Volume 21, Number 2, June 2020 ISSN 1234-7655

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

Safari-Katesari H., Zaroudi S., Count copula regression model using generalized beta distribution of the second kind

Hurairah A., Alabid A., Beta transmuted Lomax distribution with application

Chwila A., Żądło T., On the choice of the number of Monte Carlo iterations and bootstrap replicates in empirical best prediction

Pal S., Chaudhuri A., Patra D., How privacy may be protected in optional randomized response surveys

Tharshan R., Wijekoon P., A comparison study on a new five-parameter generalized Lindley distribution with its sub-models

Yadav A. S., Singh S. K., Singh U., Statistical properties and different estimation methods for weighted inverted Rayleigh distribution

Motoryn R., Prykhodko K., Ślusarczyk B., Asymmetry of foreign trade turnover in Ukraine and Poland

Łuczak A., Just M., Positional MEF-TOPSIS method in the assessment of the development level of complex economic phenomena for territorial units

Szczepocki P., Application of iterated filtering to stochastic volatility models based on non-Gaussian Ornstein-Uhlenbeck process

Shukla A. K., Yadav S. K., New linear model for optimal cluster size in cluster sampling

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, *Statistical Office in Rzeszów, Poland, e-mail: m.cierpial-wolan@stat.gov.pl*
Secretary
Patryk Barszcz, *Statistics Poland, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 — 608 33 66*
Technical Assistant
Rajmund Litkowiec, *Statistical Office in 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

From the Editor III

Submission information for authors VIII

Research articles

Safari-Katesari H., Zaroudi S., Count copula regression model using generalized beta distribution of the second kind 1

Hurairah A., Alabid A., Beta transmuted Lomax distribution with application 13

Chwila A., Żądło T., On the choice of the number of Monte Carlo iterations and bootstrap replicates in empirical best prediction 35

Pal S., Chaudhuri A., Patra D., How privacy may be protected in optional randomized response surveys 61

Tharshan R., Wijekoon P., A comparison study on a new five-parameter generalized Lindley distribution with its sub-models 89

Yadav A. S., Singh S. K., Singh U., Statistical properties and different estimation methods for weighted inverted Rayleigh distribution 119

Other articles:

Motoryn R., Prykhodko K., Ślusarczyk B., Asymmetry of foreign trade turnover in Ukraine and Poland 143

Łuczak A., Just M., Positional MEF-TOPSIS method in the assessment of the development level of complex economic phenomena for territorial units 157

Multivariate Statistical Analysis 2018, Łódź. Conference Papers

Szczepocki P., Application of iterated filtering to stochastic volatility models based on non-Gaussian Ornstein-Uhlenbeck process 173

Research Communicates and Letters

Shukla A. K., Yadav S. K., New linear model for optimal cluster size in cluster sampling 189

Conference reports

The XXXVIII International Conference on Multivariate Statistical Analysis 5–7 November, 2019, Łódź, Poland (Baszczyńska A., Bolonek-Lasoń K.) 201

About the Authors 209

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. III–VI

From the Editor

According to the traditional convention, the articles contained in this issue are divided into three parts: research articles, other articles and research communicates and letters, and in such a sequence they are briefly characterized here.

However, it is worthwhile mentioning first that in parallel to this – a regular ('summer') issue – a special issue of the *Statistics in Transition new series* is under preparation (the announcement of which was in the Vol. 20, No. 3, September 2019) on Statistical Data Integration (SDI), featuring papers that address theoretical, methodological and practical issues. It is edited by a group of leading experts invited by Professor Partha Lahiri, the Guest Editor of this issue envisioned as the state-of-the-art in this timely topic and one of the important transitions taking place in statistics. It will appear in August 2020.

Research articles

The major section containing research articles is opened **by Hadi Safari-Katesari's** and **Samira Zaroudi'd (USA)** paper entitled ***Count copula regression model using generalized beta distribution of the second kind***. Starting with observation that modelling claims severity for obtaining insurance premium is one of the major concerns of the insurance industry – as it is evidenced by considerable amount of literature devoted to the actuarial application of the copula model to calculate the pure premium – authors model claims severity for computing the pure premium in the collision market. They apply a regression model using a generalized Beta distribution of the second kind (GB2) to compute the premium for an average claim and the conditional computation for all coverage levels, under assumption that that the number of accidents is independent from the size of claims. Application is demonstrated using a portfolio of a major automobile insurer in Iran in 2007–2008, with a subsample of 59,547 policies available in their portfolio, showing that there is strong positive dependency between the real premium and the estimated one.

Ahmed Hurairah and **Abdelhakim Alabid (YEMEN)** present ***Beta transmuted Lomax distribution with applications***. In particular, they propose and test a composite generalizer of the Lomax distribution. Specifically, they use the beta distribution and transmuted map to develop the so-called beta transmuted Lomax (BTL) distribution. They discuss the properties of the distribution and provide explicit expressions derived for the moments, mean deviations, quantiles, distribution of order statistics and

reliability. The maximum likelihood method is used for estimating the model parameters, and the finite sample performance of the estimators is assessed by simulation. The authors also demonstrate the usefulness of the new distribution in analysing real data.

In the paper ***On the choice of the number of Monte Carlo iterations and bootstrap replicates in empirical best prediction*****, Adam Chwila** and **Tomasz Żądło (POLAND)** discuss the properties of the EBPs (Empirical Best Predictors) in the context of small area estimation. In the case of longitudinal surveys, this class of predictors can be used to predict any given population's or subpopulation's characteristics, for any time period, including future periods. Generally, the value of an EBP is computed by means of Monte Carlo algorithms, whereas its MSE is usually estimated using the parametric bootstrap method. Model-based simulation studies of the properties of the predictors require numerous repetitions of the random generation of population data. This leads to a question about the dependence between the number of iterations in all the procedures and the stability of the results. Authors aim to show this dependence along with proposing methods of choosing the appropriate number of iterations using a set of real economic longitudinal data, which are available on the US Census Bureau website.

Sanghamitra Pal's, Arijit Chaudhuri's and **Dipika Patra's (INDIA)** article ***How privacy may be protected in optional randomized response surveys*** addresses the sensitive issue of protection privacy on such stigmatizing features like alcoholism, history of tax-evasion habits, testing positive for AIDS-related testing, etc., in survey research through a proper application of randomized response (RR) techniques (RRT). Authors discuss how the approach needs amendments while applying optional RRT's covering qualitative characteristics, permitting a sampled respondent either to directly reveal the sensitive characteristic or go for a randomized response respectively with complementary probabilities. They conclude that all of the competing ORR methods show satisfactory results in terms of ACP and ACV values

Ramajeyam Tharshan and **Pushpakanthie Wijekoon (SRI LANKA)** present ***A comparison study on a new five-parameter generalized Lindley distribution with its sub-models*** using a new generalization of the Lindley distribution based on a mixture of exponential and gamma distributions with different mixing proportions, and compare its performance with its sub-models. The new distribution accommodates the classical Lindley, Quasi Lindley, Two-parameter Lindley, Shanker, Lindley distribution with location parameter, and Three-parameter Lindley distributions as special cases. Various structural properties of the new distribution are discussed and the size-biased and the length-biased are derived. A simulation study is conducted to examine the mean square error for the parameters by means of the method of maximum likelihood. Finally, simulation studies and some real-world data sets are used to illustrate its flexibility in terms of its location, scale and shape parameters.

In the paper ***Statistical properties and different estimation methods for weighted inverted Rayleigh distribution*****, Abhimanyu Singh Yadav., S. K. Singh** and **Umesh Singh (INDIA)** introduce a new weighted probability distribution to model the non-monotone failure rate pattern for survival data. The proposed distribution is generalized by considering inverted Rayleigh distribution as a baseline distribution called an extended weighted inverted Rayleigh distribution. Different statistical properties such as moment, quantile function, moment generating function, entropy measurement, Bonferroni and Lorenz curve, stochastic ordering and order statistics have been derived. Authors discuss procedures to estimate the unknown parameters of the proposed probability distribution and conduct the Monte Carlo simulation study to compare the performances of the proposed estimators obtained through various methods of estimation. Using two real data sets they also show the applicability of the proposed model in a real-life situation.

Other articles

The next paper ***Asymmetry of foreign trade turnover in Ukraine and Poland*** by **Ruslan Motoryn, Kateryna Prykhodko** and **Bogusław Ślusarczyk (POLAND** and **UKRAINE)** is devoted to searching for determinants of the asymmetry of foreign trade turnover between Ukraine and Poland, based on an analysis of competitiveness indicators of the studied countries in the period 2003–2017. The emphasis is on calculation of the comparative advantages of particular commodity headings in Polish exports in the domestic market of Ukraine. Potential directions of the intensification of bilateral trade were evaluated.

Aleksandra Łuczak and **Małgorzata Just (POLAND)** in the paper entitled ***Positional MEF-TOPSIS method in the assessment of the development level of complex economic phenomena for territorial units*** propose a new methodological approach to the construction of a synthetic measure for the case where the objects are described by variables with strong asymmetry, and extreme values (outliers) are present. Authors observe that extreme value (very large or very small) of a variable may significantly affect the attribution of an excessively high or low rank in the final ranking of objects. This dependence is particularly apparent when using the classical TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method. The aim of the study is to present the application potential of the positional MEF-TOPSIS method for the assessment of the level of development of complex economic phenomena for territorial units. The proposed approach is used to assess the financial self-sufficiency of Polish municipalities in 2016. The study finally compares the results of applications of positional MEF-TOPSIS and the classic and positional TOPSIS methods.

Multivariate Statistical Analysis 2018, Łódź. Conference Papers

Piotr Szczepocki in the paper on ***Application of iterated filtering to stochastic volatility models based on non-Gaussian Ornstein-Uhlenbeck process (POLAND)*** discusses a class of Barndorff-Nielsen's and Shephard's (2001) stochastic volatility models in which the volatility follows the Ornstein–Uhlenbeck process driven by a positive Levy process without the Gaussian component. Of particular interest is the problem of parameter estimation of these models in the case when the likelihood function is not available in a closed-form expression. The main goal of the paper is to present an application of iterated filtering for parameter estimation of such models. Iterated filtering is a method for maximum likelihood inference based on a series of filtering operations, which provide a sequence of parameter estimates that converges to the maximum likelihood estimate. An application to S&P500 index data shows the model performs well and diagnostic plots for iterated filtering ensure convergence iterated filtering to maximum likelihood estimates. Empirical application is accompanied by a simulation study that confirms the validity of the approach utilizing Barndorff-Nielsen's and Shephard's stochastic volatility models.

Research Communicates and Letters

In the paper ***New linear model for optimal cluster size in cluster sampling*** by **Shukla Alok Kumar,** and **Yadav Subhash Kumar, (INDIA)** a nonlinear model has been proposed for an improved relationship between the size of the cluster and the variance within the cluster. This model describes the most appropriate functional relation between the within-cluster variance and the cluster size. Using this model, authors obtain the optimum size of the cluster and the estimate of the variance between the clusters. The proposed model leads to further improvement in the estimation of the optimum size of the cluster and the formula for determination of the optimum cluster size also leads to explicit solution of models.

Conference reports

The XXXVII International Conference on Multivariate Statistical Analysis 5–7 November, 2019), Łódź, Poland (Baszczyńska Aleksandra, Bolonek-Lasoń Katarzyna).

