented the estimation procedures for different fuzzy numbers (not reported here) and found that the rationale for such fuzzy numbers, which are characterized by the membership functions $\mu_{\tilde{x}}(.)$ will not influence the estimate results.

Table 2: The AV and MSE of the estimate of parameters $\theta = (c,k) = (2,3)$, and $R = 0.5120$.

| n | | | | *priorI* | | |
|---|---|---|---|---|---|---|
| | AV(c) | MSE(c) | AV(k) | MSE(k) | AV(R) | MSE(R) |
| 10 | 2.0763 | 0.0333 | 6.2226 | 1.3855 | 0.3228 | 0.0357 |
| 20 | 2.0096 | 0.0058 | 3.9389 | 1.1535 | 0.4333 | 0.0126 |
| 30 | 1.9341 | 0.0043 | 4.0740 | 0.8816 | 0.3995 | 0.0067 |
| 40 | 2.0035 | 0.0029 | 3.8693 | 0.7557 | 0.4298 | 0.0061 |
| 50 | 1.8174 | 0.0002 | 3.0318 | 0.0300 | 0.4725 | 0.0015 |
| 70 | 1.9459 | 0.0002 | 3.1732 | 0.0010 | 0.4832 | 0.0008 |
| 100 | 1.9836 | 0.0001 | 3.0189 | 0.0003 | 0.5076 | 0.0001 |
| n | | | | *priorII* | | |
| 10 | 2.0566 | 0.0309 | 6.0537 | 9.3255 | 0.3306 | 0.0328 |
| 20 | 2.0021 | 0.0049 | 3.9133 | 1.1654 | 0.4336 | 0.0129 |
| 30 | 1.9297 | 0.0032 | 4.0795 | 0.8342 | 0.3982 | 0.0068 |
| 40 | 1.9996 | 0.0031 | 3.8646 | 0.7476 | 0.4292 | 0.0061 |
| 50 | 1.8241 | 0.0001 | 3.0440 | 0.0260 | 0.4727 | 0.0015 |
| 70 | 1.9438 | 0.0002 | 3.1613 | 0.0019 | 0.4841 | 0.0007 |
| 100 | 1.9867 | 0.0001 | 3.0146 | 0.0002 | 0.5088 | 0.0001 |
| n | | | | *priorIII* | | |
| 10 | 1.4479 | 0.3048 | 2.1876 | 0.6598 | 0.5174 | 0.0017 |
| 20 | 1.6671 | 0.1170 | 2.4551 | 0.2968 | 0.5173 | 0.0010 |
| 30 | 1.6578 | 0.1108 | 2.7872 | 0.1753 | 0.4702 | 0.0001 |
| 40 | 1.7714 | 0.0880 | 2.8966 | 0.0610 | 0.4794 | 0.0001 |
| 50 | 1.7032 | 0.0522 | 2.5812 | 0.0487 | 0.5034 | 0.0001 |
| 70 | 1.8432 | 0.0245 | 2.7791 | 0.0452 | 0.5069 | 0.0001 |
| 100 | 1.9136 | 0.0074 | 2.7528 | 0.0106 | 0.5242 | 0.0001 |

Table 3: The AV and MSE of the estimate of parameters $\theta = (c,k) = (1,1)$, and $R = 0.6667$.

| n | | | | *priorI* | | |
|---|---|---|---|---|---|---|
| | AV(c) | MSE(c) | AV(k) | MSE(k) | AV(R) | MSE(R) |
| 10 | 1.4393 | 0.1930 | 2.0352 | 1.0717 | 0.5414 | 0.0156 |
| 20 | 1.2573 | 0.1049 | 1.3105 | 0.0964 | 0.6350 | 0.0053 |
| 30 | 1.1349 | 0.0873 | 1.1559 | 0.0243 | 0.6497 | 0.0036 |
| 40 | 1.2375 | 0.0662 | 1.1183 | 0.0140 | 0.6740 | 0.0011 |
| 50 | 1.1726 | 0.0564 | 0.9733 | 0.0110 | 0.6999 | 0.0009 |
| 70 | 1.2954 | 0.0298 | 0.9305 | 0.0048 | 0.7274 | 0.0002 |
| 100 | 1.3239 | 0.0182 | 0.8951 | 0.0007 | 0.7399 | 0.0001 |
| n | | | | *priorII* | | |
| 10 | 1.4468 | 0.1996 | 2.0446 | 1.0913 | 0.5417 | 0.0156 |
| 20 | 1.2510 | 0.1058 | 1.3056 | 0.0934 | 0.6348 | 0.0052 |
| 30 | 1.1340 | 0.0891 | 1.1553 | 0.0241 | 0.6497 | 0.0036 |
| 40 | 1.2430 | 0.0630 | 1.1158 | 0.0134 | 0.6754 | 0.0012 |
| 50 | 1.1782 | 0.0590 | 0.9689 | 0.0103 | 0.7019 | 0.0010 |
| 70 | 1.2986 | 0.0317 | 0.9340 | 0.0043 | 0.7269 | 0.0002 |
| 100 | 1.3254 | 0.0179 | 0.8980 | 0.0009 | 0.7393 | 0.0001 |
| n | | | | *priorIII* | | |
| 10 | 1.2976 | 0.0903 | 1.5289 | 0.2797 | 0.5985 | 0.0046 |
| 20 | 1.2102 | 0.0886 | 1.2267 | 0.0514 | 0.6454 | 0.0045 |
| 30 | 1.1188 | 0.0732 | 1.1325 | 0.0175 | 0.6527 | 0.0029 |
| 40 | 1.2164 | 0.0468 | 1.1010 | 0.0102 | 0.6750 | 0.0009 |
| 50 | 1.1589 | 0.0441 | 0.9768 | 0.0086 | 0.6969 | 0.0004 |
| 70 | 1.2705 | 0.0252 | 0.9427 | 0.0032 | 0.7210 | 0.0002 |
| 100 | 1.3005 | 0.0141 | 0.9071 | 0.0005 | 0.7339 | 0.0001 |

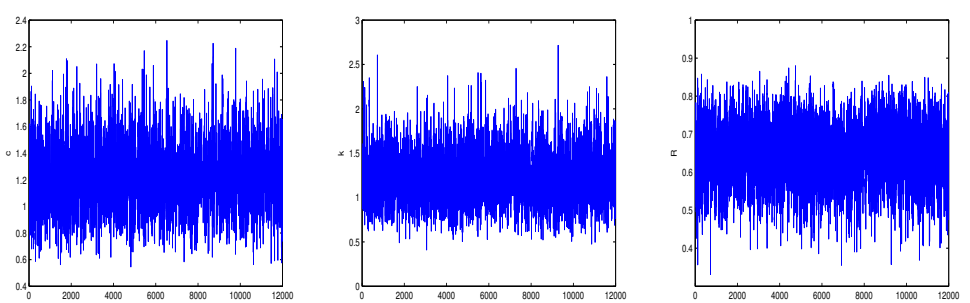


Figure 3: Plots of generated c versus iteration of MCMC (Gibss algorithm).

## Acknowledgements

The authors would like to thank Editors and the referees for their constructive comments and suggestions which improved and enriched the presentation of the paper.

## References

Akbari, M. G., Rezaei, A., (2007). A uniformly minimum variance unbiased point estimator using fuzzy observations. *Austrian Journal of Statistics*, 36 (4), pp. 307–317.

Ali Mousa, M. A. M., Jaheen, Z. F., (2002). Statistical inference for the Burr model based on progressively censored data functions. *Comput. Math. Applic.*, 43, pp. 1441–1449.

Burr, I. W., (1942). Cumulative frequency functions. *Annals of Mathematical Statistics*, 13 , pp. 215–222.

Burr, I. W., (1973). Parameters for a general system of distributions to match a grid of $\alpha_3$and $\alpha_4$. *Communications in Statistics-Theory and Methods*, 2 , pp. 1–21.

Coppi, R., Gilb, M. A., Kiersc, H. A. L., (2006). The fuzzy approach to statistical analysis. *Computational Statistics and Data Analysis*, 51 (1), pp. 1–14.

Denoeux, T., (2011). Maximum likelihood estimation from fuzzy data using the EM algorithm, *Fuzzy Sets and Systems*, doi:10.1016/j.fss.2011.05.022.

Gertner, G. Z., Zhu, H., (1996). Bayesian estimation in forest surveys when samples or prior information are fuzzy. *Fuzzy Sets and Systems.*, 77(3), pp. 277–290.

Gil, M. A., López-Diaz, M., Ralescu, D. A., (2006). Overview on the development of fuzzy random variables. *Fuzzy Sets and Systems*, 157, pp. 2546–2557.

Hanagl, D. D., Ahmadi, K. A., (2009). Bayesian estimation of the parameters of bivariate exponential distribution, *Communication in Statistics-Simulation and Computation*, 38, pp. 1391–1413.

Hate, M. A., (1949). A certain cumulative probability function. *Ann. Math. Statist.*, 20, pp. 461–63.

Huang, H., Zuo, M., Sun, Z., (2006). Bayesian reliability analysis for fuzzy lifetime data. *Fuzzy Sets and Systems*, 157(3), pp. 1674–1686.

Moore, D., Papadopoulos, A. S., (2000). The Burr type XII distribution as a failure model under various loss functions. *Microelectron. Reliab.*, 40, pp. 2117–2122, doi: 10.1016/S0026-2714(00)00031-7.

Pak, A., Parham, G. H., Saraj, M., (2013). Inference for the Weibull Distribution Based on Fuzzy Data. *Revista Colombiana de Estadistica*, 36(2), pp. 339–358.

Pak, A., Parham, G. H., Saraj, M., (2014). Inferences on the Competing Risk Reliability Problem for Exponential Distribution Based on Fuzzy Data. *IEEE Transactions on reliability*, 63(1), pp. 2–13.

Pak, A., (2016). Statistical Inference for the Parameter of Lindley Distribution Based on Fuzzy data. To appear in *Brazilian Journal of Probability and Statistics*.

Rodriguez, R. N., (1977). A guide to the Burr type XII Distributions. *Biometrika*, 64(1) , 129-134, doi:10.1093/biomet/64.1.129.

Rubinstein, R. Y., Kroese, D. P., (2006). *Simulation and the Monte Carlo method*. 2nd edition, John Wiley and Sons, Inc., Hoboken, New Jersey.

Shafiq, M., Viertl, R. (2014). Maximum likelihood estimation for Weibull distribution in case of censored fuzzy life time data. http://www.statistik.tuwien.ac.at/forschung/SM/SM-2014-2complete.pdf. pp. 1–17.

Shafiq, M., Atif, M., (2015). On the survival models for step-stress experiments based on fuzzy life time data. *Qual Quant*, doi: 10.1007/s11135-015-0295-9.

Singh S., Singh U., Kumar M., (2016). Bayesian estimation for Poission-exponential model under Progressive type-II censoring data with Binomial removal and its application to ovarian cancer data, *Communications in Statistics-Simulation and Computation*, 45, pp. 3457–3475.

Singpurwalla, N. D., Booker, J. M., (2004). Membership functions and probability measures of fuzzy sets, *Journal of the American Statistical Association*, 99(467), pp. 867–877.

Soliman, A. A., (2005). Estimation of parameters of life from progressively censored data using Burr-XII model. *IEEE Trans. Reliab.*, 54, 34–42.

Tadikamalla, P. R., (1980). A Look at the Burr and Related Distributions. *International Statistical Review*, 48(3) , pp. 337–344.

Tanaka, H., Okuda, T., Asai, K., (1979). Fuzzy information and decision in statistical model. *In: Advances in Fuzzy Sets Theory and Applications*, North-Holland, Amsterdam, pp. 303–320.

Tierney, L., Kadane, J. B., (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, pp. 82–86.

Viertl, R., (2011). *Statistical Methods for Fuzzy Data*, Wiley, Chichester.

Viertl, R., (2006). Univariate statistical analysis with fuzzy data. *Computational Statistics & Data Analysis*, 55(1), pp. 133–147.

Wingo, D. R., (1983). Maximum likelihood methods for fitting the Burr Type XII distribution to life test data. *Biom J.*, 25, pp. 77–84.

Wingo, D. R., (1993). Maximum likelihood methods for fitting the Burr type XII distribution to multiply(progressively) censored life test data. *Metrika*, 40, pp. 203–210.

Wu, H. C. (2004), Fuzzy Bayesian estimation on lifetime data. *Computational Statistics*, 19, pp. 613–633.

Wu, J. W., Yu, H. Y., (2005). Statistical inference about the shape parameter of the Burr type XII distribution under the failure-censored sampling plan. *Applied Math. Computat.*, 163, pp. 443–482.

Xiuchun, L., S. Yimin, W. Jieqiong, C., Jian, (2007). Empirical Bayes estimators of reliability performances using LINEX loss under progressively Type-II censored samples. *Math. Comput. Simulat.*, 73, pp. 320–326.

Zadeh, L., (1965). Fuzzy sets *Information and Control*, 8(3), pp. 338–353.

Zadeh, L. A., (1968). Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications*, 10, pp. 421–427.

Zarei, R., Amini, M., Taheri, S.M., Rezaei, A.H., (2012). Bayesian estimation based on vague lifetime data. *Soft Computing*, 16, pp. 165–174.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 145–165,* https://doi.org/10.59170/stattrans-2024-009
Received – 14.09.2022; accepted – 11.07.2023

# Estimation of quantiles with the exact bootstrap method

**Joanna Kisielińska**[1]

## Abstract

A problem with the estimation of quantiles occurs when the sample comes from an unknown distribution. The estimation uses the bootstrap method in the version that the literature refers to as exact. Three bootstrap estimators were used: two of them based on one order statistic, and the third on a linear combination of two order statistics (for an integer). The distribution of the exact bootstrap estimator based on a single order statistic is known. It has been shown that there is no general form of the distribution of the exact bootstrap estimator based on two order statistics. However, it is possible to calculate such a distribution – the article presents the algorithm that performs such a task. The bootstrap confidence intervals were constructed using the exact percentile method. It has been shown that if the estimator is based on a single order statistic, it is known in advance which elements of the primary sample are the limits of the confidence intervals, so there is no need to resample. The intervals determined by the exact percentile method were compared with those constructed using other methods. It has been shown that the information on the direction of the asymmetry of the distribution that the sample comes from is worth considering when selecting the rank of the order statistic used as an estimator. Attention is paid to the influence of the quality of the pseudorandom number generators on the results of the Monte Carlo simulation.

**Key words:** quantile estimation, confidence intervals for quantile, exact bootstrap method, exact percentile method, Monte Carlo method.

## 1. Introduction

Let $X$ be a continuous random variable with cumulative distribution (CDF) $F(x)$ and density function (PDF) $f(x)$. Let $p \in (0,1)$ be given and let $\xi_p$ be $p$-quantile of $F$, such that $p = F(\xi_p)$, $\xi_p = F^{-1}(p)$, and $f(\xi_p)$ (e.g. Bahadur (1966, p. 577), Nagaraja and Nagaraja (2020, p. 75)). Bahadur (1966, p. 577), gives the conditions to be satisfied by $F$ so that $\xi_p$ be unique.

The $p$-quantile is most often defined as the left quantile (e.g. Serfling, 1980, p. 3):

$$\xi_p = F^{-1}(p) = inf\{x: F(x) \geq p\}. \tag{1}$$

[1] Warsaw University of Life Sciences, Poland. E-mail: joanna_kisielinska@sggw.edu.pl.
ORCID: https://orcid.org/0000-0003-3289-1525.

Sample quantiles are used to estimate quantiles. For a sample $(X_1, X_2, \dots, X_n)$ from distribution $F$ Serfling (1980, p. 74) defines the sample $p$-quantile $\xi_{pn}$ as $p$-quantile of empirical distribution $F_n$:

$$\xi_{pn} = F_n^{-1}(p) = inf\{x: F_n(x) \geq p\}. \tag{2}$$

Sample $p$-quantiles, used as quantile estimators (i.e. when $\hat{\xi}_{pn} = \xi_{pn}$), are presented using order statistics. Let $X_{nz}$ denote the $z$th order statistic, i.e. the smallest $z$th element of the sample of size $n$, and then (Serfling, 1980, p. 88):

$$\hat{\xi}_{pn} = \begin{cases} X_{n,np,} & \text{if } np \text{ is integer} \\ X_{n,[np]+1,} & \text{if } np \text{ is not integer'} \end{cases} \tag{3}$$

where $[\cdot]$ denotes the floor function.

Nagaraja and Nagaraja (2020, p. 75) identify the sample p-quantile with the following order statistic:

$$\hat{\xi}_{pn} = X_{n,[np]+1}. \tag{4}$$

Hyndman and Fan (1996, p. 361) give many other definitions of sample quantiles based on order statistics. Their general form is:

$$\xi_{pn} = (1-\gamma)X_{nj} + \gamma X_{n,j+1,} \tag{5}$$

where $\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$ for some $m \in$ R and $0 \leq \gamma \leq 1$. The $\gamma$ parameter is a function of $j$ and $g$, where $j = [pn + m]$ and $g = pn + m - j$. Formula (5) includes the definitions (3) and (4).

With some assumptions (Serfling, 1980 p. 74), the sample $p$-quantile $\hat{\xi}_{pn}$ defined by (2) is strongly consistent for estimation of $\xi_p$.

It is known that the sample $p$-quantile is asymptotically normal if $f$ is continuous and positive at $\xi_p$ (e.g. Serfling, 1980, p. 77), (Nagaraja and Nagaraja, 2020, p. 77). The limit distribution has a mean $\xi_p$ and a variance $\frac{p(1-p)}{f^2(\xi_p)n}$. The sample quantile vector $\left(\hat{\xi}_{p1}, \dots, \hat{\xi}_{pk}\right)$ is also asymptotically normal for $0 < p_1 < \cdots < p_k < 1$ if $f$ is continuous and positive at $\xi_{p1}, \dots, \xi_{pk}$. The parameters of this distribution are a mean vector $\left(\xi_{p1}, \dots, \xi_{pk}\right)$ and a covariance matrix with elements: $\left(\frac{p_i(1-p_j)}{f\left(\xi_{p_i}\right)f\left(\xi_{p_j}\right)n}\right)$ (Serfling, 1980 p. 80). The consequence of the asymptotic normality of the sample quantile vector is the asymptotic normality of any linear combination of these quantiles.

Serfling (1980, p. 94) based on the Bahadur (1966) article indicates that the order statistic $X_{nk_n}$ (where $\{k_n\}$ is a sequence of positive integers ($1 \leq k_n \leq n$) such that $k_n/n$ tends to $p$ sufficiently fast) and the sample $p$-quantile $\hat{\xi}_{pn}$ are roughly equivalent as estimates of $\xi_p$. Despite this, in the general case difference $X_{nk_n} - \xi_p$ has a limit normal distribution not centered at 0 (Serfling, 1980, p. 94).

If the sample of size $n$ comes from a continuous distribution with CDF $F(x)$ and PDF $f(x)$, then the PDF of the $z$th order statistics $X_{nz}$ is (e.g. David and Nagaraja, 2003, p. 10); (Evans, Leemis and Drew, 2006, p. 20):

$$f_{X_{nz}}(x) = \frac{n!}{(z-1)!(n-z)!} f(x)[F(x)]^{z-1}[1-F(x)]^{n-z} \tag{6}$$

This formula is the same as that given by Serfling (1980 p. 85), which specifies PDF of the sample $p$-quantile.

The practical application of the expression (6) is cumbersome (Serfling, 1980, p. 87). First, it requires knowledge of the distribution the sample comes from, and secondly, the distribution of order statistics is not usually in the class of known and commonly used distributions. Pekasiewicz (2015, p. 23) gives the density functions of the order statistics $X_{nz}$ for selected distributions the sample comes from. Using limit distribution is also troublesome due to the necessity of knowing $f(\xi_p)$. The bootstrap method proposed by Efron in 1979 does not have these disadvantages. It does not require knowledge of the distribution a sample comes from. Falk and Kaufmann (1991), Falk and Reiss (1989), Bickel and Freedman (1981) and Singh (1981) (among others) studied the convergence of the bootstrap estimators of the parameters (also quantiles). They showed that bootstrap error converges to 0 with probability one. This indicates the correctness of this approach, although one can discuss the order of this convergence.

In the bootstrap method, empirical distribution $F_n$ is an estimator of the distribution $F$. And therefore, the bootstrap estimator distribution (dependent on the $F_n$) is an estimator of the estimator distribution (dependent on the $F$). Efron (1979, p. 4) proposes three methods of computing the bootstrap estimator distribution. The first is a theoretical calculation, the second is the Monte Carlo (MC) approximation, and the third is the Taylor series expansion. The MC approximation invloves selecting many resamples of size $n$ with replacement from the $n$-element primary sample. Fisher and Hall (1991) pointed out that instead of drawing resamples[2] (especially for small samples), one can generate all resamples. One can then determine all the realizations of the bootstrap estimator. This method was called the exact bootstrap method in order to distinguish it from the commonly used MC approximation with resampling. It should be noted that the distributions of the bootstrap estimators gained with the exact bootstrap method are equivalent to those obtained with the first method proposed by Efron. The difference is only in the method of their determination. The exact method relies on numerical calculations, Efron method on theoretical calculations. In the following considerations, the bootstrap method based on all resamples will be called the exact method, no matter how the calculations were made.

[2] There are $n^n$ resamples in total, but the different resamples are $\binom{2n-1}{n}$ (Fisher and Hall 1991 p. 160). To calculate the number of resamples with the same elements, one should compute the number of its permutations. One should permute only the elements on positions with non-repeating elements.

Taking into consideration all resamples allows to eliminate errors caused by resampling (Hutson and Ernst, 2000, p. 94). Resampling may be interpreted as drawing bootstrap samples from their entire population ($n^n$). In the MC approximation, some resamples may be omitted, while others used multiple times. Kisielinska (2013, p. 1068) presented the comparison of the exact bootstrap method and the bootstrap method with resampling for any parameter.

The bootstrap $p$-quantile estimators $\hat{\xi}^*_{pn}$ are also based on order statistics. The order statistics of the bootstrap resample $X^*_{nz}$ is its $z$th smallest element. Evans, Leemis, Drew (2006, p. 23) give the distribution of such statistic – a case of a finite population, sampling with replacement. The distribution thus determined is of course the exact bootstrap distribution (formula (16) in section 2).

The bootstrap method is not the only method for estimating quantiles, which does not require to know the distribution the sample comes from. For an ample review of distribution-free methods to construct confidence intervals, see Nagaraja and Nagaraja (2020).

Confidence intervals for quantiles can be determined using an asymptotic approach based on the sample $p$-quantile. In simulation experiments, the samples come from a known distribution, and therefore the values of $\xi_p$ and $f(\xi_p)$ are known. One can determine 1-$\alpha$ confidence interval of the sample $p$-quantile from the limit distribution:

$$I^{Aa}_{pn} = [F_A^{-1}(\propto/2), F_A^{-1}(1-\propto/2)] \tag{7}$$

where $F_A$ is the normal distribution with mean $\xi_p$ and variance $\frac{p(1-p)}{f^2(\xi_p)n}$.

Serfling (1980, p. 130) proposes to use an asymptotic approach based on order statistics to determine confidence intervals for quantiles. The confidence interval is as follows:

$$I^{Ab}_{pn} = \left[X_{nk_{1n}}, X_{nk_{2n}}\right] \tag{8}$$

where $k_{1n} = n \cdot \left(p - \frac{u_\propto \sqrt{p(1-p)}}{\sqrt{n}}\right)$, $k_{2n} = n \cdot \left(p + \frac{u_\propto \sqrt{p(1-p)}}{\sqrt{n}}\right)$, and $u_\alpha$ is the 100·(1-α/2)th percentile point of standard normal distribution. If $n \to \infty$ confidence coefficient of the interval $I^{Ab}_{pn} \to$ 1-$\alpha$ (Serfling, 1980 p. 104).

The percentile method enables to construct confidence intervals when using the bootstrap approach. The main objections to this method relate to applications in the cases of small samples. Many authors note that the percentile method produces confidence intervals of first-order accuracy only (e.g Falk and Kaufmann (1991), Efron and Tibshirani (1993), Nagaraja and Nagaraja, 2020). For this reason, many proposals for better solutions have been created.

Efron (1987) proposed the BCa method (i.e. bias-corrected and accelerated), which is second-order accurate (Efron and Tibshirani, 1993) and has higher coverage probability.

Nagaraja and Nagaraja (2020) proposed a method of adjacent spacings to construct quantiles confidence intervals. The method is easy to apply, yet it requires experimental selection of two parameters *s* and *t*, and knowledge of the critical values of the statistics $W_{(s+t)}$. Critical values of the $W_{(s+t)}$ are given by Nagaraja and Nagaraja (2020, p. 88). Parameters *s* and *t* determine the rank of order statistics, used to construct the confidence interval. Nagaraja and Nagaraja (2020) conducted simulation studies to compare quantiles confidence intervals obtained using various distribution free-methods. They assessed the effectiveness of the methods based on the width of confidence intervals and coverage probability.

The problem presented in the article is in the estimation of quantiles when a distribution the sample comes from is not known. The novelty of the approach presented in the paper consists in estimating the quantiles with an exact bootstrap quantile estimator based on a linear combination of two order statistics. The algorithm presented in Section 2 allows for determining its distribution exactly, not only in an approximate manner. It is also shown that in the case of quantile estimation, confidence intervals are much easier to determine with the exact percentile method than with the percentile method with resampling, if the estimator is based on a single order statistics. Moreover, it is shown that the information about the direction of asymmetry of the distribution the sample comes from can be used to determine the rank of the single order statistics used as an estimator.

All calculations were made in Excel using the VBA language for Application.

## 2. Distributions of quantiles bootstrap estimators

Let the *n*-element resample, drawn with replacement from the original sample $(x_1, x_2, \dots, x_n)$, be marked as $(X_1^*, X_2^*, \dots, X_n^*)$. Each variable $X_i^*$ has a discrete empirical distribution $F_n$. Efron, 1979 assumed equal probabilities $p_i = 1/n$ for each element of the primary sample $x_i$. Due to a finite measurement accuracy of any values, elements of the observed sample can be repeated. The empirical distribution is determined by probabilities $p_i$ for each $x_i$ where $\sum_{i=1}^{k} p_i = 1$ and $k$ is the number of distinct elements in a primary sample.

The elements of the resample are the discrete random variables $X_i^*$ with PDF *fn*, CDF *Fn*, and survival function (SF) *Sn*:

$$f_n(x) = P(X_i^* = x) = \begin{cases} p_j & x = x_j, j = 1, \dots, k \\ 0 & \text{for others } x \in R \end{cases}, \tag{9}$$

$$F_n(x) = P(X_i^* \le x) = \sum_{j-1; x_j \le x}^{k} p_j, \tag{10}$$

$$S_n(x) = P(X_i^* \ge x) = 1 - F_n(x) + f_n(x). \tag{11}$$

The bootstrap quantile estimator according to (5) can be written in the general form as:

$$\hat{\xi}^{*}_{pn} = (1-\gamma)X^{*}_{nj} + \gamma X^{*}_{n,j+1,} \tag{12}$$

where: $0 \le \gamma \le 1$, $j=[np]$ (wherein $[np]=np$ for integer $np$), and $X^{*}_{nj}$ is the $j$th order statistics of the resample.

In the research three bootstrap quantile estimators were used:

$$\hat{\xi}^{1*}_{pn} = \begin{cases} X^{*}_{n,np,} & \text{if } np \text{ is integer} \\ X^{*}_{n,[np]+1} & \text{if } np \text{ is not integer} \end{cases}, \tag{13}$$

$$\hat{\xi}^{2*}_{pn} = X^{*}_{n,[np]+1}, \tag{14}$$

$$\hat{\xi}^{3*}_{pn} = \begin{cases} (1-\varepsilon)X^{*}_{n,np} + \varepsilon X^{*}_{n,np+1,} & \text{if } np \text{ is integer} \\ X^{*}_{n,[np]+1} & \text{if } np \text{ is not integer} \end{cases}, \tag{15}$$

where: $\varepsilon = (n+1)p-[(n+1)p]$ as Hutson 2002 p. 332 suggests.

The estimator $\hat{\xi}^{1*}_{pn}$ was obtained assuming $\gamma = 0$ for $np$ integer and $\gamma = 1$ for $np$ not integer, the estimator $\hat{\xi}^{2*}_{pn}$ assuming $\gamma = 1$, and the estimator $\hat{\xi}^{3*}_{pn}$ assuming $\gamma=\varepsilon$ for $np$ integer and $\gamma = 1$ for $np$ not integer.

In the formulae given below, it was assumed that the primary sample is ordered, viz. $x_{1\le} x_2, \ldots, x_{n-1} \le x_n$.

Distributions of bootstrap quantile estimators based on one order statistics result directly from the formula given by Evans, Leemis, Drew (2006, p. 23) and are as follows:

$$P(X^{*}_{nz} = x_l) = \begin{cases} \text{for } l = 1 \\ \sum_{w=0}^{n-z} \binom{n}{w} [f_n(x_1)]^{n-w} [S_n(x_2)]^{w} \\ \text{for } l = 2, \ldots, k-1 \\ \sum_{u=0}^{z-1} \sum_{w=0}^{n-z} \binom{n}{u, n-u-w, w} [F_n(x_{l-1})]^{u} [f_n(x_l)]^{n-u-w} [S_n(x_{l+1})]^{w} \\ \text{for } l = k \\ \sum_{u=0}^{z-1} \binom{n}{u} [F_n(x_{k-1})]^{u} [f_n(x_k)]^{n-u} \end{cases} \tag{16}$$

where $z$ is the rank of the order statistic used as the estimator.

When $np$ is not integer the estimators $\hat{\xi}^{1*}_{pn}$, $\hat{\xi}^{2*}_{pn}$, and $\hat{\xi}^{3*}_{pn}$ are the same. The rank of the order statistic used as the bootstrap estimator of $p$-quantile is $z=[np] + 1$. When $np$ is not integer, the rank of the order statistic used as the bootstrap estimator of $p$-quantile is $z=np$ for estimator $\hat{\xi}^{1*}_{pn}$ and $z=np$ for estimator $\hat{\xi}^{2*}_{pn}$.

Only elements of a primary sample can be realizations of the estimators based on one order statistics. Realizations of the estimator in the form of a linear combination of two order statistics may also be weighted means of all two-element combinations

chosen therefrom. This means that the estimator $\hat{\xi}_{pn}^{3*}$ for integer $np$ has a considerably higher number of realizations. It is impossible to give a general expressions determining these estimator. Nevertheless, one can determine the probabilities that on positions $np$ and $np$+1 in resamples either any $l$th primary sample element will occur or any two of its elements: $l_1$ and $l_2$, with $l_1 < l_2$.

The probability that in an ordered resample element $x_l$ occurs at least on two positions $z = np$ and $z + 1$ is:

$$P\left((X_{nz}^* = x_l) \wedge \left(X_{n,z+1}^* = x_l\right)\right) =$$

$$= \begin{cases} \text{for } l = 1 \\ \sum_{w=0}^{n-z-1} \binom{n}{w} [f_n(x_1)]^{n-w} [S_n(x_2)]^w \\ \text{for } l = 2, \dots, k-1 \\ \sum_{u=0}^{z-1} \sum_{w=0}^{n-z-1} \binom{n}{u, n-u-w, w} [F_n(x_{l-1})]^u [f_n(x_l)]^{n-u-w} [S_n(x_{l+1})]^w \\ \text{for } l = k \\ \sum_{u=0}^{z-1} \binom{n}{u} [F_n(x_{k-1})]^u [f_n(x_k)]^{n-u} \end{cases} \quad (17)$$

The probability that in an ordered resample, element $x_1$ occurs exactly $z$ times, and the element $x_l$, for $l = 2, \dots, k-1$ occurs at least once on position $z$ +1 is equal to:

$$P\left((X_{nz}^* = x_1) \wedge \left(X_{n,z+1}^* = x_l\right)\right) =$$

$$= \sum_{w=0}^{n-z-1} \binom{n}{z, n-z-w, w} [F_n(x_1)]^z [f_n(x_l)]^{n-z-w} [S_n(x_{l+1})]^w. \quad (18)$$

The probability that in an ordered resample, element $x_l$, for $l = 2, \dots, k-1$ occurs at least once on position $z$, and element $x_k$ occurs exactly $n$-$z$ times, is:

$$P\left((X_{nz}^* = x_l) \wedge \left(X_{n,z+1}^* = x_k\right)\right) =$$

$$= \sum_{u=0}^{z-1} \binom{n}{u, z-u, n-z} [F_n(x_{l-1})]^u [f_n(x_l)]^{z-u} [S_n(x_k)]^{n-z}. \quad (19)$$

The probability that element $x_1$ occurs in an ordered resample exactly $z$ times, and the elements $x_k$ exactly $n$-$z$ times, is:

$$P\left((X_{nz}^* = x_1) \wedge \left(X_{n,z+1}^* = x_k\right)\right) = \binom{n}{z, n-z} [f(x_1)]^z [f_n(x_k)]^{n-z}. \quad (20)$$

The probability that in an ordered resample element $x_{l_1}$ occurs at least once on position $z$, and element $x_{l_2}$ at least once on position $z$+1 $\bigwedge_{l_1<l_2}(l_1<l_2) \in \{2,3,..,k-2\} \times \{3,4,..,k-1\}$, is:

$$P\left(\left(X_{nz}^* = x_{l_1}\right) \wedge \left(X_{n,z+1}^* = x_{l_2}\right)\right) =$$

$$= \sum_{u=0}^{z-1} \sum_{w=0}^{n-z-1} \binom{n}{u, z-u, n-z-w, w} W \quad (21)$$

$$W = \left[F_n\left(x_{l_1-1}\right)\right]^u \left[f_n\left(x_{l_1}\right)\right]^{z-u} \left[f_n\left(x_{l_2}\right)\right]^{n-z-w} \left[S_n\left(x_{l_2+1}\right)\right]^w$$

The exact distribution of bootstrap $p$-quantile estimator based on two order statistics (estimator $\hat{\xi}_{pn}^{3*}$ when $np$ is an integer) is:

$$P\left(\hat{\xi}_{pn}^{3*} = x_l\right) = P\left(\left(X_{nz}^* = x_l\right) \wedge \left(X_{n,z+1}^* = x_l\right)\right), \text{ for } l = 1, \dots, k$$
$$P\left(\hat{\xi}_{pn}^{3*} = (1-\varepsilon)x_1 + \varepsilon x_l\right) = P\left(\left(X_{nz}^* = x_1\right) \wedge \left(X_{n,z+1}^* = x_l\right)\right),$$
$$\text{for } l = 2, \dots, k-1$$
$$P\left(\hat{\xi}_{pn}^{3*} = (1-\varepsilon)x_l + \varepsilon x_k\right) = P\left(\left(X_{nz}^* = x_l\right) \wedge \left(X_{n,z+1}^* = x_k\right)\right),$$
$$\text{for } l = 2, \dots, k-1$$
$$P\left(\hat{\xi}_{pn}^{3*} = (1-\varepsilon)x_1 + \varepsilon x_k\right) = P\left(\left(X_{nz}^* = x_1\right) \wedge \left(X_{n,z+1}^* = x_k\right)\right)$$
$$P\left(\hat{\xi}_{pn}^{3*} = (1-\varepsilon)x_{l_1} + \varepsilon x_{l_2}\right) = P\left(\left(X_{nz}^* = x_{l_1}\right) \wedge \left(X_{n,z+1}^* = x_{l_2}\right)\right),$$
$$\bigwedge_{l_1<l_2}(l_1<l_2)\in\{2,3,..,k-2\}\times\{3,4,..,k-1\}. \tag{22}$$

Some realizations of the $\hat{\xi}_{pn}^{3*}$ estimator may repeat themselves, therefore the probabilities corresponding to these realizations should be added. As the number and order of the ordered realizations of the estimator based on two order statistics depend on the primary sample, one cannot give the general form of its distribution.