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. VII

Submission information for Authors

Statistics in Transition new series (SiT) 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: http://stat.gov.pl/en/sit-en/guidelines-for-authors/

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. IX–X

Editorial Policy

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:

Databases indexing the journal:

- 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
- Ulrichsweb & Ulrich's Periodicals Directory
- WanFang Data
- WorldCat (OCLC)
- Zenodo.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 1–12, DOI 10.21307/stattrans-2020-011
Submitted – 13.04.2019; paper accepted for published – 10.03.2020

Count copula regression model using generalized beta distribution of the second kind

Hadi Safari-Katesari[1], **Samira Zaroudi**[2]

ABSTRACT

Modelling claims severity for obtaining insurance premium is one of the major concerns of the insurance industry. There is a considerable amount of literature on the actuarial application of the copula model to calculate the pure premium. In this paper, we model claims severity for computing the pure premium in the collision market by means of the count copula model. Moreover, we apply a regression model using a generalized beta distribution of the second kind (GB2) to compute the premium for an average claim and the conditional computation for all coverage levels. Like many other researchers, we assume that the number of accidents is independent from the size of claims. For real data application, we use a portfolio of a major automobile insurer in Iran in 2007-2008, with a subsample of 59,547 policies available in their portfolio. We then proceed to compare the estimated premiums with the real premiums. The results demonstrate that there is strong positive dependency between the real premium and the estimated one.

Key words: count copula, GB2 regression, pure premium, collision insurance.

1. Introduction

Premium is the payment that a policyholder pays for buying full or partial insurance coverage versus a specified risk. Premium ratemaking is a vital subject to balancing insurance payments (Zhang et al., 2015). In confronting with financial outcomes of the random phenomenon, insurance plays the role of supporting policyholders. It includes the accumulation of a big bunch of policyholder risks such that, within a given time cycle, a number of insurance claims and an accumulated loss to the insurer can be determined. Nowadays, estimating premium plays a pivotal role for insurance companies in the competitive markets. The biased computation may lead to losing the market share and confronting ruin. There is a range of works in this field such as Weisberg (1982), David (2015), Marton et al. (2015), Zhang et al. (2015), Schirmacher (2016), Yang et al. (2017), Shi and Yang (2018), Lesmana et al. (2018), Wolny-Dominiak et al. (2018) and Avanzi et al. (2019). However, using the copula model in ratemaking and actuarial application is to some extent new. Frees et al. (2013) used a multivariate two-part regression model such that the correlation ratio and copula regression for the claims and severity modelling were considered, respectively.

[1]Corresponding Author: Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA. E-mail: hadi.safari@siu.edu. . ORCID: https://orcid.org/0000-0003-2630-3133
[2]Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, USA. ORCID: https://orcid.org/0000-0001-8290-6137

For more information, one can refer to Shi (2016). We cannot distinguish risky policyholders beforehand but the severity and the number of claims in a portfolio of an insurance company are predictable. In this paper, our aim is to calculate pure premium by using the insured coverage selection preferences and the number of accidents during the policy period. For this goal, the authors use a generalized beta distribution of the second kind (GB2) regression model to model the average claims for each level of coverage preferences. In the actuarial literature, the assumption between the number of accidents and the size of the claim is common, which is used here as well. The wide variety application of the probabilistic model for claim severities can be justified by the "long-tail" nature of insurance losses, which appear as a result of the delay in reporting and long settlement periods of claims. So, this matter makes it difficult to evaluate the exact price of some liability insurance products and actuary job is to compute the average loss, or pure premium, for different classes of insurance products for fairly rating insurance policies. They use the observed past claim data from a portfolio of an insurance company for predicting the future pure premium for a determined period. Shi and Valdez (2011) and Katesari and Vajargah (2015) used count Copula model for examining asymmetric information in the insurance industry. The former computed pure premium using the information of selected coverage level and loss number for a specific year, and here with following the latter we try to compute the premium. Frees and Valdez (2008) computed premiums under alternative reinsurance coverage. However, Katesari and Zarodi (2016) predicted accident probability after observing the accidents for a specific year by using the copula model in the latter.

In this paper, we use the count copula model for computing the pure premium of the severity data from a major insurance company in Iran. Specifically, we consider the generalized beta distribution of the second kind (GB2) regression model for the severity claims. For this, we need the joint distribution of coverage selection and the risk of policyholders. An ordered multinomial model is used to measure the coverage levels and a negative binomial regression model is used to measure the risk of policyholders for the specific year. Moreover, a copula regression model is used to measure the linear and nonlinear dependence between these two margins and the estimated results are presented. The estimation results of the fitted model using Frank copula is available in Katesari and Vajargah (2015). Instead, we use another tow famous members of the Archimedean copula family that is Clayton and Gumbel to measure this dependence. The benefit of our bivariate copula regression model is that it provides the joint distribution of coverage levels and the risk of policyholders. We exploit this joint distribution in conditional expectation for computing the pure premium of the severity data. For real data application, we use a portfolio of major automobile insurer in Iran in the calendar year 2007-2008 with a subsample of 59,547 policies in their portfolio. Also, this dataset was used to the work of Katesari and Vajargah (2015) to test asymmetric information in the collision insurance portfolio of this company.

We have organized the remains of the article as follows. In Section 2, the data description is given. In Section 3, the count copula regression model will be considered for computing the pure premium and the estimation results are given. In Section 4, premium estimation is presented and the results are compared with the actual premium. Finally, in Section 5 we provide some concluding remarks.

2. Data attributes

For fitting the model, we use a portfolio of a major automobile insurer in Iran in 2007-2008 with a total of 59,547 policies available in their portfolio. According to the policy of the company, policyholders buy insurance policy from these main and overall claims: overall accident, overall theft and overall fire. Furthermore, policyholders are able to purchase one or more coverage options from the below items:

1. Damaged caused by flood, earthquake and hurricanes,
2. Broken glass,
3. Stolen parts and accessories of vehicle,
4. Damage caused by spills or splashes of paint, acid and chemicals,
5. Compensation by not using the vehicle in repair period,
6. Slippage (only in minor damage).

For our purpose, we ordered the levels as follows:

1. overall coverage of collision insurance,
2. overall coverage of collision insurance as well as one or two more item(s),
3. (comprehensive) overall coverage of collision insurance plus three, four, five or all of more item(s).

Note that with increasing the levels (from 1 to 3), the insured coverage will increase. The dataset comes from a major insurer in Iran and we use a subsample of 59,547 cases from more than 800,000 recorded cases the portfolio in 2007-2008 for this insurer. One can find frequency statistics of policy selection and the number of losses in Table (1).

Table 1: Frequency statistics of policy selection and number of losses

	Levels				
Claims	1	2	3	Total Number	Percent
0	30176	20033	4879	55088	92.51
1	405	1497	2130	4032	6.77
2	39	161	192	392	0.66
3	2	11	21	34	0.06
4	0	1	0	1	0.00
Total Number	30622	21703	7222	59547	
Percent	51.42	36.45	12.13		100

Like every insurance database, more than 90 percent of the policyholders did not have an accident during the considered year. Moreover, Table (2) elaborates the available covariates

Table 2: Descriptive statistics of the covariates

Variable	Explanation			Level 1		Level 2		Level 3	
		Mean	StdDev	Mean	StdDev	Mean	StdDev	Mean	StdDev
Driver attributes									
Sexinsured	=1 F, 0 M	0.2014		0.1694		0.2306		0.2779	
NCD	=1(0-15%)	0.4383		0.4538		0.4289		0.3827	
	=2(15-30%)	0.2156		0.2187		0.2131		0.2066	
	=3 (30-45%)	0.2529		0.2393		0.2684		0.2725	
	=4 ($\geq$ 45%)	0.9320		0.0882		0.0896		0.1382	
Vehicle attributes									
Vage		4.2951	3.4921	4.5838	4.0601	3.8714	2.5323	4.2597	2.8845
Vtype	=Sedan	0.8849		0.8002		0.9882		0.9818	
	=Others	0.1151		0.1998		0.0117		0.0182	
Vapplication	=Personal	0.8632		0.7672		0.9798		0.9741	
	=Non-Personal	0.1368		0.2328		0.0202		0.0259	

Table 3: Severity size by months for the calendar year 2007-2008

	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
Severity	1.3	4	6.3	9.5	13.1	13.8	19.2	21.1	26.2	28.5	26.8	17.2

in the dataset. One can classify each of these covariates as a driver or vehicle attributes. Vehicle age (Vage), vehicle type (Vtype: Sedan or Non-Sedan) and vehicle application (Vapplication: Personal or Non-Personal) are vehicle attributes while sex (Female and Male) and No Claim Discount (NCD) are driver attributes. As can be seen from Table (2), many of these covariates are categorical, which demonstrates the proportion of an observed variable in each class. Moreover, both mean and standard deviation are presented for vehicle age, which is the only continuous covariate in this dataset. Like Shi and Valdez (2011), we used average claims in the observed calendar year. Table (3) provides summary of the severity claims for different months of the year 2007-2008. As demonstrated in Table (3), the majority of the policyholder's loss, nearly 28.5 in this case, occurred in January and the minority of the policyholder's loss, roughly 1.3, occurred in April. One of our restrictions is that the amounts of these losses are adjusted and we cannot distribute the exact amounts to all.

3. Count copula model fitted to the data

A bivariate copula $C(.,.)$ is a joint cumulative distribution function $C : [0,1] \longrightarrow [0,1]^2$. The application of copula comes from Sklar's theorem. Sklar (1959) says that for random variables y_1 and y_2 with corresponding marginal distributions $F_1(y_1)$ and $F_2(y_2)$, the bivariate distribution $F(y_1, y_2)$ can be stated as

$$F(y_1, y_2) = C(F_1(y_1), F_2(y_2); \theta) \tag{1}$$

where C is a copula function with dependence parameter θ. If the marginal distributions are continuous, then the copula in Equation (1) is unique, otherwise C is uniquely determined on $RanF_1 \times RanF_2$. In Shi and Valdez (2011) and Katesari and Vajargah (2015), count copula models were used for testing asymmetric information and adverse selection in automobile insurance market. Here, according to our database, we take y_{i1} and y_{i2} as selected coverage level and loss number, correspondingly, for each policyholder. y_{i1} shows the selected coverage level such that first level (overall), second level, and third level (comprehensive) coverages are connected with possible values 1, 2 or 3, correspondingly. We use latent variables y_{i1}^* and y_{i2}^*, for modelling y_{i1} and y_{i2} with a parametric copula $C(.,.)$.The joint probability mass function of y_{i1} and y_{i2} can be express as:

$$\begin{aligned} f_i(y_{i1},y_{i2}) =& C(F_{i1}(y_{i1}),F_{i2}(y_{i2})) - C(F_{i1}(y_{i1}-1),F_{i2}(y_{i2})) \\ & - C(F_{i1}(y_{i1}),F_{i2}(y_{i2}-1)) + C(F_{i1}(y_{i1}-1),F_{i2}(y_{i2}-1)) \end{aligned} \tag{2}$$

where F_{i1} and F_{i2} are the CDF of y_{i1} and y_{i2}, correspondingly. Now, we need to calibrate the marginal distribution functions of F_{i1} and F_{i2} for model identification (Shi and Valdez, 2011). For coverage level and catching the connection between y_{i1} and y_{i1}^*, we use an ordered multinomial model as follows:

$$y_{i1} = \begin{cases} 1, & \text{if } y_{i1}^* \le \alpha_1 \\ 2, & \text{if } \alpha_1 \le y_{i1}^* \le \alpha_2 \\ 3, & \text{if } y_{i1}^* > \alpha_2 \end{cases},$$

where α_1 and α_2 are unknown and should be estimated. Also, for estimating y_{i1}, we fit an ordered logistic regression model as follows:

$$F_{i1}(y_{i1}) = \begin{cases} \dfrac{1}{1+exp(-(\alpha_1 - \mathbf{x_i}'\beta))}, & \text{if } y_{i1}=1 \\ \dfrac{1}{1+exp(-(\alpha_2 - \mathbf{x_i}'\beta))}, & \text{if } y_{i1}=2 \\ 1, & \text{if } y_{i1}=3 \end{cases} \tag{3}$$

where $\mathbf{x_i}$ is the vector of covariates used for the coverage level of the ith policyholder. Another marginal variable y_{i2} can be calibrated by using a negative binomial regression model. Like Shi and Valdez (2011), we define its probability mass function as follows:

$$f_{i2}(y_{i2}) = Pr(Y_{i2}=y_{i2}) = \frac{\Gamma(y_{i2}+\psi)}{\Gamma(\psi)\Gamma(y_{i2}+1)}\left(\frac{\psi}{\psi+\lambda_i}\right)^{\psi}\left(\frac{\lambda_i}{\psi+\lambda_i}\right)^{y_{i2}} \tag{4}$$

where ψ is the dispersion parameter for policyholder i, and we use a log link function for the conditional mean that is $Y_{i2}|\mathbf{z}_i$. Note that $\mathbf{z}_i$ is the vector of covariates used for the risk of the ith policyholder. For estimating this model, we can use maximum likelihood method. The copula functions of the Gumbel and Clayton can be expressed, respectively, as follow:

$$C(u_1,u_2;\theta) = exp\{-[(-logu_1)^{\theta} + (-logu_2)^{\theta}]^{1/\theta}\}, \theta \ge 1 \tag{5}$$

Table 4: Estimation results of Clayton copula model for all reported accidents

Choice-Cumulative Logit			Risk-Negative Binomial		
	Estimate	StdErr		Estimate	StdErr
Choice -α_1	-0.9033	0.0131			
Choice-α_2	0.3767	0.0125	Risk-intercept	-2.1585	0.0194
Choice-sex (F)	0.4938	0.0194	Risk-sex (F)	0.0189	0.9339
Choice-Vage	0.0614	0.0012	Risk-Vage	0.0235	0.0015
Choice-(NCD=2)	0.0675	0.0199	Risk-(NCD=2)	-0.4443	0.0389
Choice-(NCD=3)	0.2301	0.0190	Risk-(NCD=3)	-0.7799	0.0441
Choice-(NCD=4)	0.1362	0.0269	Risk-(NCD=4)	-1.3939	0.0826
Choice-Vapplication (2)	-0.0858	0.1036	Risk-Vapplication (2)	-0.3052	0.2911
Choice-Vtype (2)	-0.1704	0.0698	Risk-Vtype (2)	-0.3052	0.2911
			Dispersion	0.8326	0.0734
Dependence parameter θ	**0.2615**	0.0184			
-2Loglikelihood	160165.1				

$$C(u_1,u_2;\theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, \theta > 0 \tag{6}$$

where θ is the dependence parameter that shows the amount of association between two marginals. For more details about application of copula in finance and actuarial science, see Frees and Valdez (1998), Cherubini et al. (2004), Joe (2014), Zaroudi et al. (2018a), Zaroudi et al. (2018b), and Shi and Yang (2018). In the similar work of Katesari and Vajargah (2015), they explained the problems arising from adverse selection based on copula model. They estimated parameter of Frank copula with $\theta = 1.3$ referred to the existence of adverse selection in their dataset. In this paper, we are interested in modelling the severity of claims and we will use the modelled copula for computing the pure premium with the same database.

The estimation results of the fitted model by using the maximum likelihood method for Frank copula is available in Table 2 of Katesari and Vajargah (2015). Here, we fit the aforementioned model using the maximum likelihood method for two other members of the Archimedean copula family, which are Gumbel and Clayton in equations (5) and (6), respectively. The estimation results of these two famous copulas are presented in Table (4) and Table (5). As can be seen from the results of Table (4) and Table (5), the dependence parameter θ for Clayton and Gumbel copula is 0.2615 and 1.10207, respectively. These results show a strong dependence between coverage level and the risk of policyholders in the portfolio of the insurance company in Iran.

4. Computing premium

Here, we describe, discuss and compute the pure premium formula from a mathematical viewpoint and then compare it with the gross premium in the original data. We define the premium by $\prod_X$ that an insurance company charges to pay a loss X, which is a random variable. Thus, a premium formula is of the form $\prod_X = \phi(X)$ where ϕ is some function. At first, we consider the mean of X and the simplest premium, which is called pure risk premium ($\prod_X = E(X)$), which means the pure premium is equal to the insurer's expected claims under the considered risk (Dickson, 2016). Additional statistical properties of the

Table 5: Estimation results of Gumbel copula model for all reported accidents

Choice-Cumulative Logit			Risk-Negative Binomial		
	Estimate	StdErr		Estimate	StdErr
Choice-α_1	-0.9032	0.0131			
Choice-α_2	0.3768	0.0126	Risk-intercept	-2.1590	0.0193
Choice-sex (F)	0.4947	0.0194	Risk-sex (F)	0.0188	0.0151
Choice-Vage	-0.0194	0.2013	Risk-Vage	0.0172	1.6611
Choice-(NCD=2)	0.0671	0.0209	Risk-(NCD=2)	-0.4435	0.0391
Choice-(NCD=3)	0.2299	0.0191	Risk-(NCD=3)	-0.7780	0.0413
Choice-(NCD=4)	0.1633	0.0276	Risk-(NCD=4)	-1.3928	0.1828
Choice-Vapplication (2)	-1.7275	0.0765	Risk-Vapplication (2)	0.3179	0.0280
Choice-Vtype (2)	-1.8164	0.1371	Risk-Vtype (2)	-0.5796	0.2811
			Dispersion	0.8321	0.0739
Dependence parameter θ	**1.10207**	0.4579			
-2Loglikelihood	160164.8				

premium computation were explored in Dickson (2016). Our data comes from a big insurance company in Iran. In this section, we use the severity of losses for one year (in the year 2007-2008) for this insurer with a sample of 62,602 policyholders out of total 800,769 recorded cases. Here, we compute the pure premium by using the selected coverage level and the number of losses for the year in the work of Katesari and Vajargah (2015). The common method for price evaluation in the automobile insurance market is modelling the number and severity of losses separately. In reality, the independence assumption between the number and severity of losses is straightforward and we need to model the size of claims to compute the pure premium. So, we compute the claims mean for each of the three levels of coverage using a regression model of Generalized Beta distribution of the second kind (GB2) . The density function of GB2 with four positive parameter goes as follows (Kleiber and Kotz, 2003):

$$f(x) = \frac{ax^{ap-1}}{b^{ap}B(p,q)\{1+(x/b)^a\}^{p+q}}, \qquad x>0, \qquad a,b,p,q>0, \tag{7}$$

where b is a scale parameter and a, b, c are shape parameters and $B(p,q)$ is the usual Euler beta function. For more information about GB2, one can refer to McDonald and Butler (1987), Sun et al. (2008), Frees and Valdez (2008) and Shi and Valdez (2011). Here, we follow the same way of Shi and Valdez (2011) with taking as $b_i = exp(l_i'\beta)$, where l_i' and β show covariates vector for each policyholder and the coefficients, respectively. In our GB2 regression model, sets of parameters for estimation purpose are (β^j, a^j, p^j, q^j) , with possible values $j = 1,2,3$, which show the three selected coverage levels, correspondingly. Table (6) shows the results of estimating the three sets of parameters by using the likelihood-based estimation method. Figure (1) demonstrates the pp-plots of the residuals from the three regression models of GB2 for showing the quality of the fitted model. According to the copula method that was used in Shi and Valdez (2011), we can additionally compute the impact of the policyholder's coverage preference y_{i1} on the number of losses (accidents) y_{i2},

Table 6: Estimate results of the GB2 regressions for all coverage levels

	1st level Estimate	StdError	2nd level Estimate	StdError	3rd level Estimate	StdError
a	3.1465	0.0016	4.6151	0.0100	3.7952	0.7432
Intercept	2.3920	0.0037	3.3120	0.0071	3.3122	0.0650
sexinsured (F)	0.0124	0.0095	0.0088	0.0003	- 0.0445	0.0286
NCD=2	0.0357	0.0024	0.0421	0.0416	- 0.0107	0.0099
NCD=3	0.1068	0.0052	0.0449	0.0003	0.0741	0.0683
NCD=4	0.2026	0.0010	0.1367	0.1164	- 0.0204	0.0835
Vapplication (2)	0.1193	0.0088	- 0.0802	0.0016	- 0.0056	0.0477
Vage	- 0.0001	0.0001	- 0.0063	0.0000	- 0.0063	0.0059
Vtype (2)	- 0.1443	0.0117	0.0736	0.0014	0.0256	0.1407
p	9.9716	0.0001	0.8890	0.0429	1.0804	0.2853
q	0.3398	0.0091	0.2567	0.0079	0.3388	0.0808
-2Loglikelihood	10127.15		14029.72		21469.40	

conditionally by using Bayes' formula:

$$Pr(Y_{i2} = y_{i2}|Y_{i1} = y_{i1}) = f_{i2|1}(y_{i2}|y_{i1},x,z) \times \frac{f_i(y_{i2},y_{i1}|x,z)}{f_{i1}(y_{i1}|x)}. \tag{8}$$

By applying this conditional formula, we can anticipate the likelihood of the number of claims, condition on the policy selection. In the above equation, the joint probability distribution in the numerator can be computed by copula distribution in equation (2) and obviously the marginal distribution of y_{i1} in the denominator by equation (3). According to the coverage selection for y_{i1}, we can conditionally compute the pure premium for the ith policyholder as follows:

$$\begin{aligned} \prod_i &= E(Y_{i2}|Y_{i1} = y_{i1}) \times E(X_i|Y_{i1} = y_{i1}) \\ &= \sum_{y_{i2}=0}^{\infty} y_{i2} f_{i2|1}(y_{i2}|y_{i1},x,z) \times \frac{exp(l_i^{'}\beta^{y_{i1}})B(p^{y_{i1}} + (1/a)^{y_{i1}}, q^{y_{i1}} - (1/a)^{y_{i1}})}{B(p^{y_{i1}}, q^{y_{i1}})} \end{aligned} \tag{9}$$

where $B(p,q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$, $\prod_X$ is the pure premium for the ith policyholder and $\Gamma(.)$ is the Gamma function. Using the above formula, we are able to compute the pure premium for each policyholder in our dataset.

Dependency coefficients among the real gross premium and the estimated one for all coverage levels have been computed by Spearman's rho and demonstrated in Table (7), which shows strong positive dependency. This strong positive correlation shows that the actual premium paid by the policyholder is according to the conditional computation. In comparison with the results of Shi and Valdez (2011), one can see the same positive dependency in the portfolio of automobile insurance in Singapore. More precisely, the dependency between real and computed premiums for the first, second and third levels of our work is 0.6636, 0.2328, and 0.8372, respectively. This is while that the dependency between real and computed premiums for the first, second and third levels of the work of Shi and Valdez (2011) is 0.58282, 0.62215, and 0.80632, respectively. Also, descriptive statistics of the real

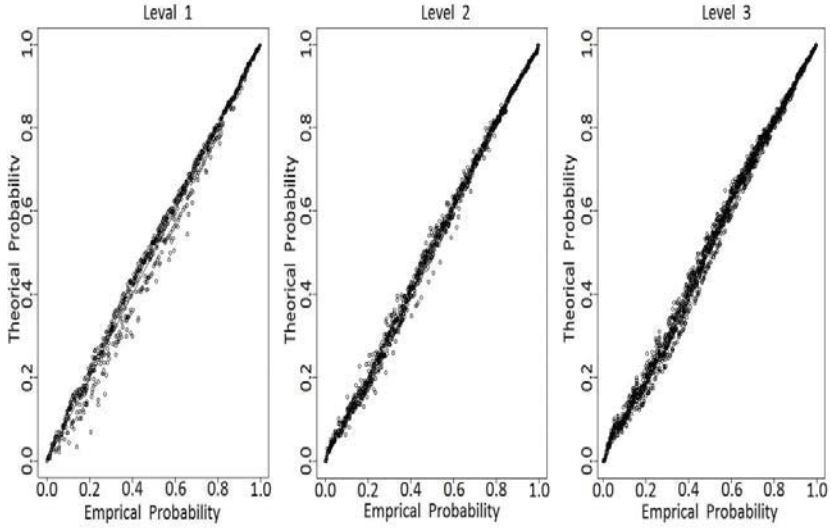


Figure 1: pp-plot for GB2 regression models.