The algorithm for determining the distribution of the estimator based on two order statistics is as follows:

1. For each pair $(l_1, l_2) \in \{1,2, \dots, k\} \times \{1,2, \dots, k\}$ such that $l_1 \le l_2$, the corresponding realization of the estimator should be calculated: $y_j = (1-\varepsilon)x_{l_1} + \varepsilon x_{l_2}$ and probability $P\left(\left(X_{nz}^* = x_{l_1}\right) \wedge \left(X_{n,z+1}^* = x_{l_2}\right)\right)$.
2. Calculate the sum of probabilities determined in point 1 for each unique $y_j$.
3. If necessary, the estimator realizations should be sorted (e.g. to use the percentile method).

The presented algorithm allows for an exact calculation of the distribution of a linear combination of two consecutive order bootstrap statistics. Nagaraja and Nagaraja (2020, p. 81) based on the previous work of other authors (Nyblom, 1992), (Hettmansperger and Sheather, 1986) give the formulas that allow calculating this distribution approximately.

A useful attribute of quantile estimators distributions based on single order statistics is that the probabilities for all estimator realizations are the same for all primary samples of a given size, provided that $k = n$ (probabilities given by the expression (16) depend only on $n$ and $p$). Distributions of estimators in the form of a linear combination of two order statistics do not have such property. It is worth noting, however, that the probabilities given by the expressions (17)–(21) also depend only on $n$ and $p$ – if there were no repetition in the sample. The occurrence of repetitions only causes that the probabilities for an element occurring multiple times in the sample are added together.

Knowing the exact bootstrap distribution of the quantile estimator can be useful for constructing confidence intervals for quantile. Determining the expected value and variance does not require knowing it. These values can be calculated using exact analytical expressions for any *L*-estimator given by Hutson and Ernst (2000). All variants of estimators (13)–(15) are *L*-estimators. If an estimator is based on a single order statistics, the expressions for multipliers for a mean (Hutson and Ernst (2000, p. 91)) are equivalent to the probabilities given in (16). If the estimator is based on two order statistics, the multipliers can be easily obtained from formulas (17)–(21). It is worth recalling that the expressions for the mean and variance of the median bootstrap estimators were given by Maritz and Jarrett as early as 1978. It was before Efron presented the concept of the bootstrap method.

## 3. Confidence intervals for quantiles by the exact bootstrap percentile method

One may construct the quantiles bootstrap confidence intervals by the percentile method described in the paper Wilcox (2001, p. 88), among others. It should be noted that the resamples do not need to be drawn from their entire population (size $n^n$). The distributions of the bootstrap quantile estimators can be calculated. On this basis, one may easily find the limits of the confidence interval. We know in advance numbers of the primary sample elements, constituting the limits of confidence intervals when the estimator is based on a single order statistic[3]. The determination of the limits of the quantiles confidence intervals requires much larger calculations (when $n$ is big[4]) if the estimator is based on two order statistics, due to the sorting of possible realizations – in that case, resampling may be justified but is not necessary. The percentile method using all resamples can be called the exact percentile method (by analogy with the exact bootstrap method).

Let $y_j$ be a realization of a bootstrap $p$-quantile estimator $\hat{\xi}^*_{pn}$, for $j = 1, \ldots, o$. If an estimator is based on a single order statistic, $o$ is equal to $k$ (or $n$ if there were no repetitions in the primary sample). Let us mark the bootstrap confidence interval as $I^*_{pn} = \left[y^*_{z_1}, y^*_{z_2}\right]$. For a given confidence level of 1-$\alpha$, the lower limit is:

$$\mathrm{y}^*_{\mathrm{z}_1} = \sup\left\{\mathrm{y_j}: F_{\hat{\xi}^*_{pn}}(\mathrm{y_j}) \le \frac{\alpha}{2}\right\}, \tag{23}$$

where $F_{\hat{\xi}^*_{pn}}$ is the bootstrap quantile estimator distribution. The upper limit is:

$$\mathrm{y}^*_{\mathrm{z}_2} = \inf\left\{\mathrm{y_j}: F_{\hat{\xi}^*_{pn}}(\mathrm{y_j}) \ge 1 - \frac{\alpha}{2}\right\}, \tag{24}$$

[3] It results from the properties of the exact distribution of the bootstrap percentile estimator (the probabilities of individual realizations of the estimator are the same for all samples of a given size). Table 4 lists these numbers for $p = 0.5$ and 1-$\alpha$= 0.95.

[4] Currently, due to the high computing efficiency of computers, computations even for big $n$ are not long-lasting.

If the bootstrap $p$-quantile estimator is based on a single order statistic, the elements of the primary sample are the limits of the confidence intervals. If the estimator is based on a linear combination of order statistics, the linear combinations of these elements may also be the limits.

Due to the discrete nature of distributions of the bootstrap estimator, besides the assumed confidence level 1-$\alpha$, we have a confidence level that can be called actual.

The number of realizations of bootstrap quantile estimators based on single order statistics is much smaller than that based on their linear combination. This has two important consequences. First, estimators based on two order statistics will allow building narrower confidence intervals than those based on one. Secondly, we can suspect that the discrepancy between the assumed and the actual confidence level is smaller for the estimator based on two order statistics than for that based on a single one (Nagaraja and Nagaraja (2020, p. 81)[5] pay attention to this discrepancy).

## 4. Monte Carlo method for quantile estimation

The bias and variance of the estimators, the widths of the confidence intervals, and the coverage probability can be estimated by the Monte Carlo (MC) simulation method. These measures can be used, for example, to compare different estimators. Calculating them requires drawing $R$ random samples, the so-called replication. If the bootstrap estimators are used, a single replication is a single primary sample (which may be resampled).

For sampling, pseudorandom number generators are used, which generate real numbers from a uniform distribution on the interval [0; 1]. Let the drawn number ($\omega_i$) be the value of a CDF of the distribution the sample comes from. Elements of the sample can be designated as $x_i = F^{-1}(\omega_i)$, for $i = 1,\ldots,n$.

Let $\beta$ denote some target quantity of interest, $\hat{\beta}_R$ its MC estimate from simulation experiment with $R$ replications, and $\hat{\beta}_r$ the estimate based on the $r$th replication, $r = 1,\ldots, R$ (Koehler et al. (2009)). The MC estimate of $\beta$ is then:

$$\hat{\beta}_R = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_r. \tag{25}$$

In statistical experiments, the distributions that the samples come from are known. So, it is possible to calculate the MC approximation of the estimator bias relative to the true value of the $p$-quantile. The MC estimate of the bias and variance of some $p$-quantile estimator is:

$$\widehat{bias}_{MC} = \frac{1}{R}\sum_{r=1}^{R}\sum_{j=1}^{o}\left(y_j^r \cdot P\left(\xi_{pn}^r = y_j^r\right)\right) - \xi_p, \tag{26}$$

[5] Nagaraja and Nagaraja point out the discrepancy between the assumed confidence level and the coverage probability (coverage probability will be discussed in Section 4). Since the coverage probability can be regarded as an estimate of the actual level of confidence, both statements are roughly equivalent.

$$\hat{V}_{MC} = \frac{1}{R}\sum_{r=1}^{R}\left(\sum_{j=1}^{o}\left(\left(y_j^r\right)^2 \cdot P\left(\hat{\xi}_{pn}^r = y_j^r\right)\right) - \left(\sum_{j=1}^{o}\left(y_j^r \cdot P\left(\hat{\xi}_{pn}^r = y_j^r\right)\right)\right)^2\right), \tag{27}$$

where $\hat{\xi}_{pn}^r$ is the $p$-quantile estimator in the $r$th replication.

Let the confidence interval determined in the $r$th replication be marked as $I_{pn}^r = \left[y_{z_1}^r, y_{z_2}^r\right]$. The MC estimate of its width and the coverage probability is:

$$\hat{d}_R = \frac{1}{R}\sum_{r=1}^{R}\left(y_{z_2}^r - y_{z_1}^r\right) \tag{28}$$

$$\varphi_R = \frac{\#\{[y_{z_1}^r, y_{z_2}^r] : \varphi_p \in [y_{z_1}^r, y_{z_2}^r]\}}{R}. \tag{29}$$

When the coverage probability is close to the assumed confidence level but not lower than it, the method of determining the confidence intervals properly fulfills its task.

## 5. The median estimation – comparison of estimators

The simulation research using the Monte Carlo method was carried out. Samples come from six distributions: two with right asymmetry (LogNormal(1.0.75), Gamma(2.2)), two with left asymmetry (-LogNormal(1.0.6) + 5, Gamma(1.25, 2.5) + 5) and two symmetrical (N(3.0.5) and N(3.2)). The sample sizes were selected to include both small and large samples.

For different sample sizes $n$, $R$ = 2,000 times $n$ pseudorandom numbers from the interval [0,1] were drawn, which were treated as a CDF value. The same CDF values were used for all distributions and methods, which allows for a better comparability of results. Such selection makes the results for individual cases independent of the quality of the pseudorandom number generator. Based on the $n$ values of the CDF, random samples were determined for six distributions.

The first stage of the simulation studies was to estimate the bias and variance of the three bootstrap median estimators $\hat{\xi}_{0.5n}^{1*}$, $\hat{\xi}_{0.5n}^{2*}$, and $\hat{\xi}_{0.5n}^{3*}$, defined by formulas (13), (14) and (15). The bias and the variance were estimated according to the formulas (26) and (27) by the MC method.

In the second stage, confidence intervals for quantiles were determined using the following methods:

M1 – exact percentile method and estimator $\hat{\xi}_{0.5n}^{1*}$,

M2 – exact percentile method and estimator $\hat{\xi}_{0.5n}^{1*}$,

M3 – exact percentile method and estimator $\hat{\xi}_{0.5n}^{3*}$,

M4 – $BC_a$ method (Efron and Tibshirani, 1993 p. 185) and estimator $\hat{\xi}_{0.5n}^{3*}$,

M5 – the adjacent spacings method (Nagaraja and Nagaraja, 2020 p. 89) (calculations were made for several combinations of parameters $s$ and $t$[6], choosing those for which confidence intervals were narrowest),
M6 – using the limit distribution of the order statistics corresponding to the sample $p$-quantile (formula (8)),
M7 – using the limit distribution of $p$-quantile (formula (7)).

The widths and coverage probabilities were used to compare confidence intervals constructed with different methods, calculated according to (28) and (29). The use of the M7 method requires a comment. The confidence intervals constructed using the M1 to M6 methods were estimated by the MC method based on $R$ replications. The intervals constructed using the M7 method (from the limit distribution of p-quantile) are calculated from formula (7). One may suspect (especially for large samples) that the confidence intervals determined in this way are close to the real ones and may constitute a reference point for the intervals obtained with other methods. Note, however, that the use of formula (7) requires knowledge of $\xi_p$ and $f(\xi_p)$. In fact, we know them very rarely.

In Figure 1, the bias of the bootstrap median estimators is given, depending on the sample size (for $n$ = 10, 15,…, 205). The bias was calculated by the MC method based on $R$ = 2,000 samples from six distributions. To improve the readability of the graph, the data series are presented as continuous lines.

If np is not integer (which in the case of the median corresponds to the odd n), the tree bootstrap median estimators are based on the same order statistics, so they are the same. Estimators differ when n is even. The case n odd was extracted as a separate data series to avoid oscillations when the sample size changes from even to odd. This was made because those oscillations would completely obscure the image (as in Parrish (1990 p. 253)). It is obvious that as the sample size increases, the bias on all estimators usually decreases (if bias jumps are omitted when the sample size changes from even to odd and vice versa). This does not mean, however, that the increase in n in the case of simulation by the MC method is always accompanied by a decrease in bias. The possible increase in bias results from the random selection of $R$ samples.

The estimator $\hat{\xi}_{0.5n}^{3*}$ shows the smallest jumps in the bias with the change in the sample size from odd to even (and vice versa). The data series marked as E123 and E3 for all distributions almost coincide.

When samples came from right asymmetry distributions (for even $n$), the absolute value of the bias of the estimator $\hat{\xi}_{0.5n}^{1*}$ was usually the smallest, while that of the

[6] Five parameter combinations were used: $s$ = 1 and $t$ = 2, $s$ = 2 and $t$ = 1, $s$ = 2 and $t$ = 2, $s$ = 2 and $t$ = 3, $s$ = 3 and $t$ = 2. The narrowest confidence intervals were obtained for the last two variants in all simulation experiments. The combination of $s$ = 3 and $t$ = 2 was best in the case of samples from right asymmetry distributions, while the combination of $s$ = 2 and $t$ = 3 in the case of samples from left asymmetry and symmetrical distributions.

estimator $\hat{\xi}^{2*}_{0.5n}$ the largest. When samples came from left asymmetry distributions, the opposite was true usually – the absolute value of the bias of the estimator $\hat{\xi}^{2*}_{0.5n}$ was the smallest, while that of the estimator $\hat{\xi}^{1*}_{0.5n}$ the largest. When samples came from symmetrical distributions, the absolute value of the bias of the estimator $\hat{\xi}^{3*}_{0.5n}$ was the smallest for almost all *n*, while that of the estimator $\hat{\xi}^{1*}_{0.5n}$ or $\hat{\xi}^{2*}_{0.5n}$ was the largest.

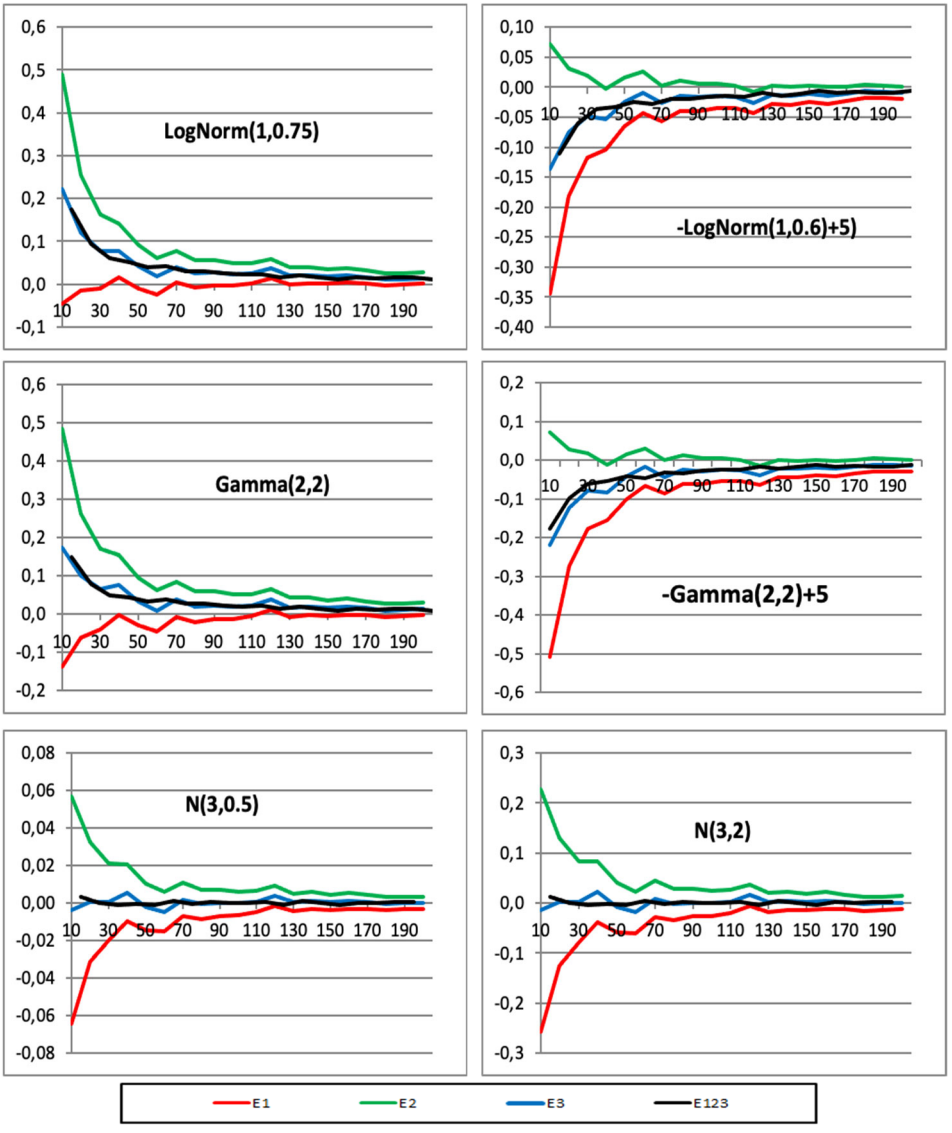


**Figure 1:** The bias of the bootstrap median estimators depending on the sample size (n = 10, 15,…, 205), calculated by the MC method. Note: In the charts, the data series marked E1, E2 and E3 correspond to the estimators $\hat{\xi}^{1*}_{0.5n}$, $\hat{\xi}^{2*}_{0.5n}$, and $\hat{\xi}^{3*}_{0.5n}$ for even n, while E123 corresponds to all estimators for odd n.

When samples came from right asymmetry distributions, the expected value of almost all estimators is greater than the median (except for even $n$ and the estimator $\hat{\xi}^{1*}_{0.5n}$). When samples came from left asymmetry distributions, the expected value of almost all estimators is less than the median (except for even $n$ and the estimator $\hat{\xi}^{2*}_{0.5n}$). When samples came from symmetrical distributions, the bias oscillates around zero – except for even $n$ and the estimator $\hat{\xi}^{1*}_{0.5n}$ (always negative bias) or the estimator $\hat{\xi}^{2*}_{0.5n}$ (always positive bias).

Figure 2 shows the variance of the bootstrap median estimators depending on the sample size. The graphs were prepareded only for small samples (n = 10, 11,…,35). The differences in the case of large samples were very small. When samples came from symmetrical distributions, the variance of the $\hat{\xi}^{3*}_{0.5n}$ estimator was the smallest. When samples came from right asymmetry distributions, the variance of the $\hat{\xi}^{2*}_{0.5n}$ estimator was the biggest. When samples came from left asymmetry distributions, the variance of the $\hat{\xi}^{1*}_{0.5n}$ estimator was the biggest.