Table 7: Dependency between real and computed premiums

	dependency	p-value
first level	0.6636	0.0098
second level	0.2328	0.0223
third level	0.8372	0.0025

and computed premiums for all coverage levels have been shown in Table (8). These results are not surprising at all and we expected the positive difference between the two premiums. This positive difference can be justified by covering loading expenses such as profits, taxes and other administrative charges, which the policyholder should pay for them as well.

Table 8: Comparison of real and computed premiums

	first level		second level		third level	
	Mean	StdDev	Mean	StdDev	Mean	StdDev
Real	23.3024	17.4888	17.6657	14.7873	19.6919	18.4127
Estimated	15.6222	9.1653	13.4348	7.8841	14.3349	16.8840

5. Conclusions

The main focus of this paper is to compute pure premium by using copula models in the automobile insurance market. We applied a GB2 regression model to compute the claims mean and conditional computation for all coverage levels. This model permits us to compare

real and estimated premiums. For this comparison, the coverage level of policyholders is fitted using an ordered multinomial model and the risk of the policyholder is measured with a negative binomial regression model in the specific year. The difficulty of this method is to modelling two count variables for finding the joint distribution, which is useful in computing the pure premium for the *i*th policyholder. To address this problem, we used a copula regression model, which builds a bivariate distribution function and measures both linear and nonlinear dependency between marginal distributions. For testing the quality of our model we used pp-plots of residuals of the fitted model. The estimation results of our model showed a strong positive dependence between real and estimated premiums.

One of our restrictions in this research is that we used a cross-sectional dataset to fit our model. If we could use a longitudinal dataset that followed each policyholder's records during the years, we would reach out to more knowledgeable results.

REFERENCES

AVANZI, B., TAYLOR, G., WONG, B., YANG, X., (2019). A Multivariate Micro-Level Insurance Counts Model With a Cox Process Approach, *UNSW Business School Research Paper*, (2019ACTL02).

CHERUBINI, U., LUCIANO, E., VECCHIATO, W., (2004). *Copula methods in finance*. John Wiley and Sons.

DAVID, M., (2015). Auto insurance premium calculation using generalized linear models, *Procedia Economics and Finance*, 20, pp. 147–156.

DICKSON, D. C., (2016). *Insurance risk and ruin*. Cambridge University Press.

FREES, E. W., VALDEZ, E. A., (1998). Understanding relationships using copulas, *North American actuarial journal*, 1; 2(1), pp. 1–25.

FREES, E. W., VALDEZ, E. A., (2008). Hierarchical insurance claims modeling, *Journal of the American Statistical Association*, 103(484), pp. 1457–1469.

FREES, E. W., JIN, X., LIN, X., (2013). Actuarial applications of multivariate two-part regression models, *Annals of Actuarial Science*, 7(2), pp. 258–287.

JOE, H., (2014). *Dependence modeling with copulas*. Chapman and Hall/CRC.

KATESARI, H. S., VAJARGAH, B. F., (2015). Testing Adverse Selection Using Frank Copula Approach in Iran Insurance Markets, *Mathematics and Computer Science*, 15, pp. 154–158.

KATESARI, H. S., ZARODI, S., (2016). Effects of Coverage Choice by Predictive Modeling on Frequency of Accidents, *Caspian Journal of Applied Sciences Research*, 5, pp. 28–33.

KLEIBER, C., KOTZ, S., (2003). *Statistical size distributions in economics and actuarial sciences*, Vol. 470, John Wiley and Sons.

LESMANA, E., WULANDARI, R., NAPITUPULU, H., SUPIAN, S., (2018). Model estimation of claim risk and premium for motor vehicle insurance by using Bayesian method, *In IOP Conference Series: Materials Science and Engineering* (Vol. 300, No. 1, p. 012027), IOP Publishing.

MARTON, J., KETSCHE, P. G., SNYDER, A., ADAMS, E. K., ZHOU, M., (2015). Estimating premium sensitivity for children's public health insurance coverage: selection but no death spiral, *Health services research*, 50(2), pp. 579–598.

MCDONALD, J. B., BUTLER, R. J. (1987). Some generalized mixture distributions with an application to unemployment duration, *The Review of Economics and Statistics*, pp. 232–240.

SCHIRMACHER, E., (2016). Pure Premium Modeling Using Generalized Linear Models, *Predictive Modeling Applications in Actuarial Science*: Volume 2, Case Studies in Insurance, 1.

SHI, P., VALDEZ, E. A., (2011). A copula approach to test asymmetric information with applications to predictive modeling, *Insurance: Mathematics and Economics*, 49(2), pp. 226–239.

SHI, P., (2016). Insurance ratemaking using a copula-based multivariate Tweedie model, *Scandinavian Actuarial Journal*, 2016(3), pp. 198–215.

SHI, P., YANG, L., (2018). Pair copula constructions for insurance experience rating, *Journal of the American Statistical Association*, 113(521), pp. 122–133.

SUN, J., FREES, E. W., ROSENBERG, M. A., (2008). Heavy-tailed longitudinal data modeling using copulas, *Insurance: Mathematics and Economics*, 42(2), pp. 817–830.

SKLAR, M., (1959). Fonctions de repartition an dimensions et leurs marges, *Publ. inst. statist. univ. Paris*, 8, pp. 229–231.

WEISBERG, H. I., TOMBERLIN, T. J., (1982). A statistical perspective on actuarial methods for estimating pure premiums from cross-classified data, *Journal of Risk and Insurance*, pp. 539–563.

WOLNY-DOMINIAK, A., WANAT, S., SOBIECKI, D., (2018). *Modelling Quantile Premium for Dependent LOBs in Property/Casualty Insurance*, In Finance and Sustainability, Springer, Cham, pp. 265–272.

YANG, Y., QIAN, W., ZOU, H., (2017). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models, *Journal of Business and Economic Statistics*, pp. 1–15.

ZAROUDI, S., BEHZADI, M. H., FARIDROHANI, M. R., (2018a). Application of Copula in Life Insurance, *International Journal of Applied Mathematics and Statistic*, 57(3), 162–168.

ZAROUDI, S., FARIDROHANI, M., BEHZADI, M., (2018b). A Copula Approach for Finding the Type of Dependency with Mortality Force Function in Insurance Market, *Journal of Advances and Applications in Statistics*, 53(2), pp. 103–121.

ZHANG, X., YIN, W., WANG, J., YE, T., ZHAO, J., (2015). Crop insurance premium ratemaking based on survey data: a case study from Dingxing county, China, *International Journal of Disaster Risk Science*, 6(3), pp. 207–215.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 13–34, DOI 10.21307/stattrans-2020-012
Submitted – 18.05.2018; accepted for publishing – 13.03.2020

Beta transmuted Lomax distribution with applications

Ahmed Hurairah[1], **Abdelhakim Alabid**[2]

ABSTRACT

In this paper we propose and test a composite generalizer of the Lomax distribution .The genesis of the beta distribution and transmuted map is used to develop the so-called beta transmuted Lomax (BTL) distribution. The properties of the distribution are discussed and explicit expressions are derived for the moments, mean deviations, quantiles, distribution of order statistics and reliability. The maximum likelihood method is used for estimating the model parameters, and the finite sample performance of the estimators is assessed by simulation. Finally, the authors demonstrate the usefulness of the new distribution in analysing positive data.

Key words: Lomax distribution, beta Lomax distribution, transmuted distribution, maximum likelihood estimation.

1. Introduction

Lomax (1954) proposed Pareto Type – II (the shifted Pareto) distribution, also known as Lomax distribution, and used it for the analysis of business failure lifetime data. The Lomax distribution is widely applicable in reliability and life testing problems in engineering as well as in survival analysis as an alternative distribution. After the work of Lomax (1954), various authors studied the Lomax distribution.

The cumulative distribution function of the Lomax arises as a limit distribution for the residual lifetime at a great age (Balkema and De Haan, 1974). Myhre and Saunders (1982) gave application of the Lomax distribution using the right censored data. Lingappaiah (1986) proposed various procedures of estimation for the Lomax distributions. Nayak (1987) proposed a multivariate Lomax distribution and discussed its various properties and usefulness in reliability theory. Ahsanullah (1991) and Balakrishnan and Ahsanullah (1994) investigated distributional properties and moments of record values from the Lomax distribution respectively. Vidondo et al. (1997) used this distribution for modelling size spectra data in aquatic ecology.

[1] Department of Statistics, Sana'a University, Yemen. E-mail: Hurairah69@yahoo.com.
[2] Department of Statistics, Sana'a University, Yemen. E-mail: hakimdeput@gmail.com.

Childs et al. (2001) gave order statistics from nonidentical right-truncated Lomax distributions and gave applications for these situations. The Lomax distribution was used to obtain a discrete Poisson Lomax distribution (Al-Awadhi and Ghitany, 2001). Bayesian method of estimation was used for the estimation of the Lomax survival function (Howlader and Hossain, 2002). Non-Bayesian and Byesian estimators of the sample size in the case of type I censored samples from the Lomax distribution were proposed by Abd-Elfattah et al. (2007). Ghitany et al. (2007) proposed the properties of a new parametric distribution, which was investigated by Marshall and Olkin (1997) and comprehensively extended the distributions family which is being applied to the model of the Lomax. Hassan and Al-Ghamdi (2009) used the Lomax distribution for the determination of optimal times of changing level of stress for simple stress plans under a cumulative exposure model. Abd-Elfattah and Alharbey (2010) estimated the parameters of the Lomax distribution based on generalized probability weighted moments. Abdul-Moniem and Abdel-Hameed (2012) studied exponentiated Lomax (EL), Abdullahi and Ieren (2018) introduced transmuted Exponential Lomax distribution (TEL), Ghitany et al. (2007) introduced Marshall-Olkin extended Lomax (MOEL), Bindu and Sangita (2015) studied double Lomax (DL).
The probability density function (pdf) and the cumulative distribution function (cdf) of a Lomax distribution is given by

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

$$G(x) = \left[1-(1+\gamma x)^{-\theta}\right], \;\; x > 0, \;\; \gamma, \theta > 0. \tag{2}$$

In this article we propose a new generalizer, which is obtained by the composition of the genesis of beta distribution and transmutation map. We will execute this generalizer to the Lomax distribution to develop the so-called beta transmuted Lomax distribution. This will be the beta generalizer of the transmuted Lomax (TL) distribution studied by Ashour and Eltehiwy (2013). Consider a baseline cumulative distribution function (cdf) $G(x)$ with corresponding probability density function (pdf) $g(x)$ and parameter vector Θ. Then, the cdf of the transmuted Lomax (TL) family of distributions (for $x > 0$) is

$$\mathrm{G(x)} = \left(1-(1+\gamma x)^{-\theta}\right) \;\; \left(1+\lambda\,(1+\gamma x)^{-\theta}\right). \tag{3}$$

The corresponding pdf of the transmuted Lomax distribution is given by

$$\mathrm{g(x)} = \frac{\gamma\theta}{(1+\gamma x)^{\theta+1}} \;\; \left(1-\lambda+2\lambda(1+\gamma x)^{-\theta}\right), \tag{4}$$

where $|\lambda| \leq 1$.