Figure 3 presents the width of 0.95 median confidence intervals depending on the sample size. The intervals were calculated by the MC method (the M1-M6 methods) and using the limit distribution (the M7 method). The graphs were made up only for small samples ($n$ = 10, 15,.., 35). For large samples, the widths of confidence intervals constructed with different methods are very similar (except those obtained using the M5 method).

Confidence intervals constructed with the M7 method were usually narrowest, especially for samples from left asymmetry distributions and large samples ($n$ above 115) from symmetrical distributions. Interval widths for M7, M4, and M3 (but only for even $n$) are very similar. If $n$ was even, narrower confidence intervals were usually obtained using the M3 method rather than using the M4 for large samples (n above 180) and samples from left asymmetry distribution.

There are jumps in the widths of confidence intervals constructed with the exact bootstrap estimators (that is for M1, M2, and M3 methods) when $n$ changes from even to odd. This is due to the changing the rank of the order statistic used as an estimator (we do not observe it for other methods). The narrowest confidence intervals were obtained by the M3 method, regardless of the distribution asymmetry type the samples came from. This is because the estimator based on two order statistics has much more realizations than when based on one order statistic only. If the samples came from right asymmetry distribution, narrower confidence intervals were obtained with the M1 method than with the M2. If the samples came from left asymmetry distribution, the effect was opposite. If the samples came from symmetrical distributions, the M1 method gave the narrower intervals for about half of the cases and the M2 for the other half. These conclusions are similar to those obtained for the variance and apply of course only to cases when $n$ is even.

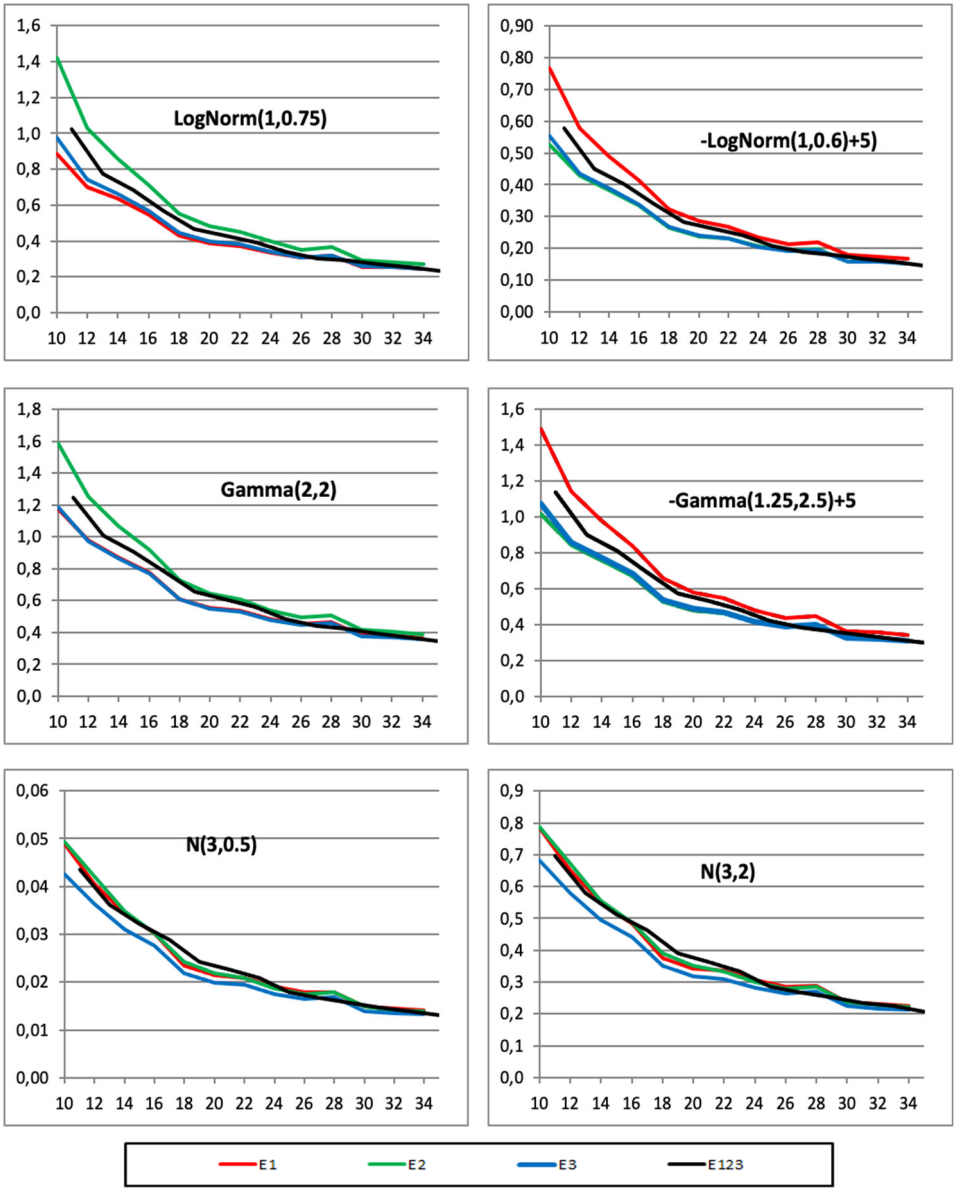


**Figure 2:** The variance of the bootstrap median estimators depending on the sample size (n = 10, 11, …, 35), calculated by the MC method. Note: as for Figure 1.

For almost all sample sizes, the widest confidence intervals were obtained by the M5 method (despite using a combination of parameters giving the best results). The authors of the method (Nagaraja, Nagaraja (2019 p. 75)) point out that although this method gives wider intervals than other methods, it can be used in the case of extreme quantiles even if the sample has only a few observations. The M6 method usually gave

wider confidence intervals than the M7, M4, and M3 methods (for $n$ even), although the differences were small for large samples.

Figure 4 presents 1-$\varphi$ (1-coverage probability) calculated for confidence intervals estimated by M1-M6 methods, depending on the sample size for a 0.95 confidence level. This probability was illustrateed in three variants. The first variant covers all sample sizes, the second only odd sizes, and the third only even sizes. The charts are presented only for LogNorm(1.0.75) distribution. The results for the remaining distributions were very similar. The chart shows strong fluctuations in the coverage probability when the sample size changes, both for all sample sizes and separately for even and odd sizes. De Angelis et al. 1993 p. 526 believed the discrete nature of the bootstrap quantile estimators distributions to be the reason for the fluctuation. It should be noted that fluctuations occur for all methods. Therefore they can only result from the quality of random samples generated (for all methods the same pseudorandom numbers were used). Although relatively many samples were drawn ($R$ = 2,000), the effect of the work of the pseudorandom number generator indicates its imperfection[7]. 4,000, 8,000, and 16,000 samples were drawn several times to check whether increasing the number of drawn samples would reduce the observed fluctuations. It turned out that the differences in the calculated 1-$\varphi$ values in individual experiments were large. This means that a comparison across methods by simulation experiments requires the same conditions. It is advisable to use the same generated pseudorandom numbers in all experiments.

## 6. Conclusions

1. Information about the asymmetry direction of the distribution the sample came from may be a valuable indication when choosing a bootstrap quantile estimator (when $np$ is an integer). The sample skewness coefficient can be used for this purpose. If the estimator is based on a single order statistic and a sample comes from a right asymmetry distribution, the order statistic of the rank $np$ used as an estimator gives a smaller bias and narrower confidence intervals. If a sample comes from a left asymmetry distribution, an order statistic of the rank $np + 1$ is a better estimator. Most asymmetric distributions used in statistical experiments have a right asymmetry distribution. This is why the most common quantiles and sample quantiles are defined as left quantiles. For the distributions with left asymmetry, a right quantile would be more appropriate.

[7] When using simulation methods using random numbers, one should take into account the limited possibilities of pseudorandom number generators. It is worth conducting experiments using various pseudorandom number generators. Examples of such studies were presented by Sulewski 2019.

2. Quantile estimators in the form of single order statistics are very simple to apply (they have the same PDF values for the same sample sizes). To use a linear combination of two order statistics as the estimator requires more effort of calculations. As the research shows, this effort pays off - the estimated confidence intervals are narrower, and the coverage probability is closer to the assumed

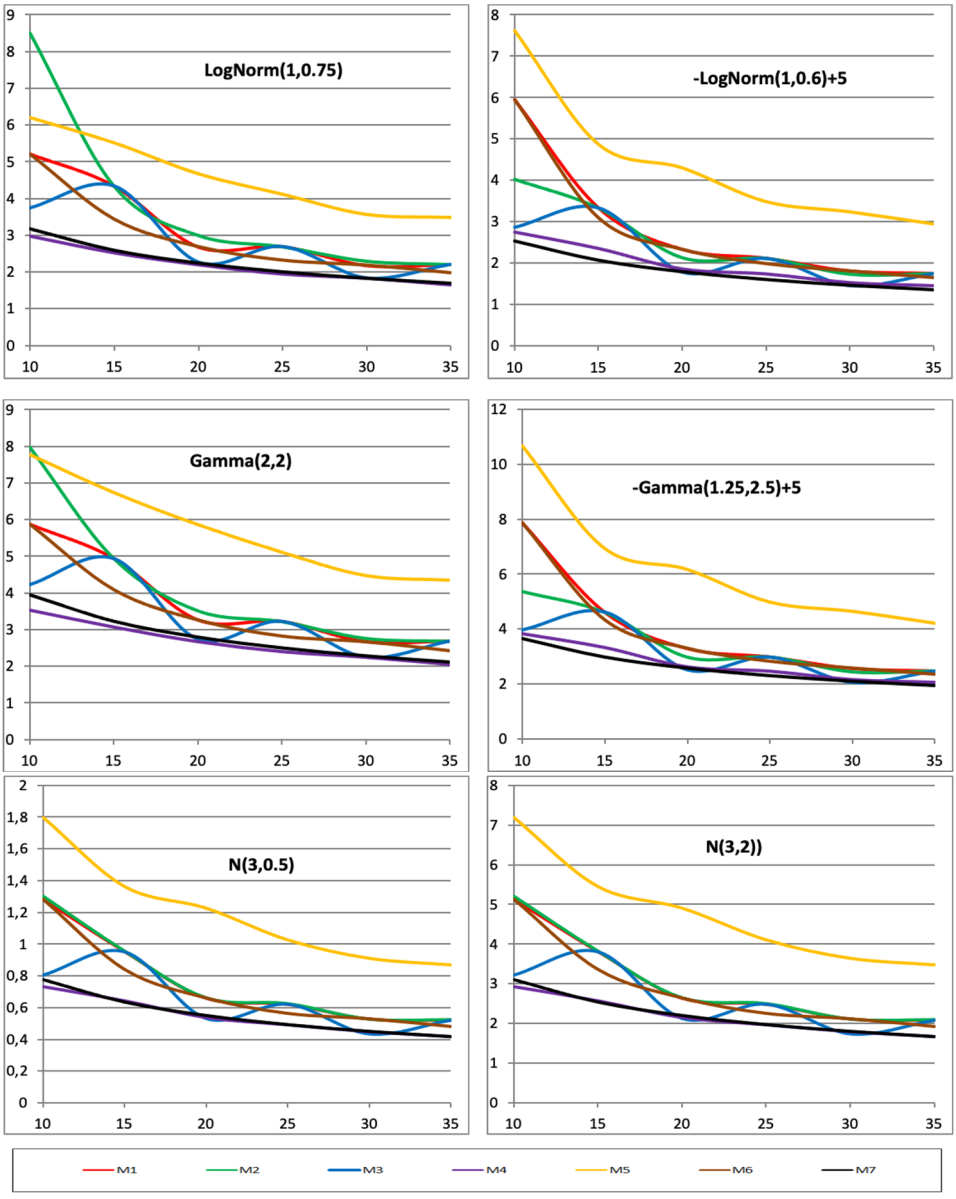


**Figure 3:** The width of median confidence intervals (1-α =0.95) depending on the sample size (n = 10, 15, …, 35).

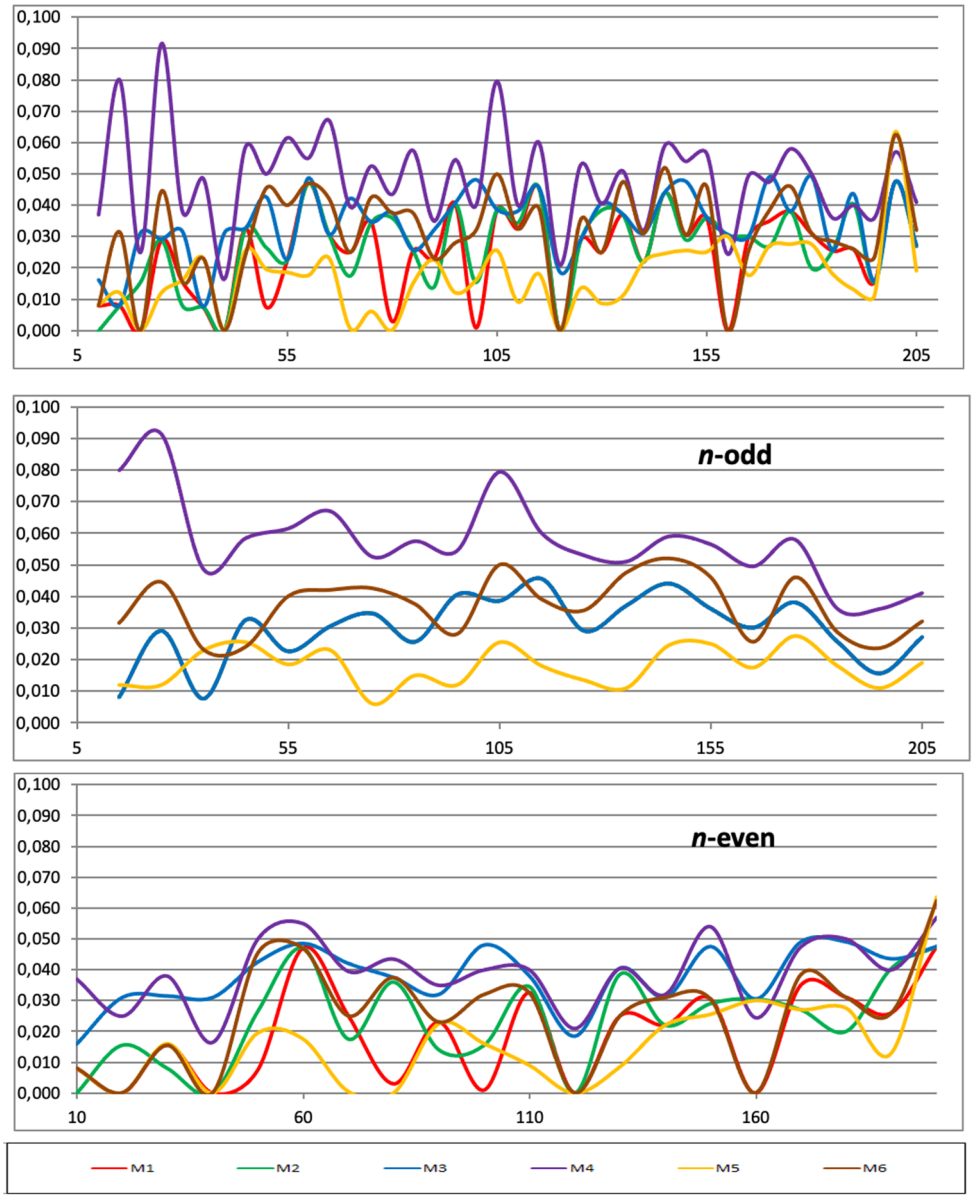


**Figure 4:** 1-φ depending on the sample size (n = 10, 15,…, 205) for a 0.95 confidence level. Samples came from the LogNorm(1.0.75) distribution. Note: The top chart is for all numbers, the middle chart for odd n, and the bottom chart for even n.

confidence level. The possibility to construct narrower confidence intervals results from a bigger number of realizations of the estimator based on two order statistics. When $np$ is not an integer, you may consider using the estimator as a linear combination of the three order statistics of the ranks $[np]$, $[np]$ +1, and $[np]$ +2.

The algorithm for calculating the distribution of such an estimator would be similar to the algorithm given in Section 2. Also, in this case, the probability that the given elements of the primary sample will occur on three positions: $[np]$, $[np]+1$, and $[np]+2$ in the ordered resample, is the same for all samples without repetition of the same size. This makes it possible to construct statistical tables one can use to compute the exact distribution of the estimator for a given primary sample (as it is possible for a combination of two order statistics).

3. There is no need to use the percentile method with resampling for interval estimation of quantiles. The application of the exact percentile method is much simpler. When one uses an estimator based on a single order statistic, it is known in advance which elements of the ordered primary sample constitute the limits of the confidence interval. When one uses an estimator based on two order statistics, the computational effort resulting from sorting all its possible realizations is probably comparable with the time needed to sort its realizations determined from the drawn resamples.[8]\
4. The coverage probability fluctuations (on changes in the sample size) result from the limited capabilities of the pseudorandom number generator. One can conclude so because fluctuations occur for all methods. The conducted experiments indicate that increasing the number of repetitions in Monte Carlo simulations does not reduce the fluctuations. This conclusion was made based on experiments with $R$ = 2,000, 4,000, and 8,000. This means that it is better to use the same samples when you compare different estimation methods.
5. The bias and the variance of the bootstrap median estimators, as well as the width of the median confidence intervals were estimated using the MC method. Fluctuations of these parameters resulted mainly from the change in rank of order statistics used as estimators when the sample size changed from even to odd. This fact was particularly clear for the estimators in the form of a single order statistic. If even and odd samples are considered separately, there are no fluctuations. There are also no fluctuations in the width of the confidence intervals estimated with the other methods. This is because the discussed measures are calculated as average from all replications. The coverage probability is determined for all $R$ repetitions.

The research conducted and presented in the article is based on a limited set of distributions. However, one can assume that conclusions can be generalized for their wider collection. Only one pseudorandom number generator was used – an Excel generator. The results showed that it is worth researching various pseudorandom numbers generators and examining their impact on the quality of the Monte Carlo simulations.

[8] For a sample with 50 elements, the number of realizations of the estimator based on two order statistics is maximally equal to $50 + 49 \cdot 25 = 1275$.