A class of generalized distributions F(x) has received considerable attention over the last few years, in particular, after the studies by Eugene, Lee, and Famoye (2002)

and Jones (2004). If G denotes the baseline cumulative distribution function (cdf) of a random variable, then the beta generalized distribution is defined as

$$F(x) = I_{G(x)}(a,b) = \frac{1}{B(a,b)} \int_0^{G(x)} t^{a-1} (1-t)^{b-1} dt \,, \tag{5}$$

where $a > 0$ and $b > 0$ are shape parameters. Note that $I_y(a,b) = \frac{B_y(a,b)}{B(a,b)}$ is the incomplete beta function ratio, and $B_y(a,b) = \int_0^y t^{a-1}(1-t)^{b-1} \ dt$ is the incomplete beta function, $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the beta function and $\Gamma(.)$ is the gamma function. The probability density function (pdf) of the Beta generated distribution has the form

$$\mathrm{f(x)} = \frac{\mathrm{g(x)}}{\mathrm{B}(a,\mathrm{b})} \ [\mathrm{G}(x)]^{\mathrm{a}-1}[1-\mathrm{G}(x)]^{\mathrm{b}-1}. \tag{6}$$

Based on the above generalization, Lemonte and Cordeiro (2013) investigated beta Lomax (BL), Gupta et al. (2016) introduced Lomax-Gumbel and studied expressions for its characteristic function, Terna and David (2018) introduced the new extension of the exponential distribution is called Lomax-Exponential distribution (LED), Kawsar et al. (2018) introduced Rayleigh Lomax distribution, Kumaraswamy Lomax (KwL) and McDonald Lomax (McL) and Cordeiro et al. (2013) introduced gamma Lomax (GL) distributions. Recently Tahir et al. (2015) introduced the Weibull Lomax (WL) distribution and studied its mathematical and statistical properties, Tahir et al. (2016) introduced the Gumbel Lomax (GL) distribution and Oguntunde et al. (2018) developed a new compound distributions which is Gompertz Lomax distribution.

The main aim of this paper is to define and study a new family of distributions by adding two extra shape parameters in (3) to provide more flexibility to the generated family. The additional advantage of the new distribution is that it has more parameters to have a better control. The rest of the paper is organized as follows. In Section 2 we define the BTL distribution and discuss some of its sub-models. In Section 3 we present the mixture representation of the BTL distribution. Section 4 discusses mathematical characteristics of the BTL distribution such as the moments, quantile, mean deviation, order statistics and stress-strength model. Estimation of parameters by the maximum likelihood method and the performance of the estimators is assessed by simulation in Section 5. In Section 6, the distribution is used for analysing real data. Finally, in Section 7, we make some concluding remarks on our study.

2. The beta transmuted Lomax distribution

In this section we provide the formulation of the beta transmuted Lomax (BTL) distribution. By inserting (3) into (5) the cumulative distribution function of the beta-transmuted Lomax distribution with five parameters is given by

$$F(x) = I_{\left(1-(1+\gamma x)^{-\theta}\right) \ \left(1+\lambda(1+\gamma x)^{-\theta}\right)} \quad (a,b)$$

$$= \frac{1}{B(a,b)} \int_0^{\left(1-(1+\gamma x)^{-\theta}\right)\ \left(1+\lambda(1+\gamma x)^{-\theta}\right)} \quad t^{\alpha-1}\,(1-t)^{b-1} dt, \qquad (7)$$

where $x > 0,\ \gamma, \theta > 0, |\lambda| \leq 1$ and $a > 0, b > 0$.

The cdf can be expressed in a closed form using the hypergeometric function (see Cordeiro and Nadarajah 2011) as follows:

$$F(x) = \frac{\left[\left(1-(1+\gamma x)^{-\theta}\right)\ \left(1+\lambda(1+\gamma x)^{-\theta}\right)\right]^a}{aB(a,b)} . 2F1\big(a, 1-b; a+1; \big(1 - (1+\gamma x)^{-\theta}\big)\ \big(1+\lambda(1+\gamma x)^{-\theta}\big)\big),$$

where $2F1(c,d;e;z) = \sum_{k=0}^{\infty} \frac{(c)_k\,(d)_k}{(e)_k} \frac{(z^k)}{k!}$ is the Gaussian hypergeometric function, where $(c)_k$ is the ascending factorial defined by (assuming that $(c)_0 = 1$)

$$(c)_k = \begin{cases} c(c+1)(c+2)\ldots.(c+k-1) & k = 1,2,3,\ldots \\ 1 & k = 0 \end{cases}$$

Differentiating (7) with respect to x, we get the probability density function of the BTL distribution given by

$$\mathrm{f(x)} = \frac{\gamma\theta}{\mathrm{B(a,b)}}(1+\gamma x)^{-\theta-1}\big(1-\lambda+2\lambda(1+\gamma x)^{-\theta}\big)\left[\big(1-(1+\gamma x)^{-\theta}\big)\right]^{\mathrm{a}-1}\left[\big(1+\lambda(1+\gamma x)^{-\theta}\big)\right]^{\mathrm{a}-1}\left[1-\big(1-(1+\gamma x)^{-\theta}\big)\big(1+\lambda(1+\gamma x)^{-\theta}\big)\right]^{\mathrm{b}-1}, \qquad (8)$$

where $x > 0,\ \gamma, \theta > 0, |\lambda| \leq 1$ and $a > 0, b > 0$.

The beta transmuted Lomax (BTL) distribution includes the following distributions as a special case:

- for $\lambda = 0$, beta transmuted Lomax reduces to beta Lomax distribution.
- For $\mathrm{a} = \mathrm{b} = 1$, beta transmuted Lomax reduces to transmuted Lomax distribution.
- For $\lambda = 0$ and $\mathrm{b} = 1$, beta transmuted Lomax reduces to exponentiated Lomax distribution.
- For $\mathrm{a} = \mathrm{b} = 1$ and $\lambda = 0$, beta transmuted Lomax reduces to Lomax distribution.

Figure 1 illustrates some of the possible shapes of the density function of the BTL distribution for selected values of the parameters θ, λ, a and b with $\gamma = 1$.

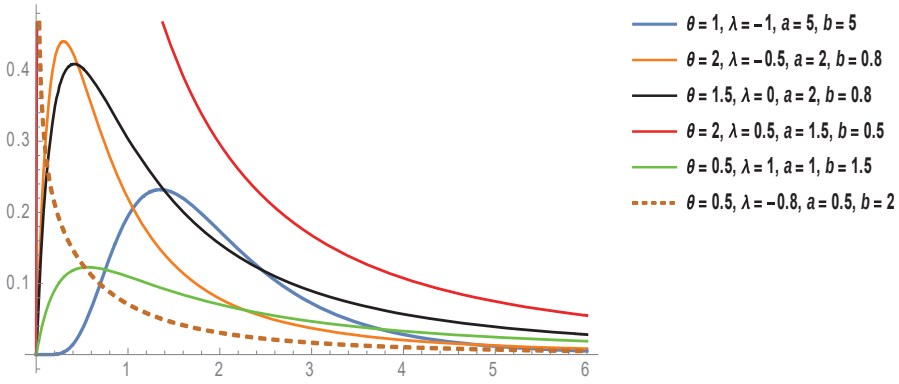


Figure 1. pdf of the BTL distribution for selected values of the parameters

The plot for PDF reveals that the BTL distribution is positively skewed and therefore will be a good model for positively skewed data sets.

Figure 2 illustrates some of the possible shapes of the cumulative distribution function of the BTL distribution for selected values of the parameters θ, λ, a and b with $\gamma = 1$.

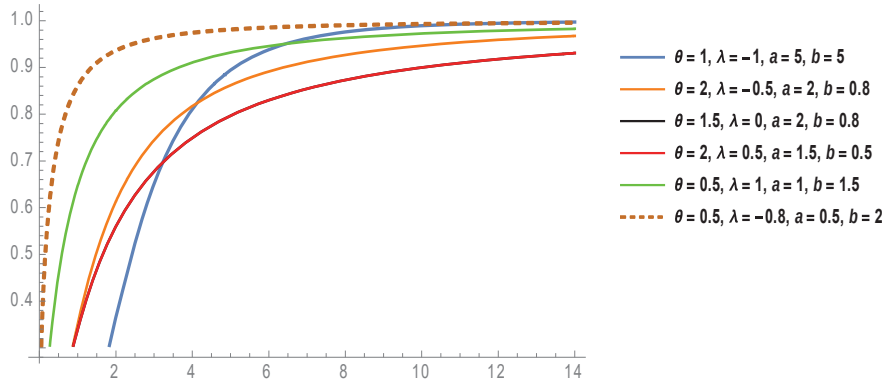


Figure 2. cdf of the BTL distribution for selected values of the parameters

The graphical representation of the cumulative function for different possible values of the parameters is shown in Figure 2, which is always an increasing function.

3. Mixture representation

In this section we find the series representations of cdf and pdf of the BTL distribution, which will be useful for studying its mathematical characteristics. As we shall see both pdf and cdf of BTL distribution can be expressed in terms of the Lomax distribution. By using (5) and the power series expansion of $(1-t)^{b-1}$, we get

$$F(x)=\frac{1}{B(a,b)}\int_0^{G(x)} t^{a-1}(1-t)^{b-1}dt=\frac{1}{B(a,b)}\sum_{j=0}^{\infty}(-1)^j\binom{b-1}{j}\frac{[G(x)]^{a+j}}{(a+j)}$$

with the binomial term $\binom{b-1}{j}=\frac{\Gamma(b)}{\Gamma(b-j)\,j!}$ defined for any real b. Hence, (7) reduces to

$$F(x)=\sum_{j=0}^{\infty}(-1)^j\binom{b-1}{j}\frac{\left[\left(1-(1+\gamma x)^{-\theta}\right)\left(1+\lambda\,(1+\gamma x)^{-\theta}\right)\right]^{a+j}}{B(a,b)(a+j)}. \tag{9}$$

Again, using the binomial expansion of $\left[\left(1-(1+\gamma x)^{-\theta}\right)\left(1+\lambda\,(1+\gamma x)^{-\theta}\right)\right]^{a+j}$, we have

$$\begin{aligned}F(x)&=\sum_{j,k,l=0}^{\infty}(-1)^{j+k}\binom{b-1}{j}\binom{a+j}{k}\binom{a+j}{l}\lambda^l\frac{(1+\gamma x)^{-\theta(k+l)}}{B(a,b)(a+j)}\\&=\sum_{j,k,l=0}^{\infty}(-1)^{j+k}\binom{b-1}{j}\binom{a+j}{k}\binom{a+j}{l}\lambda^l\frac{\left(1-G_1(x;\gamma,\theta(k+l))\right)}{B(a,b)(a+j)},\end{aligned} \tag{10}$$

where $G_1\left(x;\gamma,\theta(k+l)\right)$ is the Lomax cdf with scale γ and shape $\theta(k+l)$ parameter. Differentiating (10) with respect to x gives a useful expansion of $f(x)$ as

$$f(x)=\sum_{k,l=0}^{\infty}w_{kl}\;g\left(x;\gamma,\theta(k+l)\right),\qquad x>0, \tag{11}$$

where

$$w_{kl}=\sum_{j=0}^{\infty}(-1)^{j+k+l}\binom{b-1}{j}\binom{a+j}{k}\binom{a+j}{l}\frac{\lambda^l}{B(a,b)(a+j)}$$

and $g\left(x;\gamma,\theta(k+l)\right)$ is the Lomax pdf with scale γ and shape $\theta(k+l)$ parameters. If $b>0$ is an integer, the index j in the sum stops at $b-1$, and if a is an integer, then the indices k and l in the sum stop at $a+j$.

Thus, several mathematical properties of the BTL distribution can be obtained simply from those properties of the exp-G family. Equations (10) and (11) are the main result of this section.

4. Mathematical characteristics

In this section we provide some mathematical properties of the BTL distribution including the moments and moment generating function, quantiles, mean deviations, order statistics and stress-strength model.

4.1. Moments and moments generating

Moments are necessary and important in any statistical analysis, especially in applications. They can be used to study the most important features and characteristics of a distribution (e.g. tendency, dispersion, skewness and kurtosis). Using the mixture representation described in Section 3, the r-*th* moment of the BTL random variable X is given by

$$E(X^r) = \int_{-\infty}^{\infty} x^r \, \mathrm{f}(x)\mathrm{dx} = \int_0^{\infty} x^r \sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \; f\big(x;\gamma,\theta(k+l)\big)\,\mathrm{dx}$$

$$= \int_0^{\infty} \sum_{k,l=0}^{\infty} \mathrm{w}_{kl}\, x^r \;\; \gamma\theta\,(k+l)(1+\gamma x)^{-\theta(k+l)-1}\mathrm{dx}$$

$$= \gamma^{-r} \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \frac{\Gamma[1+r]\Gamma[-r+(k+l)\theta]}{\Gamma[(k+l)\theta]}, \;\; (k+l)\theta > r, \tag{12}$$

$$E(X) = \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \left[\frac{1}{\gamma[-1+(k+l)\theta]}\right], \;\; (k+l)\theta > 1, \tag{13}$$

$$E(X^2) = \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \left[\frac{2(k+l)\theta\Gamma[-2+(k+l)\theta]}{\gamma^2[1+(k+l)\theta]}\right], \;\; (k+l)\theta > 2,$$

$$E(X^3) = \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \left[\frac{6(k+l)\theta\Gamma[-3+(k+l)\theta]}{\gamma^3[1+(k+l)\theta]}\right], \;\; (k+l)\theta > 3,$$

$$E(X^4) = \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \left[\frac{24(k+l)\theta\Gamma[-4+(k+l)\theta]}{\gamma^4[1+(k+l)\theta]}\right], (k+l)\theta > 4.$$

The variance, skewness and kurtosis measures can now be calculated using the relations

$$Var(x) = E(x^2) - [E(x)]^2 = \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \left(\frac{(k+l)\theta}{\gamma^2(-2+(k+l)\theta)(-1+(k+l)\theta)^2}\right), \tag{14}$$

$$Skewness(x) = \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \left[\frac{2+2(k+l)\theta}{\gamma(-3(k+l)\theta)\sqrt{\frac{(k+l)\theta}{\gamma^2(-2+(k+l)\theta)(-1+(k+l)\theta)^2}}\;(-1+(k+l)\theta)}\right], \tag{15}$$

$$Kurtosis(x) = \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \left[9 - \frac{1}{(k+l)\theta} + \frac{81}{-4+(k+l)\theta} - \frac{32}{-3+(k+l)\theta}\right]. \tag{16}$$

Similarly, the moment generating function of X may be obtained as below:

$$M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \sum_{k,l=0}^{\infty} w_{kl} \left(e^{-\frac{t}{\gamma}} (k+l)\theta\, E_{\theta(k+l)+1}\left(-\frac{t}{\gamma}\right)\right), \tag{17}$$

where $E_n(z) = \int_1^{\infty} \frac{e^{-tz}}{t^n}\, dt$.

4.2. Quantiles

Quantiles are the points in a distribution that relate to the rank order of values. The quantile function of a distribution is the real solution of $F(x_q) = q$ for $0 \le q \le 1$. The quantiles of beta transmuted Lomax distribution are obtained from (7) as

$$X = \frac{\left[\frac{\lambda - 1 + \sqrt{(1+\lambda)^2 - 4\lambda\left(I_q^{-1}(a,b)\right)}}{2\lambda}\right]^{-\frac{1}{\theta}} - 1}{\gamma} \tag{18}$$

where $I_q^{-1}(a,b)$ is the inverse of the incomplete beta function with parameters a and b. The following expansion for the inverse of the beta incomplete function $I_q^{-1}(a,b)$ can be found on the Wolfram website http://functions.wolfram.com/06.23.06.0004.01

$$I_u^{-1}(a,b) = w + \frac{b-1}{a+1} w^2 + \frac{(b-1)(a^2+3ab-a+5b-4)}{2(a+1)^2(a+2)} w^3 + \frac{(b-1)[a^4+(6b-1)a^3+(b+2)(8b-5)a^2]}{2(a+1)^2(a+2)} w^4 + \frac{(b-1)[(33a^2-30b+4)a+b(31a-47)+18]}{3(a+1)^3(a+2)(a+3)} w^4 + O\left(P^{\frac{5}{a}}\right),$$

where $w = \{aB(a,b)q\}^{\frac{1}{a}}, a > 0$.

4.3. Mean deviation

The amount of scatter in a population is evidently measured to some extent by the totality of deviations from the mean and median. If X has a BTL distribution, then we can derive the mean deviations about the mean $\mu = E(X)$ and about the median M as

$$\delta_1(x) = \int_0^{\infty} |x - \mu| f(x) dx,$$

and

$$\delta_2(x) = \int_0^{\infty} |x - M| f(x) dx.$$

The mean of the distribution is obtained from (12), and the median is obtained by solving the equation

$$I_{(1-(1+\gamma x)^{-\theta})\left(1+\lambda-\lambda\left(1-(1+\gamma x)^{-\theta}\right)\right)}(a,b) = \frac{1}{2}.$$

These measures can be calculated using the relationships that

$$\delta_1(x) = \int_0^{\mu} (\mu - x) f(x) dx + \int_{\mu}^{\infty} (x - \mu) f(x) dx$$

$$= 2 \int_0^{\mu} (\mu - x) f(x) dx$$

$$= 2\{\mu F(\mu) - \textstyle\int_0^{\mu} x f(x) dx\}$$

$$\delta_1(x) = 2\{\mu F(\mu) - J(\mu)\},$$

and

$$\delta_2(x) = \mu - 2\, J(\mu),$$

where $J(t) = \int_0^t x f(x) dx$. From (11) we have

$$J(t) = \sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \int_0^t (k+l)\theta\gamma \; x\,(1+\gamma x)^{-\theta(k+l)-1}\, dx$$

$$= \textstyle\sum_{k,l=0}^{\infty} \mathrm{w}_{kl} \left(\frac{(1+\gamma t)^{-\theta(k+l)}\left(-1+(1+\gamma t)^{-\theta(k+l)}-(k+l)t\gamma\theta\right)}{\gamma(-1+(k+l)\theta)} \right). \qquad (19)$$

Using (10), one can easily find $\delta_1(x)$ and $\delta_2(x)$.

4.4. Order statistics

Let $X_1, X_2, X_3, \ldots \ldots, X_n$ be a simple random sample from the BTL distribution with cumulative distribution function (7) and probability density function (8).

Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the order statistics from this sample. The pdf $f_{i:n}(x)$ of *i*-th order statistics $X_{(i)}$ is given by

$$f_{i:n}(x) = \frac{1}{B(i, n-i+1)} f(x)[F(x)]^{i-1}[1-F(x)]^{n-i}, \quad i = 1,2, \ldots, n$$

and cdf is given by

$$F_{i:n}(x) = \sum_{k=i}^{n} \binom{n}{k} [F(x)]^k [1-F(x)]^{n-k}$$

$$= \textstyle\int_0^{F(x)} \frac{1}{B(i,n-i+1)}\; t^{i-1}[1-t]^{n-i} dt.$$

The pdf of the j^{th} order statistic for the beta transmuted Lomax distribution is given by

$$f_{i:n}(x) = \frac{1}{B(i, n-i+1)}\, f(x) \sum_{s=0}^{n-i} (-1)^s \binom{n-i}{s} [F(x)]^{i+s-1}$$

$$f_{i:n}(x) = \frac{\theta\gamma}{\mathrm{B(i,n-i+1)}}\left(\sum_{\mathrm{k,l=0}}^{\infty} \mathrm{w_{kl}}\,(k+l)(1+\gamma x)^{-\theta(k+l)-1}\right)$$
$$\sum_{\mathrm{s=0}}^{\mathrm{n-i}}(-1)^{\mathrm{s+1}}\binom{\mathrm{n-i}}{\mathrm{s}}\left[\sum_{\mathrm{k,l=0}}^{\infty}\mathrm{w_{kl}}\left(1-(1+\gamma x)^{-\theta(k+l)}\right)\right]^{\mathrm{i+s-1}}. \quad (20)$$

Writing $= (1+\gamma x)^{-\theta}$, $f_{i:n}(x)$ can be expressed as

$$f_{i:n}(x) = \frac{\theta\gamma}{\mathrm{B(i,n-i+1)}}\left(\sum_{\mathrm{k,l=0}}^{\infty}\mathrm{w_{kl}}\ (\mathrm{k+l})\,\mathrm{u}^{(\mathrm{k+l})-1}\right)\sum_{s=0}^{n-i}(-1)^{s+1}\binom{n-i}{s}\left[\sum_{k,l=0}^{\infty}\mathrm{w}_{kl}\ u^{(k+l)}\right]^{i+s-1}. \quad (21)$$

We note that in (21) we can write

$$\sum_{k,l=0}^{\infty}\mathrm{w}_{kl}\,u^{(k+l)} = \sum_{m=0}^{\infty} w_m^*\,u^m$$

and

$$\sum_{k,l=0}^{\infty}\mathrm{w}_{kl}\ (k+l)\,u^{(k+l)} = \sum_{m=0}^{\infty} m\,w_m^*\,u^m,$$

where $w_m^* \sum_{k,l:k+l=m}\mathrm{w}_{kl}$. Further, from Gradshteyn and Ryzhik (2000), for any positive integer r,

$$\left(\sum_{k=0}^{\infty}\mathrm{a}_k\,u^k\right)^r = \sum_{k=0}^{\infty} d_{r,k}\,u^k, \quad (22)$$

where the coefficients $d_{r,k}$, for $k = 1,2,..,$ can be determined from the recurrence equation

$$d_{r,k} = \left(k_{a_0}\right)^{-1}\sum_{m=1}^{k}[m(r+1)-k]\,\mathrm{a}_m d_{r,k-m} \quad (23)$$

and $d_{r,0} = a_0^r$. Hence, $d_{r,k}$ comes directly from $d_{r,0},\ldots,d_{r,k-1}$ and, therefore, from $\mathrm{a}_0,\ldots,\mathrm{a}_k$. Using (22) and (23) it follows that

$$f_{i:n}(x)\,\frac{\theta\gamma}{B(i,n-i+1)}\left(\sum_{m=0}^{\infty} m\,w_m^*\,u^m\right)\sum_{s=0}^{n-i}(-1)^{s+1}\binom{n-i}{s}\left(\sum_{m=0}^{\infty}\mathrm{d}_{i+s-1,m}u^m\right),$$

where

$$\mathrm{d}_{i+s-1,m} = (m\,w_0^*)^{-1}\sum_{q=1}^{k}[q(i+s)-m]w_m^*\mathrm{d}_{i+s-1,m-q},$$

$$\mathrm{d}_{i+s-1,0} = (w_0^*)^{i+s-1} = \left(\sum_{j=0}^{\infty}(-1)^{j+1}\binom{b-1}{j}\frac{1}{B(a,b)(a+j)}\right)^{i+s-1}.$$

Combining terms, we obtain

$$f_{i:n}(x) = \frac{\theta\gamma}{B(i,n-i+1)}\sum_{s=0}^{n-i}(-1)^{s+1}\binom{n-i}{s}\sum_{m=l}^{\infty}\sum_{t=0}^{\infty} m\mathrm{d}_{i+s-1,t}w^{*}t^{w^{*}}m^{u^{m+t}}$$

$$= \frac{1}{B(i,n-i+1)}\sum_{s=0}^{n-i}(-1)^{s+1}\binom{n-i}{s}\sum_{m=l}^{\infty}\sum_{t=0}^{\infty}\frac{m\mathrm{d}_{i+s-1,t}w_m^*}{m+t}\left[\theta\gamma(m+t)\,(1+\gamma x)^{-\theta(m+t)-1}\right]$$
$$= \sum_{m=l}^{\infty}\sum_{t=0}^{\infty}c_i(m,t)\;g\big(x{:}\,\gamma,\theta(m+t)\big), \tag{24}$$

where $g\big(x{:}\,\gamma,\theta(m+t)\big)$ denotes the pdf of a Lomax distribution with parameter $(m+t)\theta$ and γ parameter and

$$c_i(m,t) = \frac{1}{B(i,n-i+1)}\frac{m\,w_m^*}{m+t}\sum_{s=0}^{n-i}(-1)^{s+1}\binom{n-i}{s}\mathrm{d}_{i+s-1,t}. \tag{25}$$

4.5. Stress-strength model

A stress-strength model describes the life of a component which has a random strength X_1 and is subjected to a random stress X_2. The component functions satisfactorily as long as $X_1 > X_2$, and fails when $X_1 < X_2$. The probability $R = Pr(X_1 > X_2)$ defines the component reliability. Stress-strength models have many applications especially in engineering concepts such as structures, deterioration of rocket motors, static fatigue of ceramic components, fatigue failure of aircraft structures and the aging of concrete pressure vessels.