## References

Altman, E. I., (1968). Financial Ratios, Discriminant Analysis and the Prediction of the Corporate Bankruptcy. *The Journal of Finance*, vol. 23, pp. 589–609.

Chen, J. H., Williams, M., (1999). The determinants of business failures in the US low-technology and high-technology industries. *Applied Economics*, vol. 31, pp. 1551–1563.

European Commission, (2003). BEST project on restructuring, bankruptcy and a fresh start: Final report of the expert, Enterprise Directorate-General.

Bahadur, R. R., (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, vol. 37(3), pp. 577–580.

Bickel, P. J., Freedman, D. A., (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, vol. 9(9), pp. 1196–1217.

David, H. A., Nagaraja, H. N., (2003). Order Statistics, Wiley & Sons, Inc.

De Angelis, D., Hall, P., Young, G. A., (1993). A note on coverage error of bootstrap confidence intervals for quantiles. *Math. Proc. Camb. Phil. Soc.*, vol. 114(3), pp. 517–531.

Efron, B., (1979). Bootstrap Methods: Another look at the jackknife. *The Annals of Statistics*, vol. 7, no. 1, 1–26.

Efron, B., (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.*, vol. 82(39), pp. 171–185.

Efron, B., Tibshirani, R. J., (1993). An introduction to the Bootstrap, New York: Chapman & Hall.

Evans, D. L., Leemis, L. M., Drew, J. H., (2006). The distribution of order statistics for discrete random variables with applications of bootstrapping. *Journal on Computing*, vol. 18(1), pp. 19–30.

Falk, M., Kaufmann, E., (1991). Coverage probabilities of bootstrap-confidence intervals for quantiles. *Ann. Statist.*, vol. 19(1), pp. 485–495.

Falk, M., Reiss, R. D., (1989). Weak convergence of smoothed and nonsmoothed bootstrap quantile estimates. *Ann. Probab.*, vol. 17(1), pp. 362–371.

Fisher, N. I., Hall, P., (1991). Bootstrap algorithms for small samples. *Journal of Statistical Planning and Inference*, vol. 27, pp. 157–169.

Hettmansperger, T. P., Sheather, S. J., (1986). Confidence intervals based on interpolated order statistics. *Statist. Probab. Lett.*, vol. 4(2), pp. 75–79.

Hutson, A. D., (2002). A semi-parametric quantile function estimator for use in bootstrap estimation procedures. *Stat. Comput.*, vol. 2(4), pp. 331–338.

Hutson, A. D., Ernst, M. D., (2000). The exact bootstrap mean and variance of an L-estimator. *Journal of the Royal Statistical Society: Series B*, vol. 62(1), pp. 89–94.

Hyndman, R. J., Fan, Y., (1996). Sample quantiles in statistical packages. *The American Statistician*, vol. 50(4), pp. 361–365.

Koehler, E., Brown, E., Haneuse, J.-P.A., (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *Amer. Statist.*, vol. 63(2), pp. 155–162.

Kisielińska, J., (2013). The exact bootstrap method shown on the example of the mean and variance estimation. *Computational Statistics*, vol. 28(3), pp. 1061–1077.

Maritz, J. S., Jarrett, R. G., (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, vol. 73(361), pp. 194–196.

Nagaraja, C. H, Nagaraja, H. N., (2020). Distribution-free approximate methods for constructing confidence intervals for quantiles. *International Statistical Review*, vol. 88(1), pp. 75–100.

Nyblom, J., (1992). Note on interpolated order statistics. *Statist. Probab. Lett.*, vol. 14(2), pp. 129–131.

Parrish, R. S., (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, vol. 46, pp. 247–257.

Pekasiewicz, D., (2015). Order statistics in estimation procedures and their applications in economic research, University of Lodz, Lodz.

Serfling, R. J., (1980). Approximation theorems of mathematical statistics, John Wiley & Sons, New York, Chichester, Brisbane, Toronto, Singapore.

Sulewski, P., (2019). Comparison of normal random number generators, *Wiadomości Statystyczne. The Polish Statistician*, vol. 64, pp. 5–31.

Singh, K., (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.*, vol. 9(6), pp. 1187–1195.

Wilcox, R. R., (2001). Fundamentals of modern statistical methods, Springer, New York.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 167–177,* https://doi.org/10.59170/stattrans-2024-010
Received – 04.10.2022; accepted – 08.12.2023

# Reliability for Zeghdoudi distribution with an outlier, fuzzy reliability and application

**Thara Belhamra**[1], **Halim Zeghdoudi**[2], **Vinoth Raman**[3]

## Abstract

This study focuses on estimating reliability *P[Y<X]*, where *Y* has a Zeghdoudi distribution with parameter *a*, *X* has a Zeghdoudi distribution with one outlier present and parameter *c*, and the remaining $(n - 1)$ random variables are from a Zeghdoudi distribution with parameter *b*, in order for *X* and *Y* to be independent. Several findings of a simulation study and the maximum likelihood estimate of *R* are provided. We also present some results related to fuzzy dependability. Finally, using actual data on survival durations (in days) of 72 Algerians infected with a coronavirus, we demonstrate how the Zeghdoudi distribution may be applied to other distributions in order to demonstrate its adaptability.

**Key words:** Zeghdoudi distribution, maximum likelihood estimator, Newton-Raphson method, outlier, fuzzy reliability

## 1. Introduction

Inferences regarding $R = P(Y < X)$, when *X* and *Y* are independently distributed, are of relevance in the reliability context and play a significant role in many practical domains, including engineering, medicine, and quality control. In the statistical literature, *R* estimation is a highly common practice. It calculates the likelihood that a component's stress *Y* will be greater than its random strength *X*. Additionally, R offers the likelihood that a system would malfunction if the applied stress exceeds its capacity.

The earliest research on this issue dates back to Birnbaum (1956) and Birnbaum and McCarty (1958). Kapur and Lamberson have also discussed the reliability under stress (1977). *R* was estimated for the negative binomial distribution by Sathe and Dixit

[1] Department of Mathematics, Science Faculty, Badji Mokhtar-Annaba University, 2300, Annaba, Algeria. E-mail: thara.belhamra@yahoo.fr. ORCID: https://orcid.org/0009-0008-4771-279X.

[2] Department of Mathematics, Science Faculty, Badji Mokhtar-Annaba University, 2300, Annaba, Algeria. E-mail: halim.zeghdoudi@univ-annaba.dz. ORCID: https://orcid.org/0000-0002-4759-5529.

[3] Imam Abdulrahman Bin Faisal University Dammam, Eastern, Saudi Arabia. E-mail: vrrangan@iau.edu.sa. ORCID: https://orcid.org/0000-0002-3815-2312.

(2001). In the presence of outliers produced by uniform distributions, Dixit and Nasiri (2001) provided an estimation of the parameters of the exponential distribution. Then, in the exponential and gamma cases, respectively, Baklizi and Dayyeh (2003) performed a shrinkage estimation of *R*, Kundu and Gupta (2005) considered an estimation of $P[Y < X]$ for a generalized exponential distribution, and Deiri (2011) performed an estimation of *R* with the presence of two outliers. Jafari (2011) obtained the moment, maximum likelihood, and mixture estimators of *R* in the Rayleigh distribution in the presence of one outlier, and Jabbari (2013) discussed the estimation of *R* in the Lindley distribution with an Outlier. Deiri (2010) considered the estimation of reliability for the exponential case in the presence of one outlier. And most recently, Lindley distributions have been used to draw conclusions on stress-strength reliability by Mutairi, Guitani, and Kundu (2013).

With one outlier generated from the same distribution, we obtain the maximum likelihood estimate of *R* for the Zeghdoudi distribution in this study. The Zeghdoudi distribution with parameter a, probability density function (pdf) is given by

$$f(y,a) = \frac{a^3}{2+a} y(1+y)e^{-ay} \quad a > 0$$

In this study, it is assumed that the random variables $(Y_1, Y_2, \dots, Y_m)$ have a Zeghdoudi distribution with parameter a, while the random variables $(X_1, X_2, \dots, X_n)$ are such that one of them comes from a Zeghdoudi distribution with parameter $c$ and the remaining $(n-1)$ random variables come from a Zeghdoudi distribution with parameter $b$.

The body of the article is structured as follows. The joint distribution of $(X_1, X_2, \dots, X_n)$ in the presence of one outlier is obtained in Section 2. The MLE of *R* and the method of maximum likelihood estimators of parameters are discussed in Sections 3 and 4, respectively. The simulation studies are provided in Section 5. Section 6 discusses illustrative combinations of the Zeghdoudi distribution with other distributions to demonstrate the adaptability of this distribution. Finally, Section 7 is the conclusion of the study.

## 2. Joint distribution of $(X_1, X_2, , X_n)$ with an outlier

Consider that $(X_1, X_2, \dots, X_n)$ are distributed with p.d.f $g(x,c)$ as $Zeghdoudi\ (c)$ and remaining $(n-1)$ of them are distributed with p.d.f $f(x,b)$ as $Zeghdoudi\ (b)$. The joint distribution of $(X_1, X_2, \dots, X_n)$ can be expressed as

$$f(x_1, x_2, \dots, x_n; b, c) = \frac{(n-1)!}{n!} \prod_{i=1}^{n} f(x,b) \sum_{i=1}^{n} \frac{g(x_i;c)}{f(x_i;b)}$$

$$= \frac{(n-1)!}{n!}\prod_{i=1}^{n}\frac{b^3}{2+b}x(1+x)e^{-bx}\sum_{i=1}^{n}\left[\frac{\left(\frac{c^3}{2+c}x(1+x)e^{-cx}\right)}{\left(\frac{b^3}{2+b}x(1+x)e^{-bx}\right)}\right]$$

$$= \frac{(n-1)!}{n!}\frac{b^{3n}}{(2+b)^n}\prod_{i=1}^{n}x_i(1+x_i)e^{-b\sum_{i=1}^{n}x_i}\sum_{i=1}^{n}\left[\frac{\left(\frac{c^3}{2+c}x_i(1+x_i)e^{-cx_i}\right)}{\left(\frac{b^3}{2+b}x_i(1+x_i)e^{-bx_i}\right)}\right]$$

$$= \frac{(n-1)!}{n!}\frac{b^{3n-3}}{(2+b)^{n-1}}\frac{c^3}{2+c}\prod_{i=1}^{n}x_i(1+x_i)e^{-b\sum_{i=1}^{n}x_i}\sum_{i=1}^{n}x_i(1+x_i)e^{x_i(b-c)} \quad (1)$$

The marginal distribution of $X$ is

$$f(x;b,c) = \frac{1}{n}\frac{c^3}{2+c}x_i(1+x_i)e^{-cx} + \frac{n-1}{n}\frac{b^3}{2+b}x(1+x)e^{-bx} \quad (2)$$

Using (2) to obtain $R = (Y < X)$.

## 3. Maximum likelihood estimators of parameters

Let $(Y_1, Y_2, \dots, Y_m)$ be a random sample for $Y$ with pdf,

$$f(y;a) = \frac{a^3}{2+a}x(1+x)e^{-ay} \qquad x, a > 0$$

The log likelihood function is given by

$$L(a) = 3mlna(a+2) + \sum_{i=1}^{m}\ln(y_i + {y_i}^2) - a\sum_{i=1}^{m}y_i$$

The $MLE$ of $a$ is obtained by taking the derivative with regard to $a$ and equating it to 0.

$$\hat{a} = \frac{1}{y}\left(-\bar{y} + \sqrt{4\bar{y} + \bar{y}^2 + 1} + 1\right) \quad (3)$$

Now, consider $X_1, X_2, \dots, X_n$ as a random sample for $X$ with one outlier present and pdf,

$$f(x;b,c) = \frac{1}{n}\frac{c^3}{2+c}x_i(1+x_i)e^{-cx} + \frac{n-1}{n}\frac{b^3}{2+b}x(1+x)e^{-bx}$$

From (1), the log likelihood function is given by

$$L(b,c) = \ln\left(\frac{(n+1)!}{n!}\right) + (3n+3)lnb - (n-1)\ln(2+b) + 3lnc$$

$$-\ln(2+c) + \sum_{i=1}^{n}\ln\left(x_i(1+x_i)\right) - b\sum_{i=1}^{n}x_i + \ln\sum_{i=1}^{n}e^{x_i(b-c)}$$

We derive the normal equations by taking the derivative with regard to $b$ and $c$ and equating the results to 0.

$$\frac{\delta L(b,c)}{\delta b} = \frac{3n-3}{b} - \frac{n-1}{2+b} - \sum_{i=1}^{n} x_i + \frac{\sum_{i=1}^{n} x_i e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}} \tag{4}$$

$$\frac{\delta L(b,c)}{\delta c} = \frac{3}{c} - \frac{1}{1+c} - \frac{\sum_{i=1}^{n} x_i e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}} \tag{5}$$

This system of equations lacks a closed-form solution, so the authors use the Newton-Raphson method to iteratively find the values of $\hat{b}$and $\hat{c}$. In this instance, we will iteratively estimate $\hat{\beta} = (\hat{b}, \hat{c})$.

$$\hat{\beta}_{i+1} = \hat{\beta}_i - K^{-1}k \tag{6}$$

where $k$ is the vector of normal equations for which we want

$$k = [k_1, k_2]$$

with

$$k_1 = \frac{3n-3}{b} - \frac{n-1}{2+b} - \sum_{i=1}^{n} x_i + \frac{\sum_{i=1}^{n} x_i e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}}$$

$$k_2 = \frac{3}{c} - \frac{1}{1+c} - \frac{\sum_{i=1}^{n} x_i e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}}$$

and $K$ is the matrix of second derivatives

$$K = \begin{bmatrix} \frac{dk_1}{db} & \frac{dk_1}{dc} \\ \frac{dk_2}{db} & \frac{dk_2}{dc} \end{bmatrix}$$

where

$$\frac{dk_1}{db} = \frac{3-3n}{b^2} + \frac{n-1}{(1+b)^2} + \frac{\sum_{i=1}^{n} {x_i}^2 e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}} - \left(\frac{\sum_{i=1}^{n} {x_i}^2 e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}}\right)^2$$

$$\frac{dk_2}{db} = -\frac{\sum_{i=1}^{n} {x_i}^2 e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}} + \left(\frac{\sum_{i=1}^{n} {x_i}^2 e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}}\right)^2$$

$$\frac{dk_2}{dc} = -\frac{3}{c^2} + \frac{1}{(1+c)^2} + \frac{\sum_{i=1}^{n} {x_i}^2 e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}} - \left(\frac{\sum_{i=1}^{n} {x_i}^2 e^{x_i(b-c)}}{\sum_{i=1}^{n} e^{x_i(b-c)}}\right)^2$$

As our estimate of $b$ and $c$ fluctuate by less than a permitted amount with each subsequent iteration, the Newton-Raphson method converges to $\hat{b}$ and $\hat{c}$.

## 4. Fuzzy reliability of Zeghdoudi distribution

Let T denote the time until a system fails, which is a continuous random variable (component). The fuzzy probability in formula can then be used to compute the fuzzy dependability.

$$R_F(t) = P(T > t) = \int_t^{\infty} \mu(x) f\ (x)dx, \quad 0 \le t \le x < \infty,$$

where $\mu(x)$ is a membership function that expresses how much a given universe's elements belong to a fuzzy set. Assume now that $\mu(x)$ is

$$\mu(x) = \begin{cases} 0, & x \le t_1 \\ \dfrac{x - t_1}{t_0 - t_1}, & t_1 < x < t_2,\ t_1 \ge 0 \\ 1, & x \ge t_2 \end{cases}$$

For $\mu(x)$, by the computational analysis of the function of fuzzy numbers, the lifetime $x(\gamma)$ can be obtained corresponds to a certain value of $\gamma - Cut, \gamma \in [0,1]$, can be obtained by $\mu(x) = \gamma \to \frac{x-t_1}{t_0-t_1} = \gamma$, then

$$\begin{cases} x(\gamma) \le t_1, & \gamma = 0 \\ x(\gamma) = t_1 + \gamma(t_2 - t_1), & 0 < \gamma < 1 \\ x(\gamma) \ge t_2, & \gamma = 1 \end{cases}$$

As a result, it is possible to determine the fuzzy reliability values for all $\gamma$ values. The fuzzy reliability definition establishes the Zeghdoudi distribution's fuzzy dependability. The Zeghdoudi distribution's fuzziness dependability is defined as

$$R_F(t) = \left(\frac{x^2a^2 + a(2+a)x + a + 2}{a+2}\right)e^{-ax} - \left(\frac{x(\gamma)^2a^2 + a(2+a)x(\gamma) + a + 2}{a+2}\right)e^{-ax(\gamma)}$$

Then $R_F(t) = 0$.

### 4.1. Numerical values of fuzzy reliability

We compared traditional reliability and fuzzy reliability in this subsection, where traditional reliability is a survival function as $R(x) = \left(\frac{x^2a^2+a(2+a)x+a+2}{a+2}\right)e^{-ax}$

The comparison was discussed in Table 1. Based on findings, the following observations are made:

- when $\gamma - Cut$ is increased, the Fuzzy reliability increases.
- when $t_2$ interval of membership function is increased, the Fuzzy reliability increases.
- when $t_1$ is decreased, the fuzzy reliability increases, and vice versa.
- the traditional reliability with $t_2$ is lower than the traditional reliability with $t_1$.

A sequence of drawings from the Zeghdoudi distribution is produced by the fuzzy estimating procedure.