Consider X_1 and X_2 to be independently distributed, with $X_1 \sim BTL(\gamma,\theta_1,\lambda_1,\mathrm{a}_1,\mathrm{b}_1)$ and $X_2 \sim BTL(\gamma,\theta_2,\lambda_2,\mathrm{a}_2,\mathrm{b}_2)$. The cdf F_1 of X_1 and the pdf f_2 of X_2 are obtained from (10) and (11), respectively. Then,

$$R = Pr(X_1 > X_2) = \int_0^{\infty} f_2(y)[1-F_1(y)]dy$$
$$= 1 + \sum_{k,l=0}^{\infty} w_{kl}^{(1)} \int_0^{\infty} f_2(y)(1+\gamma y)^{-\theta(k+l)}dy$$
$$= \sum_{k,l=0}^{\infty} w_{kl}^{(1)}\, A(k,l),$$

where

$$w_{kl}^{(i)} = \sum_{j=0}^{\infty}(-1)^{j+k+l}\binom{b_i-1}{j}\binom{a_i+j}{k}\binom{a_i+j}{l}\frac{\lambda^l}{B(a,b)(a_i+j)}, \qquad i = 1,2,$$

and

$$A(k,l) = \int_0^{\infty} f_2(y)(1+\gamma y)^{-\theta(k+l)}dy.$$

Now,

$$A(k,l) = \sum_{r,s=0}^{\infty} w_{rs}^{(2)} \int_0^{\infty} (\mathrm{r}+\mathrm{s})\,\gamma\,\theta_2\left[(1+\gamma\, y)^{-[\theta_2(r+s)+\theta_2(k+l)]-1}\right]\mathrm{dy}$$
$$= \sum_{r,s=0}^{\infty} w_{rs}^{(2)}\,\frac{\gamma\,\theta_2(\mathrm{r}+\mathrm{s})}{\gamma\,\theta_1(\mathrm{k}+\mathrm{l})+\gamma\,\theta_2(\mathrm{r}+\mathrm{s})}.$$

Hence,

$$R = 1 + \sum_{k,l=0}^{\infty} w_{kl}^{(1)} \sum_{r,s=0}^{\infty} w_{rs}^{(2)}\,\frac{\gamma\,\theta_2(\mathrm{r}+\mathrm{s})}{\gamma\,\theta_1(\mathrm{k}+\mathrm{l})+\gamma\,\theta_2(\mathrm{r}+\mathrm{s})}$$

$$= 1 + \sum_{k=0}^{\infty} \sum_{r=0}^{\infty} w_k^{*(1)}\, w_r^{*(2)} \frac{r\gamma\ \theta_2}{k\gamma\ \theta_1 + r\gamma\ \theta_2}, \tag{26}$$

where

$$w_m^{*(i)} = \sum_{k,l:k+l=m} w_{kl}^{*(i)}, \quad i = 1,2.$$

5. Maximum likelihood estimation

Let $x_1, x_2, \dots, x_n$ be a random sample from the beta transmuted Lomax distribution with observed values $x_1, x_2, \dots, x_n$ and $\Theta = (\gamma, \theta, \lambda, a, b)$ be the parameter vector. The likelihood function $L(\Theta)$ is given by

$$L(\Theta) = \left(\frac{\gamma\theta}{B(a,b)}\right)^n \prod_{i=1}^{n} (1+\gamma x_i)^{-\theta-1} \left(1+\lambda-2\lambda(1+\gamma x_i)^{-\theta}\right)\left[\left(1-(1+\gamma x_i)^{-\theta}\right)\right]^{a-1}\left[\left(1+\lambda(1+\gamma x_i)^{-\theta}\right)\right]^{a-1}\left[1-\left(1-(1+\gamma x_i)^{-\theta}\right)\left(1+\lambda(1+\gamma x_i)^{-\theta}\right)\right]^{b-1}. \tag{27}$$

Then, the log-likelihood function $l(\Theta)$for the vector of parameters $\Theta = (\gamma, \theta, \lambda, a, b)$, is

$$l(\Theta) = n\log\gamma + n\log\theta - n\log[B(a,b)] + (-\theta-1)\sum_{i=1}^{n} log(1+\gamma x_i) + \sum_{i=1}^{n} Log\left[1+\lambda-2\lambda(1+\gamma x_i)^{-\theta}\right] + (a-1)\left[\sum_{i=1}^{n} Log\left(1-(1+\gamma x_i)^{-\theta}\right) + (a-1)\sum_{i=1}^{n} Log\left(1+\lambda(1+\gamma x_i)^{-\theta}\right)\right] + (b-1)\sum_{i=1}^{n} Log\left[1-\left(1-(1+\gamma x_i)^{-\theta}\right)\left(1+\lambda(1+\gamma x_i)^{-\theta}\right)\right]. \tag{28}$$

We differentiate (28) with respect to $\gamma, \theta, \lambda, a$ and b respectively to obtain the elements of score vector $\frac{\partial l(\Theta)}{\partial \Theta} = \left(\frac{\partial l(\Theta)}{\partial \gamma}, \frac{\partial l(\Theta)}{\partial \theta}, \frac{\partial l(\Theta)}{\partial \lambda}, \frac{\partial l(\Theta)}{\partial a}, \frac{\partial l(\Theta)}{\partial b}\right)^t$ as below

$$\frac{\partial l(\Theta)}{\partial \gamma} = \frac{n}{\gamma} - (\theta+1)\sum_{i=1}^{n} \frac{x_i}{1+\gamma x_i} + (a-1)\theta\sum_{i=1}^{n} \frac{x_i(1+\gamma x_i)^{-\theta-1}}{1-(1+\gamma x_i)^{-\theta}} + 2\theta\lambda\sum_{i=1}^{n} \frac{x_i(1+\gamma x_i)^{-\theta-1}}{(1+\lambda-2\lambda(1+\gamma x_i)^{-\theta})} - (a-1)\theta\lambda\sum_{i=1}^{n} \frac{x_i(1+\gamma x_i)^{-\theta-1}}{(1+\lambda(1+\gamma x_i)^{-\theta})} + (b-1)\theta\lambda\sum_{i=1}^{n} \frac{x_i(1+\gamma x_i)^{-\theta-1}(1-(1+\gamma x_i)^{-\theta})}{1-(1-(1+\gamma x_i)^{-\theta})(1+\lambda(1+\gamma x_i)^{-\theta})} - (b-1)\theta\sum_{i=1}^{n} \frac{x_i(1+\gamma x_i)^{-\theta-1}(1+\lambda(1+\gamma x_i)^{-\theta})}{1-(1-(1+\gamma x_i)^{-\theta})(1+\lambda(1+\gamma x_i)^{-\theta})} \tag{29}$$

$$\frac{\partial l(\Theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \log(1+\gamma x_i) + (a-1)\sum_{i=1}^{n} \frac{(1+\gamma x_i)^{-\theta} log(1+\gamma x_i)}{1-(1+\gamma x_i)^{-\theta}} + 2\lambda\sum_{i=1}^{n} \frac{(1+\gamma x_i)^{-\theta} log(1+\gamma x_i)}{1+\lambda-2\lambda(1+\gamma x_i)^{-\theta}} - \lambda(a-1)\sum_{i=1}^{n} \frac{(1+\gamma x_i)^{-\theta} log(1+\gamma x_i)}{1+\lambda(1+\gamma x_i)^{-\theta}} + (b-1)\sum_{i=1}^{n} \frac{\lambda(1-(1+\gamma x_i)^{-\theta})(1+\gamma x_i)^{-\theta} log(1+\gamma x_i)}{1-(1-(1+\gamma x_i)^{-\theta})(1+\lambda(1+\gamma x_i)^{-\theta})} - (b-1)\sum_{i=1}^{n} \frac{(1+\lambda(1+\gamma x_i)^{-\theta})(1+\gamma x_i)^{-\theta} log(1+\gamma x_i)}{1-(1-(1+\gamma x_i)^{-\theta})(1+\lambda(1+\gamma x_i)^{-\theta})} \tag{30}$$

$$\frac{\partial \mathrm{l}(\Theta)}{\partial \lambda} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{1-2(1+\gamma \mathrm{x_i})^{-\theta}}{1+\lambda-2\lambda(1+\gamma \mathrm{x_i})^{-\theta}} + (\mathrm{a}-1)\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{(1+\gamma \mathrm{x_i})^{-\theta}}{1+\lambda(1+\gamma \mathrm{x_i})^{-\theta}} - (\mathrm{b}-1)\sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\left(1-(1+\gamma \mathrm{x_i})^{-\theta}\right)(1+\gamma \mathrm{x_i})^{-\theta}}{1-\left(1-(1+\gamma \mathrm{x_i})^{-\theta}\right)\left(1+\lambda(1+\gamma \mathrm{x_i})^{-\theta}\right)} \tag{31}$$

$$\frac{\partial \mathrm{l}(\Theta)}{\partial \mathrm{a}} = -\mathrm{n}[\Psi(\mathrm{a}) - \Psi(\mathrm{a}+\mathrm{b})] + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Log}\left(1-(1+\gamma \mathrm{x_i})^{-\theta}\right) + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Log}\left(1+\lambda(1+\gamma \mathrm{x_i})^{-\theta}\right) \tag{32}$$

$$\frac{\partial \mathrm{l}(\Theta)}{\partial \mathrm{b}} = -\mathrm{n}[\Psi(\mathrm{b}) - \Psi(\mathrm{a}+\mathrm{b})] + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Log}\left[1-\left(1-(1+\gamma \mathrm{x_i})^{-\theta}\right)\left(1+\lambda(1+\gamma \mathrm{x_i})^{-\theta}\right)\right], \tag{33}$$

where $\Psi(x)$ is the digamma function defined by $\Psi(x) = \frac{d \log \Gamma(x)}{dx}$, and $\Gamma(x)$ is the Gamma function.

For a random sample $(x_1, x_2, \ldots, x_n)$ of size n from X, distributed with pdf (8), the sample log-likelihood is $l(\Theta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} l_i(\Theta)$, where $l_i(\Theta)$ is the log-likelihood for the ith observation ($\mathrm{i} = 1,2, \ldots, \mathrm{n}$), and the score vector is

$$\frac{\partial l(\Theta)}{\partial \Theta} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\partial l_i(\Theta)}{\partial \Theta}.$$

The maximum likelihood estimate (MLE) $\widehat{\Theta}$ of Θ is obtained by solving the system

$$\frac{\partial l(\Theta)}{\partial \Theta} = 0.$$

Under certain regularity conditions, $\sqrt{\mathrm{n}}\left(\widehat{\Theta} - \Theta\right) \xrightarrow{d} \mathrm{N}(0, \mathrm{I}(\Theta)^{-1})$, (here $\xrightarrow{d}$ stands for convergence in distribution), where $\mathrm{I}(\Theta)$ denotes the information matrix given by

$$\mathrm{I}(\Theta) = E\left(\frac{\partial^2 l(\Theta)}{\partial \Theta \partial \Theta'}\right).$$

This information matrix $\mathrm{I}(\Theta)$ may be approximated by the observed information matrix

$$\mathrm{I}\left(\widehat{\Theta}\right) = E\left(\frac{\partial^2 l(\Theta)}{\partial \Theta \partial \Theta'}\right)\Big|_{\Theta=\widehat{\Theta}}.$$

Then, using the approximation $\sqrt{\mathrm{n}}\left(\widehat{\Theta} - \Theta\right) \sim \mathrm{N}\left(0,\ \mathrm{I}^{-1}\left(\widehat{\Theta}\right)\right)$, one can carry out tests and find confidence regions for functions of some or all parameters in Θ.