**Algorithm**: fuzzy estimation algorithm

- **Input**: initial values of $a$, interval time $(t_1, t_2)$ and $\gamma$ where $0 < \gamma < 1$.
- **Calculate**: $x(\gamma) = t_1 + \gamma(t_2 - t_1)$.
- **For** each method do

**Set**: i=1.

**Estimate parameter** as $\hat{a}$.

**Calculate**

$$\hat{R}_F(t) = \left(\frac{{t_1}^2 a^2 + a(2+a)t_1 + a + 2}{a+2}\right) e^{-at_1} - \left(\frac{x(\gamma)^2 a^2 + a(2+a)x(\gamma) + a + 2}{a+2}\right) e^{-ax(\gamma)}$$

- **End**

**Table 1**: Fuzzy reliability with different values of $a, t_1, t_2, \gamma$.

| $a$ | $t_1$ | $t_2$ | $R(t_1)$ | $R(t_2)$ | $R_F$ | | |
|---|---|---|---|---|---|---|---|
| | | | | | 0.25 | 0.5 | 0.9 |
| 0.2 | 0.01 | 1 | 1 | 0.99736 | 0.00013 | 0.00057 | 0.00208 |
| 0.5 | 0.5 | 2 | 0.99297 | 0.88291 | 0.01542 | 0.03953 | 0.09393 |
| 1 | 0.1 | 3 | 0.99834 | 0.34851 | 0.58082 | 0.88103 | 0.96981 |
| 3 | 0.2 | 1 | 0.91761 | 0.28876 | 0.16824 | 0.34766 | 0.58321 |
| 5 | 0.1 | 1.5 | 0.93146 | 0.00914 | 0.51268 | 0.79802 | 0.91541 |

### 4.2. The maximum likelihood estimator of $R$

Let $Y \sim Zeghdoudi(a)$ with pdf $h(y;a)$ and $X$ be distributed with pdf $f(x;b,c)$ given in (2). The parameter $R$ is estimated as

$$R = P(Y < X) = \int_0^\infty \int_0^x h(y;a) f(x;b,c) dy dx$$
$$= \frac{1}{n} \int_0^\infty \int_0^x \frac{a^3}{2+a} x(1+x) e^{-ay} \frac{c^3}{2+c} x(1+x) e^{-cx} dy dx$$
$$+ \frac{n-1}{n} \int_0^\infty \int_0^x \frac{a^3}{2+a} x(1+x) e^{-ay} \frac{b^3}{2+b} x(1+x) e^{-bx} dy dx$$

$$= \frac{1}{n}\left[1 - \frac{c^3}{(a+2)(c+2)}\left(\frac{a+2}{(a+c)^2} + \frac{(2a^2+6a+4)}{(a+c)^3} + \frac{12a(a+2)}{(a+c)^4} + \frac{24a^2}{(a+c)^5}\right)\right]$$
$$+ \frac{n-1}{n}\left[1 - \frac{b^3}{(a+2)(b+2)}\left(\frac{a+2}{(a+b)^2} + \frac{(2a^2+6a+4)}{(a+b)^3} + \frac{12a(a+2)}{(a+b)^4} + \frac{24a^2}{(a+b)^5}\right)\right] \tag{7}$$

Thus, by invariant property for MLEs, the MLE of $R$ is

$$\hat{R} = \frac{1}{n}\left[1 - \frac{\hat{c}^3}{(\hat{a}+2)(\hat{c}+2)}\left(\frac{\hat{a}+2}{(\hat{a}+\hat{c})^2} + \frac{(2\hat{a}^2+6\hat{a}+4)}{(\hat{a}+\hat{c})^3} + \frac{12\hat{a}(\hat{a}+2)}{(\hat{a}+\hat{c})^4} + \frac{24\hat{a}^2}{(\hat{a}+\hat{c})^5}\right)\right]$$
$$+ \frac{n-1}{n}\left[1 - \frac{\hat{b}^3}{(\hat{a}+2)(\hat{b}+2)}\left(\frac{\hat{a}+2}{(\hat{a}+\hat{b})^2} + \frac{(2\hat{a}^2+6\hat{a}+4)}{(\hat{a}+\hat{b})^3} + \frac{12\hat{a}(\hat{a}+2)}{(\hat{a}+\hat{b})^4} + \frac{24\hat{a}^2}{(\hat{a}+\hat{b})^5}\right)\right]$$

Where $\hat{a}, \hat{b}$ and $\hat{c}$ can be obtained from (3) and (6).

## 5. Simulation study

In this section, using the accept-reject approach and Maple software, we generate random numbers from the Zeghdoudi distribution (both with and without an outlier). We obtain the maximum likelihood estimators of the parameters $a, b$ and $c$ using these samples and the Newton-Raphson technique. The MLE of R is then calculated using these parameters. The values of biases and MSEs of these estimates are presented in Table 2, for $a = 1$, $b = 2$ and $c = 1.3,\ 1.4,\ 1.5,\ 1.6,\ 1.9,\ 2.0,\ 2.1,\ 2.6,\ 2.7,\ 2.9,\ 3.1,\ 3.5,\ 6.0$. All the results are based on 100 replications.

**Table 2:** Biases and (MSE)s of the MLEs of *R*, for *a=1*, *b=2*, and different values of *c*

| $n \longrightarrow$ $c \downarrow$ | n=10 | n=20 | n=30 | n=50 | n=60 | n= 80 | n=90 |
|---|---|---|---|---|---|---|---|
| 1.3 | 0.0013611 | 0.0012772 | 0.0012493 | 0.0012269 | 0.0012213 | 0.0012143 | 0.001212 |
| 1.4 | 0.0013274 | 0.0012604 | 0.0012380 | 0.0012202 | 0.0012157 | 0.0012101 | 0.001208 |
| 1.5 | 0.0012977 | 0.0012455 | 0.0012281 | 0.0012142 | 0.0012107 | 0.0012064 | 0.001205 |
| 1.6 | 0.0012714 | 0.0012324 | 0.0012194 | 0.0012090 | 0.0012064 | 0.0012031 | 0.001202 |
| 1.9 | 0.0012095 | 0.0012014 | 0.0011987 | 0.0011966 | 0.0011961 | 0.0011954 | 0.0011952 |
| 2.0 | 0.0011934 | 0.0011934 | 0.0011934 | 0.0011934 | 0.0011934 | 0.0011934 | 0.0011934 |
| 2.1 | 0.0011791 | 0.0011862 | 0.0011886 | 0.0011905 | 0.0011910 | 0.0011916 | 0.0011918 |
| 2.6 | 0.0011281 | 0.0011607 | 0.0011716 | 0.0011803 | 0.0011825 | 0.0011852 | 0.0011861 |
| 2.7 | 0.0011209 | 0.0011572 | 0.0011692 | 0.0011789 | 0.0011813 | 0.0011843 | 0.0011853 |
| 2.9 | 0.0011088 | 0.0011511 | 0.0011652 | 0.0011765 | 0.0011793 | 0.0011828 | 0.0011840 |
| 3.1 | 0.0010991 | 0.0011462 | 0.0011619 | 0.0011745 | 0.0011777 | 0.0011816 | 0.0011829 |
| 3.5 | 0.0010850 | 0.0011392 | 0.0011572 | 0.0011717 | 0.0011753 | 0.0011798 | 0.0011813 |
| 6 | 0.0010638 | 0.0011269 | 0.0011490 | 0.0011668 | 0.0011712 | 0.0011712 | 0.0011786 |

## 6. Illustrative application

To demonstrate the adaptability of the Zeghdoudi distribution, we offer an example application of it with other distributions in this section. Therefore, we examine the Lindley, exponential, and Zeghdoudi distributions using real data on the survival times (in days) of 72 Algerians who had contracted a coronavirus (https://www.who.int/fr/news/item), Table 3.

**Table 3:** Comparison between Lindley, exponential and Zeghdoudi distributions

| Survival time $m = 3.2$ | Obsfreq | Lindley $\widehat{\theta} = 0.50$ | Exp $\widehat{\theta} = 0.30$ | Zeghdoudi $\widehat{\theta} = 0.6$ |
|---|---|---|---|---|
| [0, 2] | **34** | **27.30** | **33.90** | **30.05** |
| [2, 4] | **17** | **22.10** | **20.50** | **19.81** |
| [4, 6] | **11** | **12.15** | **7.43** | **10.05** |
| [6, 8] | **7** | **7.28** | **6.67** | **7.02** |
| [8, 10] | **3** | **3.17** | **3.50** | **3.07** |
| *Total* | **92** | **92** | **92** | **92** |
| $\chi^2$ | | 2.946 | **2.400** | **1.0095** |

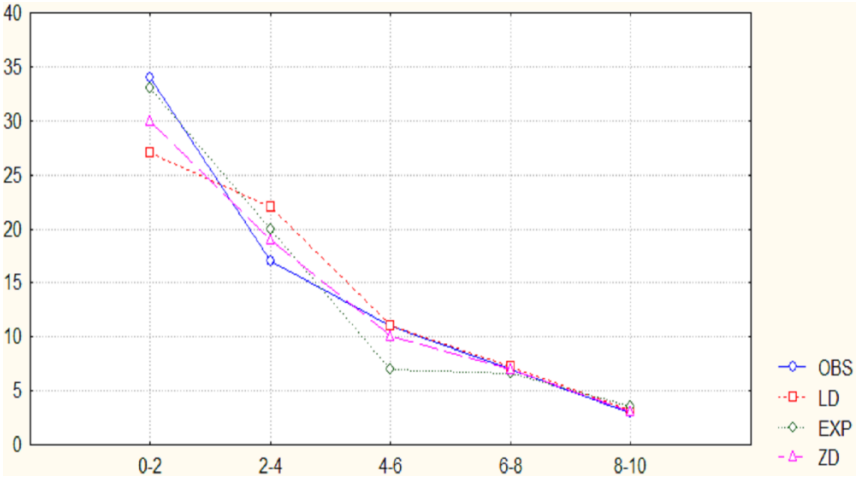


**Figure 1:** Comparison between distributions

As shown in Table 2 and Figure 1, the Zeghdoudi distribution offers the smallest $\chi^2$ value in comparison to the Lindley and exponential distributions, and as a result, best fits the data of all the distributions taken into consideration.

## 7. Conclusion

The challenge of estimating $P(Y < X)$ for the Zeghdoudi distribution in the presence of one outlier has been addressed in this study. Studies have been done on the maximum likelihood estimator for R and fuzzy dependability.

Table 1 contains all of the results, which were based on 100 replications. According to the simulation's findings, biases and MSEs frequently hover around zero when parameters b and c are close to one another, and they rise when the difference between b and c approaches one.

In order to demonstrate the adaptability of the Zeghdoudi distribution, the authors suggested an exemplary application using real data on the survival times (in days) of 72 Algerian people who were infected with coronaviruses, and then compared the outcomes with those of other distributions.

## Acknowledgment

The authors are grateful for the comments and suggestions by the referee and the Editor. Their comments and suggestions greatly improved the article.

## References

Bouhadjar M, Ahmed M. Gemeay, Ehab M. Almetwally, Zeghdoudi H, Etaf Alshawarbeh, Alanazi Talal Abdulrahman, M. M. Abd El-Raouf, and Eslam Hussam, (2022). The Power XLindley Distribution: Statistical Inference, Fuzzy Reliability, and COVID-19 Application. *Hindawi Journal of Function Spaces*, Vol. 2022, Article ID 9094078, 21 pages https://doi.org/10.1155/2022/9094078.

Deiri, E, (2010). Estimation of Reliability For Exponential Case In The Presence Of One Outlier. *Financial Mathematics and Applications,* Vol. 1, pp. 2217–7795.

Deiri, E., (2011). Estimation of *P(Y<X)* for Exponential Distribution in the Presence of Two Outliers. *International Journal of Academic Research,* Vol. 3, pp. 508–514.

Deiri, E., (2011). Estimation of $P(Y < X)$ for Generalized Exponential Distribution in the Presence of Two Outliers when Scale Parameters is Known. *International Journal of Academic Research*, Vol. 3, pp. 1179–1185.

Deiri, E., (2011). Estimation of Parameters of the Gamma Distribution in the Presence of Two Outliers. *International Journal of Academic Research*, Vol. 3, pp. 846–852.

Dixit, U. J., M. Jabbari Nooghabi, M., (2011). Estimation of Parameters of Gamma Distribution in The Presence Of Outliers In Right Censored Samples. *Aligarh Journal of Statistics*, Vol. 31, pp. 17–29.

Dixit, U. J., (1989). Estimation of Parameters of the Gamma Distribution in the Presence of Outliers. *Communications in Statistics − Theory and Methods,* Vol. 18, pp. 3071–3085.

Dixit, U. J., Moor, K. L., and Barnett, V., (1996). On the Estimation of the Power of the Scale Parameter of the Exponential Distribution in the Presence of Outlier Generated from Uniform Distribution. *Metron*, Vol. 54, pp. 201–211.

Dixit, U. J., Nasiri, P. F., (2001). Estimation of Parameters of the Exponential Distribution with Presence of Outliers Generated from Uniform Distribution. *Metron*, Vol. 49(3–4), pp. 187–198.

Eslam Hossam, Alanazi Talal Abdulrahman, Ahmed M. Gemeay, Nawaf Alshammari, Etaf Alshawarbeh, Nour Khaled Mashaqbah, (2022). A novel extension of Gumbel distribution: Statistical inference with Covid-19 application. *Alexandria Engineering Journal*, Vol. 61, Issue 11, November 2022, pp. 8823–8842.

Fathy H. Riad, Bader Alruwaili, Ahmed M. Gemeay, Eslam Hussam, (2022). Statistical modeling for COVID-19 virus spread in Kingdom of Saudi Arabia and

Netherlands. *Alexandria Engineering Journal*. Vol. 61, Issue 12, December 2022, pp. 9849–9866.

Jabbari Khamnei, H., Abolhasani, A., and Fathipour, P., (2012). Reliability for Six Parameter Generalized Burr XII Distribution with Transformation Method. *Accepted in the Journal of Advances and Applications in Statistics*.

Kapur, K. C., Lamberson, L. R., (1977). Reliability in engineering design. John Wiley and sons, New York.

Muqrin A. Almuqrin, Ahmed M. Gemeay, M. M. Abd El-Raouf, Mutua Kilai, Ramy Aldallal, and Eslam Hussam, (2022). A Flexible Extension of Reduced Kies Distribution: Properties, Inference, and Applications in Biology. *Hindawi journal*, ID 6078567, 19 pages https://doi.org/10.1155/2022/6078567.

Nasiri, P. F., Pazira, H., (2010). Bayesian and Non-Bayesian Estimations on the Generalized Exponential Distribution in the Presence of Outliers. *Journal of Statistical Theory and Practice*, Vol. 4(3), pp. 453–475.

Pak, A., Raqab M. Z., Mahmoudi, M. R., Band, S., Mosavi. A, (2022). Estimation of stress-strength reliability $R = P(X > Y)$ based on Weibull record data in the presence of inter-record times. *Alexandria Engineering Journal*, Vol. 61, pp. 2130–2144.

Ramy Aldallal, Ahmed M. Gemeay, Eslam Hussam, Mutua Kilai, (2022). Statistical modeling for COVID 19 infected patient's data in Kingdom of Saudi Arabia, https://doi.org/10.1371/journal.pone.0276688.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 179–190,* https://doi.org/10.59170/stattrans-2024-011
Received – 06.05.2022; accepted – 01.07.2023

# Composite estimators for domain estimation and sensitivity performance interval of their weights

**Piyush Kant Rai**[1], **Sweta Singh**[2]

## Abstract

Some composite estimators based on various combinations of two different existing estimators are obtained for domain estimation. The estimation of weights and thus obtaining optimum weights to combine two or more different existing direct and indirect estimators to form composite estimators are not an easy task for practitioners due to many reasons. To account for the absence of optimum weights, we obtained the sensitivity performance intervals for weights with respect to the proposed composite estimator. Subsequently, we determined the sensible values of the involved weights. The aim of this procedure was to confine the superiority for different composite combinations i.e., simple direct vs. direct ratio, simple direct vs. synthetic ratio and direct ratio vs. synthetic ratio composite estimators as compared to the existing estimators.

**Key words:** domain estimation, synthetic and composite estimation, optimum weight, sensitivity performance interval.

## 1. Introduction

Generally, sample surveys are used as a cost-effective means for data collection but they are not able to provide estimates with competent precision for domains (subpopulations). Domains may be socio-demographic or geographic subdivision of the population for which separate estimates are required. Direct estimators perform better than synthetic estimators if the sample size is large for the domain while synthetic estimator is better in terms of mean square error (MSE) than direct estimator if the sample size is small for the domain along with the corresponding synthetic assumptions being satisfied, i.e., smaller area resembles larger area in their properties (Gonzalez, 1973). Further, the composite estimator is used, which is a weighted sum of two or more

[1] Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi-221005, India. E-mail: raipiyush5@gmail.com. ORCID: https://orcid.org/0000-0001-8462-4707.

[2] Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi-221005, India. E-mail: swetasingh968@gmail.com. ORCID: https://orcid.org/ 0000-0002-3275-5139.

estimators having smaller MSE in comparison with MSE of either its component estimators. Evaluation of the optimal weight for the composite estimators is generally difficult and complex in domain estimation. One of the many causes is to involve population parameters in the formula used for its estimation.

Sometimes a difficulty occurs with the weights due to sampling frame problems, which results in some sampled elements being selected with less desired probabilities. The main purpose of weighting adjustments is to reduce the bias in the survey estimates that non-response and non-coverage can cause. Also, a challenging task in the construction of composite estimator is to set the weights of each input variable. Basically, an irritant that needs to be tackled lies in assuming the knowledge of the optimum value of the weighting factor which involves the population quantities. Thus, the main concern of the present article is to develop the performance intervals of weight which ensure the superiority of composite estimators as compared to its individual component estimators.

In the absence of optimum weights, we need an interval of weight with a view to maintaining the efficiency of the composite estimator as compared to its component estimators.In this direction many works are in progress while a very rich literature is available based on estimation of weights. Agrawal and Roy (1999) discussed the performance of efficient estimators of small domains. The generalized class of composite estimator is developed and analyzed by Tikkiwal and Ghiya (2004), including group of estimators which are convexly combined with weights. Further, Pandey and Tikkiwal (2006) also discussed the generalized class of composite estimators under Lahiri-Midzuno sampling scheme. Tikkiwal and Rai (2009) also proposed composite estimators and their sensitivity interval for small domains. King-Jong Lui (2020) discussed notes on the use of the composite estimator for improvement of the ratio estimator.

Here, in the present work we considered the situation of absence of optimum weights and thus obtained the sensitivity performance intervals for weights in respect to the proposed composite estimators and figured out sensible values of the involved weights with a view to confining superiority for different composite combinations.

## 2. Notations and Formulation of the Problem

Suppose a finite population $U=\{1, 2, ..., i, ..., N\}$ is divided into 'A' domains $U_a$ having size $N_a$ ($a=1, \ldots, A$). We represent the study characteristic by 'y' and auxiliary characteristic by 'x'. A random sample 's' of size 'n' is drawn using simple random

sampling without replacement (SRSWOR) from population U such that '$n_a$' units in the sample 's' comes from domain $U_a$ ($a$=1, …, $A$). We denote

$$\sum_{a=1}^{A} N_a = N \quad \text{and} \quad \sum_{a=1}^{A} n_a = n$$

Notations used are given as follows:

$\overline{X}$ : Mean of the population based on 'N' observations of $x$.

$\overline{X}_a$ : Mean of the domain 'a' based on '$N_a$' observations of $x$.

$\overline{x}$ : Mean of the sample 's' based on 'n' observations of $x$.

$\overline{x}_a$ : Mean of the sample of domain 'a' based on '$n_a$' observations of $x$.

$\overline{Y}$ : Mean of the population based on 'N' observations of $y$.

$\overline{Y}_a$ : Mean of the domain 'a' based on '$N_a$' observations of $y$.

$\overline{y}$ : Mean of the sample 's' based on 'n' observations of $y$.

$\overline{y}_a$ : Mean of the sample of domain 'a' based on '$n_a$' observations of $y$.

Let $X_{ai}$ ($a$=1, …, $A$; $i$=1, …, $N_a$) denote the i$^{th}$ observation of a$^{th}$ domain for the characteristic $x$ and $Y_{ai}$ ($a$=1, …, $A$; $i$=1, …, $N_a$) denote the i$^{th}$ observation of a$^{th}$ domain for the characteristic y. The corresponding various mean squares and coefficient of variations of domain $U_a$ for direct estimators for study and auxiliary characteristics are given as follows:

$$S'^2_{x_a} = \frac{1}{(N_a - 1)} \sum_{i=1}^{N_a} (X_{ai} - \overline{X}_a)^2 \, , C'_{x_a} = \frac{S'_{x_a}}{\overline{X}_a}$$

$$S'^2_{y_a} = \frac{1}{(N_a - 1)} \sum_{i=1}^{N_a} (Y_{ai} - \overline{Y}_a)^2 \, , \, C'_{y_a} = \frac{S'_{y_a}}{\overline{Y}_a}$$

$$S'_{x_a y_a} = \frac{1}{(N_a - 1)} \sum_{i=1}^{N_a} (X_{ai} - \overline{X}_a)(Y_{ai} - \overline{Y}_a) \, , \, C'_{x_a y_a} = \frac{S'_{x_a y_a}}{\overline{X}_a \overline{Y}_a}$$

The corresponding various mean squares and coefficient of variations of domain $U_a$ for synthetic estimators are given as follows:

$$S_x^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (X_i - \overline{X}_a)^2 , \, C_x = \frac{S_x}{\overline{X}_a}$$

$$S_y^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (Y_i - \overline{Y}_a)^2 , \, C_y = \frac{S_y}{\overline{Y}_a}$$

$$S_{xy} = \frac{1}{(N-1)} \sum_{i=1}^{N} (X_i - \overline{X}_a)(Y_i - \overline{Y}_a) , \, C_{xy} = \frac{S_{xy}}{\overline{X}_a \overline{Y}_a}$$

## 3. Domain Estimator under Study

As we have discussed, separate estimates are required for the domain under study. There are different direct and indirect methods of estimation for the study of domain of interest. For our case, we consider the composite estimators for the estimation of domains.

### 3.1. Composite Estimators

The following three cases of composite estimators for a[th] domain are considered:

(i) Simple direct estimator with direct ratio estimator

$$\bar{y}_{c,a(1)} = \omega \bar{y}_{d,a} + (1-\omega)\bar{y}_{d,r,a}$$

where $\bar{y}_{d,a}$ = simple direct estimator and $\bar{y}_{d,r,a}$ = direct ratio estimator.

Here, the bias and MSE terms of $\bar{y}_{d,a}$ and $\bar{y}_{d,r,a}$ can be obtained as,

$$Bias(\bar{y}_{d,a}) = 0 \tag{3.1.1}$$

$$MSE(\bar{y}_{d,a}) = \bar{Y}_a^2 \frac{(N_a - n_a)}{(N_a n_a)} \frac{S'^2_{y_a}}{\bar{Y}_a^2} \tag{3.1.2}$$

$$Bias(\bar{y}_{d,r,a}) = \bar{Y}_a \frac{(N_a - n_a)}{(N_a n_a)} \left[ \frac{S'^2_{x_a}}{\bar{X}_a^2} - \frac{S'_{x_a y_a}}{\bar{X}_a \bar{Y}_a} \right] \tag{3.1.3}$$

$$MSE(\bar{y}_{d,r,a}) = \bar{Y}_a^2 \frac{(N_a - n_a)}{(N_a n_a)} \left[ \frac{S'^2_{x_a}}{\bar{X}_a^2} + \frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a \bar{Y}_a} \right] \tag{3.1.4}$$

(ii) Simple direct estimator with synthetic ratio estimator

$$\bar{y}_{c,a(2)} = \omega \bar{y}_{d,a} + (1-\omega)\bar{y}_{syn,r,a}$$

where $\bar{y}_{d,a}$ = simple direct estimator and $\bar{y}_{syn,r,a}$ = synthetic ratio estimator.

The bias and MSE of $\bar{y}_{syn,r,a}$ will be obtained as,

$$Bias(\bar{y}_{syn,r,a}) = \bar{Y}_a \left( \frac{N-n}{Nn} \right) \left[ \frac{S_x^2}{\bar{X}_a^2} - \frac{S_{xy}}{\bar{X}_a \bar{Y}_a} \right] \tag{3.1.5}$$

$$MSE(\bar{y}_{syn,r,a}) = \bar{Y}_a^2 \left( \frac{N-n}{Nn} \right) \left[ \frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a \bar{Y}_a} \right] \tag{3.1.6}$$

(iii) Direct ratio estimator with synthetic ratio estimator

$$\bar{y}_{c,a(3)} = \omega \bar{y}_{d,r,a} + (1-\omega)\bar{y}_{syn,r,a}$$

where $\overline{y}_{d,r,a}$ = direct ratio estimator and $\overline{y}_{syn,r,a}$ = synthetic ratio estimator. The bias and MSE terms of $\overline{y}_{d,r,a}$ and $\overline{y}_{syn,r,a}$ have been already mentioned above.

### 3.2. Performance Intervals for Weight ω

Let us consider composite estimator $t_3$ as a linear combination of components $t_1$ and $t_2$ i.e.,

$$t_3 = \omega_1 t_1 + \omega_2 t_2 = \omega t_1 + (1-\omega)t_2 \tag{3.2.1}$$

Here $\omega_1 + \omega_2 = 1$, where $\omega_1 = \omega$ and $\omega_2 = 1-\omega$; $\omega$ is the assigned weight.

For better performing interval of composite estimator $t_3$, MSE($t_3$) is less than equal to either of MSE($t_1$) or MSE($t_2$).Now, we have two conditions, the first one is:

$$\begin{aligned}
&MSE(t_3) \le MSE(t_1) \qquad (3.2.2)\\
&\Rightarrow MSE\{\omega t_1 + (1-\omega)t_2\} \le MSE(t_1)\\
&\Rightarrow \omega^2 MSE(t_1) + (1-\omega)^2 MSE(t_2) + 2\omega(1-\omega)c\,\mathrm{ov}(t_1,t_2) \le MSE(t_1)\\
&\left[\because MSE\{\omega t_1 + (1-\omega)t_2\} = \omega^2 MSE(t_1) + (1-\omega)^2 MSE(t_2) + 2\omega(1-\omega)c\,\mathrm{ov}(t_1,t_2);(\text{Rao, 2003})\right]\\
&\Rightarrow \omega^2 MSE(t_1) + MSE(t_2) + \omega^2 MSE(t_2) - 2\omega MSE(t_2) + 2\omega cov(t_1,t_2) - 2\omega^2\,\mathrm{cov}(t_1,t_2) \le MSE(t_1)\\
&\Rightarrow \omega^2\{MSE(t_1) + MSE(t_2) - 2\,\mathrm{cov}(t_1,t_2)\} - 2\omega\{MSE(t_2) - \mathrm{cov}(t_1,t_2)\} + \{MSE(t_2) - MSE(t_1)\} \le 0
\end{aligned}$$

On solving the above quadratic equation and assuming that the covariance term is small relative to MSE($t_2$),we get,

$$\omega = \frac{MSE(t_2) + MSE(t_1)}{MSE(t_2) + MSE(t_1)} = 1 \text{ or } \omega = \frac{MSE(t_2) - MSE(t_1)}{MSE(t_2) + MSE(t_1)}$$

As 1 is an integer value of $\omega$ , we take the other values of $\omega$ as,

$$\omega \ge \frac{MSE(t_2) - MSE(t_1)}{MSE(t_2) + MSE(t_1)} \tag{3.2.3}$$

Again, the second condition is:

$$MSE(t_3) \le MSE(t_2) \tag{3.2.4}$$

Similarly, on solving equation (3.2.4), we get,

$$\omega \le \frac{2MSE(t_2)}{MSE(t_1) + MSE(t_2)} \tag{3.2.5}$$

So, the better performing interval of $t_3$ estimator is given as,

$$\frac{MSE(t_2) - MSE(t_1)}{MSE(t_2) + MSE(t_1)} \le \omega \le \frac{2MSE(t_2)}{MSE(t_1) + MSE(t_2)} \tag{3.2.6}$$

Now, let us consider the three cases for $t_1$ and $t_2$ estimators as discussed before in previous Section 3.1, as follows:

(i) Simple direct estimator with direct ratio estimator

$$\bar{y}_{c,a(1)} = \omega\bar{y}_{d,a} + (1-\omega)\bar{y}_{d,r,a}, \text{ where } t_1 = \bar{y}_{d,a} \text{ and } t_2 = \bar{y}_{d,r,a}.$$

Putting the formulae of MSE from the expressions (3.1.2) and (3.1.4)in the expression of the left-hand part and right-hand part of (3.2.6),we get,

$$\frac{\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a\bar{Y}_a}\right]}{\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + 2\frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a\bar{Y}_a}\right]} \le \omega \le \frac{2\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + \frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a\bar{Y}_a}\right]}{\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + 2\frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a\bar{Y}_a}\right]} \tag{3.2.7}$$

(ii) Simple direct estimator with synthetic ratio estimator

$$\bar{y}_{c,a(2)} = \omega\bar{y}_{d,a} + (1-\omega)\bar{y}_{syn,r,a}, \text{ where } t_1 = \bar{y}_{d,a} \text{ and } t_2 = \bar{y}_{syn,r,a}.$$

After putting the MSE expressions, the expression (3.2.6) provides,

$$\frac{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right] - \left(\frac{N_a-n_a}{N_a n_a}\right)\frac{S'^2_{y_a}}{\bar{Y}_a^2}}{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right] + \left(\frac{N_a-n_a}{N_a n_a}\right)\frac{S'^2_{y_a}}{\bar{Y}_a^2}} \le \omega \le \frac{2\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right]}{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right] + \left(\frac{N_a-n_a}{N_a n_a}\right)\frac{S'^2_{y_a}}{\bar{Y}_a^2}} \tag{3.2.8}$$

(iii) Direct ratio estimator with synthetic ratio estimator

$$\bar{y}_{c,a(3)} = \omega\bar{y}_{d,r,a} + (1-\omega)\bar{y}_{syn,r,a}, \text{ where } t_1 = \bar{y}_{d,r,a} \text{ and } t_2 = \bar{y}_{syn,r,a}.$$

The MSE of $\bar{y}_{d,r,a}$ and $\bar{y}_{syn,r,a}$ are given by expressions (3.1.4) and (3.1.6) respectively. Thus, expression (3.2.6) provides,

$$\frac{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right] - \left(\frac{N_a-n_a}{N_a n_a}\right)\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + \frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a\bar{Y}_a}\right]}{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right] + \left(\frac{N_a-n_a}{N_a n_a}\right)\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + \frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a\bar{Y}_a}\right]} \le \omega \le$$

$$\frac{2\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right]}{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right] + \left(\frac{N_a-n_a}{N_a n_a}\right)\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + \frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a\bar{Y}_a}\right]} \tag{3.2.9}$$

### 3.3. Sensitivity Performance Intervals for Weight ω

Let us consider 'P' as the proportional inflation in the *MSE* of $t_3$ due to use of some $\omega$ other than $\omega_{opt.}$,i.e.

$$P = \frac{MSE(t_3) - MSE_{opt.}(t_3)}{MSE_{opt.}(t_3)} \tag{3.3.1}$$

For the sake of convenience neglecting the covariance term which does not hampered the equation and on substituting the formula of MSE of $t_3$ under $\omega$ and $\omega_{opt.}$, we get:

$$P = \frac{\left\{\omega^2 MSE(t_1) + (1-\omega)^2 MSE(t_2)\right\} - \left\{\omega_{opt.}{}^2 MSE(t_1) + (1-\omega_{opt.})^2 MSE(t_2)\right\}}{\left\{\omega_{opt.}{}^2 MSE(t_1) + (1-\omega_{opt.})^2 MSE(t_2)\right\}} \tag{3.3.2}$$

Divide numerator and denominator by $(1-\omega_{opt.})^2$ and taking $(1-\omega)^2$ common from first term of numerator, we have;

$$P = \left(\frac{1-\omega}{1-\omega_{opt.}}\right)^2 P_1 - 1 \tag{3.3.3}$$

where,

$$P_1 = \frac{\left(\frac{\omega}{1-\omega}\right)^2 MSE(t_1) + MSE(t_2)}{\left(\frac{\omega_{opt.}}{1-\omega_{opt.}}\right)^2 MSE(t_1) + MSE(t_2)}$$

As *P* is a ratio of two positive quantity (as numerator and denominator of *P* are positive quantity) so, $P \geq 0$, which implies

$$P_1 \geq \left(\frac{1-\omega_{opt.}}{1-\omega}\right)^2 \tag{3.3.4}$$

Therefore,

$$\frac{\left(\frac{\omega}{1-\omega}\right)^2 MSE(t_1) + MSE(t_2)}{\left(\frac{\omega_{opt.}}{1-\omega_{opt.}}\right)^2 MSE(t_1) + MSE(t_2)} \geq \left(\frac{1-\omega_{opt.}}{1-\omega}\right)^2 \tag{3.3.5}$$

On simplifying equation (3.3.5), we have;

$$(\omega + \omega_{opt.}) \geq \frac{2MSE(t_2)}{MSE(t_1) + MSE(t_2)} \tag{3.3.6}$$

Now, we have to find the optimum weight for composite estimator $t_3$.

$$t_3 = \omega t_1 + (1-\omega) t_2$$

$$MSE(t_3) = \omega^2 MSE(t_1) + (1-\omega)^2 MSE(t_2) + 2\omega(1-\omega)\operatorname{cov}(t_1, t_2) \tag{3.3.7}$$

On differentiating eq. (3.3.7) with respect to $\omega$ and equating it to zero after neglecting the covariance term, assuming that the covariance term is relatively small, we get,

$$\omega_{opt.} = \frac{MSE(t_2)}{MSE(t_1) + MSE(t_2)} \tag{3.3.8}$$

Using (3.3.6) and (3.3.8), we have,

$$\omega \geq \frac{MSE(t_2)}{MSE(t_1) + MSE(t_2)} \tag{3.3.9}$$

Thus, the sensitivity performance interval for $\omega$ is given as:

$$\frac{MSE(t_2)}{MSE(t_1) + MSE(t_2)} \leq \omega \leq \frac{2MSE(t_2)}{MSE(t_1) + MSE(t_2)} \tag{3.3.10}$$

Now, the sensitivity performance interval of the involved weight for the above three composite estimators as discussed before in previous Section 3 are given as follows:

(i) Simple direct estimator with direct ratio estimator

$$\bar{y}_{c,a(1)} = \omega \bar{y}_{d,a} + (1-\omega)\bar{y}_{d,r,a}, \text{ where } t_1 = \bar{y}_{d,a} \text{ and } t_2 = \bar{y}_{d,r,a}.$$

Here, the expression of sensitivity interval is obtained as

$$\frac{\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + \frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a \bar{Y}_a}\right]}{\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + 2\frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a \bar{Y}_a}\right]} \leq \omega \leq \frac{2\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + \frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a \bar{Y}_a}\right]}{\left[\frac{S'^2_{x_a}}{\bar{X}_a^2} + 2\frac{S'^2_{y_a}}{\bar{Y}_a^2} - 2\frac{S'_{x_a y_a}}{\bar{X}_a \bar{Y}_a}\right]} \tag{3.3.11}$$

(ii) Simple direct estimator with synthetic ratio estimator

$$\bar{y}_{c,a(2)} = \omega \bar{y}_{d,a} + (1-\omega)\bar{y}_{syn,r,a} \text{ , where } t_1 = \bar{y}_{d,a} \text{ and } t_2 = \bar{y}_{syn,r,a}$$

The sensitivity performance interval for $\omega$ in this case is obtained as

$$\frac{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a \bar{Y}_a}\right]}{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a \bar{Y}_a}\right] + \left(\frac{N_a - n_a}{N_a n_a}\right)\frac{S'^2_{y_a}}{\bar{Y}_a^2}} \leq \omega \leq \frac{2\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a \bar{Y}_a}\right]}{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2} + \frac{S_y^2}{\bar{Y}_a^2} - 2\frac{S_{xy}}{\bar{X}_a \bar{Y}_a}\right] + \left(\frac{N_a - n_a}{N_a n_a}\right)\frac{S'^2_{y_a}}{\bar{Y}_a^2}} \tag{3.3.12}$$

(iii) Direct ratio estimator with synthetic ratio estimator

$$\bar{y}_{c,a(3)} = \omega \bar{y}_{d,r,a} + (1-\omega)\bar{y}_{syn,r,a} \text{ , where } t_1 = \bar{y}_{d,r,a} \text{ and } t_2 = \bar{y}_{syn,r,a}.$$

Here, the sensitivity performance interval will be obtained as

$$\frac{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2}+\frac{S_y^2}{\bar{Y}_a^2}-2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right]}{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2}+\frac{S_y^2}{\bar{Y}_a^2}-2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right]+\left(\frac{N_a-n_a}{N_an_a}\right)\left[\frac{S_{x_a}'^2}{\bar{X}_a^2}+\frac{S_{y_a}'^2}{\bar{Y}_a^2}-2\frac{S'_{x_ay_a}}{\bar{X}_a\bar{Y}_a}\right]} \le \omega \le \frac{2\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2}+\frac{S_y^2}{\bar{Y}_a^2}-2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right]}{\left(\frac{N-n}{Nn}\right)\left[\frac{S_x^2}{\bar{X}_a^2}+\frac{S_y^2}{\bar{Y}_a^2}-2\frac{S_{xy}}{\bar{X}_a\bar{Y}_a}\right]+\left(\frac{N_a-n_a}{N_an_a}\right)\left[\frac{S_{x_a}'^2}{\bar{X}_a^2}+\frac{S_{y_a}'^2}{\bar{Y}_a^2}-2\frac{S'_{x_ay_a}}{\bar{X}_a\bar{Y}_a}\right]} \tag{3.3.13}$$

## 4. Numerical Illustration

We consider the data from Sarndal et al. (1992), Appendix B of Sweden Municipalities which are classified into eight geographical regions. We take all eight geographical regions for the study purpose with sizes 25, 48, 32, 38, 56, 41, 15 and 29 respectively and select a sample of 5, 10, 6, 8, 11, 8, 3, 6sampling units from each domain respectively. We take the study variable 'y' as RMT85 (Revenues from the 1985 municipal taxation (in millions of kronor)) and the auxiliary variable 'x' as P85 (1985 population (in thousands)). The performance intervals for weight derived in equations (3.2.7), (3.2.8), (3.2.9)and sensitivity performance intervals for weight derived in equations (3.3.11), (3.3.12), (3.3.13)are presented in Table 4.1 and 4.2 respectively.

**Table 4.1:** Performance intervals for weight of three different composite estimators

| Domain | Simple Direct with Direct Ratio $\bar{y}_{c,a(1)}$ | Simple Direct with Synthetic Ratio $\bar{y}_{c,a(2)}$ | Direct Ratio with Synthetic Ratio $\bar{y}_{c,a(3)}$ |
|---|---|---|---|
| | $\omega_1 \le \omega \le \omega_2$ | $\omega_1 \le \omega \le \omega_2$ | $\omega_1 \le \omega \le \omega_2$ |
| 1 | [-0.9843,0.0157] | [-0.9950,0.0050] | [-0.5191,0.4808] |
| 2 | [-0.9867,0.0132] | [-0.8200,0.1799] | [0.8737,1.8737] |
| 3 | [-0.8767,0.1232] | [-0.6234,0.3765] | [0.5587,1.5587] |
| 4 | [-0.6600,0.3399] | [-0.9626,0.0374] | [-0.8297,0.1703] |
| 5 | [-0.6903,0.3097] | [-0.9818,0.0182] | [-0.9043,0.0957] |
| 6 | [-0.9712,0.0288] | [-0.4356,0.5644] | [0.9284,1.9284] |
| 7 | [-0.9748,0.0252] | [-0.8565,0.1435] | [0.7166,1.7166] |
| 8 | [-0.9703,0.0297] | [-0.7278,0.2722] | [0.8253,1.8253] |

**Table 4.2:** Sensitivity performance intervals for weight of three composite estimators

| Domain | Simple Direct with Direct Ratio $\bar{y}_{c,a(1)}$ | Simple Direct with Synthetic Ratio $\bar{y}_{c,a(2)}$ | Direct Ratio with Synthetic Ratio $\bar{y}_{c,a(3)}$ |
|---|---|---|---|
| | $\omega_1 \le \omega \le \omega_2$ | $\omega_1 \le \omega \le \omega_2$ | $\omega_1 \le \omega \le \omega_2$ |
| 1 | [0.0078,0.0157] | [0.0025,0.0050] | [0.2404,0.4808] |
| 2 | [0.0066,0.0132] | [0.0899,0.1799] | [0.9368,1.8737] |
| 3 | [0.0616,0.1232] | [0.1882,0.3765] | [0.7793,1.5587] |
| 4 | [0.1699,0.3399] | [0.0187,0.0374] | [0.0852,0.1703] |
| 5 | [0.1548,0.3097] | [0.0091,0.0182] | [0.0478,0.0957] |
| 6 | [0.0144,0.0288] | [0.2822,0.5644] | [0.9642,1.9284] |
| 7 | [0.0126,0.0252] | [0.0718,0.1435] | [0.8583, 1.7166] |
| 8 | [0.0148,0.0297] | [0.1361,0.2722] | [0.9126,1.8253] |

From the above two tables we see that the performance intervals for weight of $\bar{y}_{c,a(1)}, \bar{y}_{c,a(2)}$ and $\bar{y}_{c,a(3)}$ are ranging from -0.9867 to 0.3399, -0.9950 to 0.5644 and -0.9043 to 1.9284 respectively. It means all three composite estimators retain its superiority for values of ω ranging from -0.9950 to 1.9284. Also, we observe that the length of the performance intervals for weight of composite estimators is one which follows from the expression (3.2.6). Table 4.2 clearly shows that the sensitivity performance interval for weight of composite estimators lies between 0.0025 to 1.9284.

## 5. Conclusions

Composite estimators provide efficient estimates for the unknown population parameters as compared to their constituent estimators. The estimation of weights in the composite estimators are not easy task and due to this reason, this is not a popular estimator among users and practitioners. Here, in the present study an effort is made to get sensitivity performance intervals of the weight that guarantee the superiority of the proposed composite estimator with respect to its component estimators in the field of domain estimation also.

From the above analysis of three different composite estimators, we obtain the performance intervals of weight which ensure supremacy of composite estimators as compared to their component estimators. As an example, we show that the combination of direct ratio estimator with synthetic ratio estimator performs better within performance intervals obtained in Table 4.1in terms of MSE. It is also concluded that the composite estimators for the weights lie in the sensitivity performance intervals are less varying in terms of MSE. The outcomes of the study will be useful to develop efficient composite estimators for the domain estimation in general and for small area estimation in particular.

## References

Agrawal, M. C., Roy, D. C., (1999). Efficient Estimators for Small Domains. *Journal of the Indian Society of Agricultural Statistics*, 52(3), pp. 327–337.

Ghosh, M., Rao, J. N. K., (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 91, pp. 55–93.

Gonzalez, M. E., (1973). Use and Evaluation of Synthetic Estimates. *Proceedings of the Social Statistical Section of American Statistical Association*, pp. 33–36.

Hedayat, A. S., Sinha, B. K., (1991). Design and Inference in Finite Population Sampling. *John Wiley & Sons, New York.*

Holt, D., Smith T. M. F., and Tomberlin, T. J., (1979). A Model-Based Approach to Estimation for Small Subgroups of a Population. *Journal of the American Statistical Association*, 74, pp. 405–410.

Kalton, G., Flores-Cervantes, I., (2003). Weighting Methods. *Journal of Official Statistics*, 19(2), pp. 81–97.

L. Kung-Jong, (2020). Notes on Use of the Composite Estimator: an Improvement of the Ratio Estimator. *Journal of Official Statistics*, 36(1), pp. 137–149.

Pandey, K. K., (2010). Aspects of Small Area Estimation Using Auxiliary Information. *VDM Verlag Dr.Muller e.K.*

Pandey, P. S., Kathuria, O. P., (1995). Some Composite Estimators for Small Area Estimation. *Journal of the Indian Society of Agricultural Statistics*, 47(3), pp. 262–272.

Platek, R., Rao, J. N. K., Sarndal, C. E., and Singh, M. P., (1981). Small Area Statistics: An International Symposium. *John Wiley & Sons, New York.*

Purcell, N. J., Kish, L., (1979). Estimation for Small Domains. *Biometrics*, 35, pp. 365–384.

Rai, P. K., Pandey K. K., (2013). Synthetic Estimators Using Auxiliary Information in Small Domains, *Statistics in Transition-new series*, 14(1), pp. 31–44.

Rao, J. N. K., (2003). Small Area Estimation. *Wiley Inter-Science, New Jersey, John Wiley and Sons.*

Sarndal, C. E., Swensson, B., and Wretman, J., (1992). Model assisted survey sampling. *Springer-Verlag.*

Schaible, W. L., (1979). A Composite Estimator for Small Area Statistics. *Synthetic Estimates for Small Areas:Statistical Workshop Papers and Discussion*, National Institute on Drug Abuse (NIDA) Research Monograph 24, pp. 36–53.

Tikkiwal, G .C., Ghiya, A., (2004). A generalized Class of Composite Estimators with Application to Crop Acreage Estimation for Small Domains. *Statistics in Transition*, 6, pp. 697–711.

Tikkiwal, G. C., Rai, P. K., (2009). A Composite Estimator for Small Domains and its Sensitivity Interval for Weights $\alpha$. *Statistics in Transition*, 10(2), pp. 269–275.

*STATISTICS IN TRANSITION new series, March 2024*
*Vol. 25, No. 1, pp. 191–194*

# About the Authors

**Abimibola Victoria** is a lecturer at the Department of Statistics, Faculty of Physical Sciences, University of Nigeria, Nsukka. Her areas of research interest include computational statistics, response surface methodology, design and analysis of experiments, sample survey methods, and missing and outlier data analysis. Abimibola has published in both national and international journals and conferences. She is a Fellow, Royal Statistical Society (FRSS), and a member, Nigerian Statistical Association, International Statistical Institute, and Caucus of Women in Statistics.

**Babatunde Oluwagbenga Tobi** is a lecturer at the Department of Statistics, Faculty of Physical Sciences, University of Nigeria, Nsukka. His research interests are time series analysis, sample survey and statistical process control. He has published several research articles in international journals and presented several papers in both national and international conferences. He is a member of several professional statistical bodies.

**Belhamra Thara** is a faculty member at the Department of Mathematics at The University of Badji-Mokhtar, Annaba-Algeria. She received her PhD degree in Mathematics specializing Actuarial Science from Badji-Mokhtar University, Annaba-Algeria. Her research areas are in applied statistics and actuarial science.

**Djafar Nur Mutmainnah** is a Statistics graduate from Universitas Islam Indonesia, who currently works as a data analyst in one of the media company in Jakarta, Indonesia. She maintains and analyses the media plan that has been ordered from advertisers. Her main areas of interest right now include analysing, especially commercial data, business analysis, strategic and planning, media television, advertising and agency companies.

**Fauzan Achmad** is an Assistant Professor at the Department Statistics, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia. Simultaneously he holds the position of an expert in Spatial Statistics. His main research interests revolve around spatial statistics, remote sensing and spatial data science. Currently, he is leading the way as the head of the Risk Management and Disaster Statistics Lab at the Department of Statistics.

**Gaire Arjun Kumar** is a PhD Scholar of Population studies at the Central Department of Population Studies, Tribhuvan University, Nepal. He is also a Senior Lecturer of Mathematics and Statistics at Khwopa Engineering College, Nepal. His main research area of interest includes modification of probability distribution, their properties and

application to real world problems; and differential, determination and distributional pattern of demographic variables. He has published about a dozen research papers in international/national journals and conferences. He has also published two test books of university level. He is a life member of Population Association of Nepal (APA).

**Gurgul Henryk** is a Professor at AGH University of Science and Technology in Cracow, Department of Applications of Mathematics in Economics. His research interests include econometrics, financial econometrics, international financial markets, input-output models. Author of more than 200 publications. Visiting Professor at universities in Graz and Klagenfurt (Austria), Erlangen-Nuernberg, Greifswald, Ilmenau, Saarbruecken (Germany), Trieste and Siena (Italy), Valencia (Spain), Koper (Slovenia), Joensuu (Finland). Honours: Awarded in 2007 the most prestigious prize for economics in Poland "Bank Handlowy w Warszawie S.A. Award" (Citi Bank), in 2015 awarded the Medal of Honour, University of Graz, Austria.

**Gurung Yogendra B.**, PhD, is a Professor of Population Studies at the Central Department of Population Studies, Tribhuvan University. He has experience of more than 30 years of teaching and research. He has supervised about half a dozen PhD students. His research interest includes demographic analysis and social survey research utilizing both quantitative and qualitative methods. Poverty, social inclusion/exclusion, migration and gender are the main areas of interest in social science. He has published more than a dozen of research papers in national/international journals. He has publications in a number of research monographs and chapters in edited volumes. He is a member of Asian Population Association (APA), International Union for the Scientific Study of Population (IUSSP), and Population Association of Nepal (PAN).

**Hayne Stephen** is a Professor Emeritus in the Department of Computer Information Systems in the College of Business at Colorado State University. His research has focused on developing innovative technologies to solve real business problems. His research interests lie with social networks theory and analytics, distributed systems, big data, auctions and ecommerce and detecting and mitigating large-scale distributed denial of service attacks (DDOS).

**Kałuża-Kopias Dorota,** PhD in economics, is an Assistant Professor at the Department of Demography at the Faculty Economic and Sociological University of Lodz. She is a graduate of Computer Science and Econometrics at the University of Lodz and a member of the Center for Migration Studies at the University of Lodz. Her main research areas include population migration in the context of the labour market and aging population.

**Kant Rai Piyush** is a Full Professor of Statistics at the Department of Statistics, Banaras Hindu University, Varanasi (B.H.U.), India. His research interests include sample

survey, small area estimation (sae), demography, Bayesian modelling, biostatistics & bioinformatics. Prof. Piyush has published substantial quality research articles in many national and international journals of repute and is also an active reviewer of many of them. He has also published different books and book chapters through different national and international publishers, which include Handbook of Statistics, Elsevier, Springer and IGI, Global. He is an active associate and executive member of many national and international academic statistical societies and professional bodies such as ISBA, Indian Bayesian Society, Society of Statistics and Computer Applications, Rajasthan Statistical Society.

**Kisielińska Joanna** is a Retired Professor of Economics at the Management Institute of the Warsaw University of Life Sciences. Her main areas of interest include the application of quantitative methods and computer science in economics.

**Kokoszka Piotr** is a Full Professor of Statistics. His research interests are functional data analysis and time series analysis, with emphasis on applications to finance and network engineering. Professor Kokoszka has published over 160 research papers in statistics journals as well as in journals in other fields. He has also published a research monograph and a textbook, both on functional data analysis. He serves on editorial boards of many statistics journals and is a fellow of the Institute of Mathematical Statistics.

**Lin Menting** is a PhD student in the Department of Statistics at Colorado State University. Her interests focus on functional data analysis, object-oriented statistics and applications to engineering problems.

**Makhdoom Iman** is currently an Assistant Professor at the Department of Statistics, Payame Noor University (PNU). His current position is as a faculty member with a ranking of 21 and more than 23 years of experience in teaching and researching. He has published several articles and research projects in the applied and computational statistics field. His research favourite is Bayesian inference, fuzzy probability, data sciences, and statistical learning.

**Pak Abbas** is an Associate Professor of Statistics. His research interests are multivariate statistical analysis, analysis of fuzzy data, statistical inference and data analysis in reliability. He has published more than 50 research papers in international/national journals and conferences.

**Palma Agnieszka,** PhD in mathematics, is an Assistant Professor in the Department of Demography at the Faculty of Economics and Sociology of the University of Lodz. She is a graduate of the Faculty of Mathematics and Computer Science at the University of Lodz. Her main research areas are probability theory, statistics, stochastic processes and demographic processes.

**Raman Vinoth** is an Assistant Professor at the Deanship of Quality and Academic Accreditation Department at Imam Abdulrahman Bin Faisal University, Kingdom of Saudi Arabia. He received his PhD degree in statistics from Annamalai University, Annamalai Nagar, Chidambaram, India. Doctor Vinoth has published more than 60 research papers in international/national journals and conferences. His research areas are in: Statistical Analysis, Biostatistics and Applied Statistics.

**Sahoo Nirupama** is an Assistant Professor in the School of Statistics. She is interested in sampling theory and econometrics. She successfully managed the MPhil programme, leading to the awarding of two PhD students. Her writings have been published in nineteen national/international journals and she has released one book. She has taken part in several national and international conferences, seminars, and workshops. She is involved with a number of professional associations. She is presently an Assistant Professor of Statistics at Gangadhar Meher University at Amruta Vihar, Sambalpur-768004, Odisha, India.

**Singh Sweta** is an Assistant Professor of Statistics at the Department of Mathematics and Statistics, Banasthali Vidyapith, Jaipur, Rajasthan and pursuing her research from Banaras Hindu University (BHU), Varanasi under the supervision of Prof. Piyush Kant Rai. Her research interests include sample survey and domain estimation. She has published some research articles in reputed journals and is currently working on estimation of variance under design-model approach in small area estimation.

**Syrek Robert** (PhD) is an Assistant Professor at the Institute of Economics, Finance and Management, Faculty of Management and Social Communication at the Jagiellonian University in Cracow. His main areas of interest include time series analysis, financial econometrics and modelling the dependence structures of financial time series.

**Wang Haonan** is a Full Professor of Statistics. His research interests include functional data analysis and object-oriented statistics, with emphasis on applications to network security and biomedical engineering. Professor Wang has published over 50 research papers in statistics journals as well as in journals in other fields. He is a fellow of the American Statistical Association.

**Zeghdoudi Halim** is a faculty Professor at the Department of Mathematics at the University of Badji-Mokhtar, Annaba-Algeria. He received his PhD degree in Mathematics and the highest academic degree (HDR) specializing in Probability and Statistics from Badji-Mokhtar University, Annaba-Algeria. He also did his Post Doc at the Waterford Institute of Technology- Cork Rd, Waterford, Ireland. Professor Zeghdoudi has published more than 100 research papers in international/national journals and conferences. His research areas are in actuarial science, particles systems, dynamics systems, and applied statistics. Currently, he is a member of two editorial boards: Frontiers in Applied Mathematics and Statistics - Asian Journal of Probability and Statistics.