5.1. Simulation study

Here, we evaluate the performance of MLEs for the beta transmuted Lomax distribution. The assessments were based on simulation studies.

The assessment of the finite sample behaviour of MLEs for this distribution was based on the following:

1. use the inversion method to generate one thousand samples of size n from the BTL distribution, i.e. generate values of

$$X = \frac{\left[\frac{\lambda-1+\sqrt{(1+\lambda)^2-4\lambda\left(I_q^{-1}(a,b)\right)}}{2\lambda}\right]^{-\frac{1}{\theta}} - 1}{\gamma}$$

2. compute MLEs for one thousand samples, say $\left(\hat{\gamma}_i, \hat{\theta}_i, \hat{\lambda}_i, \hat{a}_i, \hat{b}_i\right)$ for $i = 1,2,\dots,1000$.

3. compute the standard errors of MLEs for one thousand samples, say $\left(S_{\hat{\gamma}_i}, S_{\hat{\theta}_i}, S_{\hat{\lambda}_i}, S_{\hat{a}_i}, S_{\hat{b}_i}\right)$ for $i = 1,2,\dots,1000$. The standard errors were computed by inverting the observed information matrix.

4. compute the biases and mean squared errors by

$$Bias\left(\widehat{\Theta}\right) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\widehat{\Theta}_i - \Theta\right)$$

$$MSE\left(\widehat{\Theta}\right) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\widehat{\Theta}_i - \Theta\right)^2$$

For $\widehat{\Theta}_i = \left(\hat{\gamma}_i, \hat{\theta}_i, \hat{\lambda}_i, \hat{a}_i, \hat{b}_i\right)$ and $\Theta = (\ \gamma, \theta, \lambda, a, b\ )$

5. We repeated these steps for $n = \{10,50,80,100,150,200,300\}$ with $\gamma = 0.9, \theta = 0.8,\ \lambda = 0.7, a = 0.6\ and\ b = 0.5$, so computing $Bias\left(\widehat{\Theta}\right)$, $SE\left(\widehat{\Theta}\right)$ and $MSE\left(\widehat{\Theta}\right)$ for $\Theta = (\gamma, \theta, \lambda, a, b\ )$ and for $n = \{10,50,80,100,150,200,300\}$.

From Table 1 it is observed that as the sample size increases, the average biases, the standard error and the means squared errors decrease. This verifies the consistency properties of the estimates.

Table 1. Estimated parameters, Bias, standard error, and MSE of the BTL distribution

n	$\widehat{\Theta}$	Bias	S.E	MSE
10	$\hat{\gamma} = 0.872032$	0.7720320	0.3391303	0.5960335
	$\hat{\theta} = 0.096215$	0.1037850	0.0089731	0.0107714
	$\hat{\lambda} = 0.721075$	0.4210746	1.9199212	0.1773038
	$\hat{a} = 0.725406$	0.3254056	0.0615352	0.1058888
	$\hat{b} = 0.670449$	0.1704485	0.0747316	0.0290527

Table 1. Estimated parameters, Bias, standard error, and MSE of the BTL distribution (cont.)

n	$\widehat{\Theta}$	Bias	S.E	MSE
50	$\hat{\gamma} = 0.861056$	0.7610562	0.0725915	0.5792066
	$\hat{\theta} = 0.153349$	0.0496510	0.0004640	0.0021763
	$\hat{\lambda} = 0.703408$	0.4034083	0.8194007	0.1627383
	$\hat{a} = 0.705059$	0.3050595	0.0126729	0.0940613
	$\hat{b} = 0.634201$	0.1342013	0.0124783	0.0180100
80	$\hat{\gamma} = 0.860920$	0.7609200	0.0482477	0.5789992
	$\hat{\theta} = 0.150779$	0.0494220	0.0002847	0.0021226
	$\hat{\lambda} = 0.702135$	0.4021352	0.7258839	0.1617127
	$\hat{a} = 0.705481$	0.3054812	0.0063866	0.0933187
	$\hat{b} = 0.634298$	0.1342981	0.0071282	0.0180036
100	$\hat{\gamma} = 0.860974$	0.7607430	0.0394231	0.5770819
	$\hat{\theta} = 0.151489$	0.0485110	0.0002302	0.0020533
	$\hat{\lambda} = 0.701706$	0.4017062	0.6649546	0.1613679
	$\hat{a} = 0.704974$	0.3049736	0.0045990	0.0930089
	$\hat{b} = 0.633523$	0.1335229	0.0054569	0.0178284
150	$\hat{\gamma} = 0.861044$	0.7604350	0.0273712	0.5761872
	$\hat{\theta} = 0.152401$	0.0475990	0.0001553	0.0016656
	$\hat{\lambda} = 0.701139$	0.4011389	0.5515348	0.1609124
	$\hat{a} = 0.704322$	0.3043215	0.0024106	0.0926116
	$\hat{b} = 0.632522$	0.1325220	0.0033301	0.0175621
200	$\hat{\gamma} = 0.861077$	0.7603690	0.0211253	0.5742380
	$\hat{\theta} = 0.152844$	0.0471560	0.0001171	0.0022237
	$\hat{\lambda} = 0.700853$	0.4008572	0.4719037	0.1606865
	$\hat{a} = 0.704005$	0.3040050	0.0014553	0.0924192
	$\hat{b} = 0.632034$	0.1320344	0.0023345	0.0174331
300	$\hat{\gamma} = 0.860907$	0.7601090	0.0146216	0.5712876
	$\hat{\theta} = 0.149021$	0.0467220	0.0000784	0.0021083
	$\hat{\lambda} = 0.708791$	0.4005767	0.3665787	0.1604617
	$\hat{a} = 0.709832$	0.3036949	0.0006557	0.0922306
	$\hat{b} = 0.641948$	0.1315549	0.0014122	0.0173067

6. Applications

In this section, we provide applications to three real data sets to illustrate the importance and potentiality of the BTL distribution and some of the models generated from Lomax distributions, namely the Lomax (L), transmuted Lomax (TL), beta Lomax (BL), Gamma Lomax (GaL), Marshall-Olkin Lomax (MOL) and Weibull Lomax (WL) distributions.

Data Set I: The first real data set (Ghitany et al. 2008) consists of 100 observations on waiting time (in minutes) before the customer received service in a bank. The data are: 0.8, 0.8, 1.3, 1.5, 1.8, 1.9, 1.9, 2.1, 2.6, 2.7, 2.9, 3.1, 3.2, 3.3, 3.5, 3.6, 4, 4.1, 4.2, 4.2, 4.3,

4.3, 4.4, 4.4, 4.6, 4.7, 4.7, 4.8, 4.9, 4.9, 5.0, 5.3, 5.5, 5.7, 5.7, 6.1, 6.2, 6.2, 6.2, 6.3, 6.7, 6.9, 7.1, 7.1, 7.1, 7.1, 7.4, 7.6, 7.7, 8, 8.2, 8.6, 8.6, 8.6, 8.8, 8.8, 8.9, 8.9, 9.5, 9.6, 9.7, 9.8, 10.7, 10.9, 11.0, 11.0, 11.1, 11.2, 11.2, 11.5, 11.9, 12.4, 12.5, 2.9, 13.0, 13.1, 13.3, 13.6, 13.7, 13.9, 14.1, 15.4, 15.4, 17.3, 17.3, 18.1, 18.2, 18.4, 18.9, 19.0, 19.9, 20.6, 21.3, 21.4, 21.9, 23, 27, 31.6, 33.1, 38.5. The summary of the data set is provided as follows:

Table 2. Summary Statistics for Data Set I

Min.	Q_1	Median	Q_3	Mean	Max.	Var.	Skewness	Kurtosis
0.80	4.65	8.1	13.05	9.877	38.5	52.3741	1.47277	5.54029

Data Set II: The second data set (Gross and Clark (1975), page 105) on the relief times of twenty patients receiving an analgesic is: 1.1, 1.4, 1.3, 1.7, 1.9, 1.8, 1.6, 2.2, 1.7, 2.7, 4.1, 1.8, 1.5, 1.2, 1.4, 3, 1.7, 2.3, 1.6, 2. The summary of the data set is provided as follows:

Table 3. Summary Statistics for the Data Set II

Min.	Q_1	Median	Q_3	Mean	Max.	Var.	Skewness	Kurtosis.
1.1	1.45	1.7	2.1	1.9	38.5	0.4958	1.7198	5.9241

Data Set III : Here, we consider an uncensored data set corresponding to remission times (in months) of a random sample of 128 bladder cancer patients. These data were previously studied by Lee and Wang (2003) and Lemonte and Cordeiro (2011). Bladder cancer is a disease in which abnormal cells multiply without control in the bladder. The most common type of bladder cancer recapitulates the normal histology of the urothelium and is known as transitional cell carcinoma. The data are as follows: 0.08, 0.20, 0.40, 0.50, 0.51, 0.81, 0.90, 1.05, 1.19, 1.26, 1.35, 1.40, 1.46, 1.76, 2.02, 2.02, 2.07, 2.09, 2.23, 2.26, 2.46, 2.54, 2.62, 2.64, 2.69, 2.69, 2.75, 2.83, 2.87, 3.02, 3.25, 3.31, 3.36, 3.36, 3.48, 3.52, 3.57, 3.64, 3.70, 3.82, 3.88, 4.18, 4.23, 4.26, 4.33, 4.34, 4.40, 4.50, 4.51, 4.87, 4.98, 5.06, 5.09, 5.17, 5.32, 5.32, 5.34, 5.41, 5.41, 5.49, 5.62, 5.71, 5.85, 6.25, 6.54, 6.76, 6.93, 6.94, 6.97, 7.09, 7.26, 7.28, 7.32, 7.39, 7.59, 7.62, 7.63, 7.66, 7.87, 7.93, 8.26, 8.37, 8.53, 8.65, 8.66, 9.02, 9.22, 9.47, 9.74, 10.06, 10.34, 10.66, 10.75, 11.25, 11.64, 11.79, 11.98, 12.02, 12.03, 12.07, 12.63, 13.11, 13.29, 13.80, 14.24, 14.76, 14.77, 14.83, 15.96, 16.62, 17.12, 17.14, 17.36, 18.10, 19.13, 20.28, 21.73, 22.69, 23.63, 25.74, 25.82, 26.31, 32.15, 34.26, 36.66, 43.01, 46.12, 79.05. The summary statistics of this data set is given below:

Table 4. Summary Statistics for Data Set III

Min.	Q_1	Median	Q_3	Mean	Max.	Var.	Skewness	Kurtosis
0.08	3.335	6.395	11.885	9.366	79.05	110.425	3.287	18.483

From the descriptive statistics in Tables 2, 3 and 4 for the three data sets respectively, we observed that the three data sets are positively skewed, however, the third data set is highly peaked with a higher skewness coefficient followed by the second and then the first with a low peak. To compare this distribution, we have considered some criteria: the maximized log-likelihood (**−2*l***), Akaike information criterion (AIC), corrected Akaike information criterion (CAIC), Bayesian information criterion (BIC), Hannan-Quinn information criterion (HQIC). These statistics are given as:

$$AIC = 2K - 2l\left(\widehat{\Theta}\right), CAIC = AIC + \frac{2k(k+1)}{n-k-1}, BIC = klog(n) - 2l\left(\widehat{\Theta}\right)$$

and

$$HQIC = 2klog(log(n)) - 2l\left(\widehat{\Theta}\right),$$

where k is the number of parameters in the statistical model, n the sample size and $l\left(\widehat{\Theta}\right)$ is the log-likelihood function evaluated at the maximum likelihood estimates, Θ is the parameters. The distribution with minimum values for these statistics would be chosen as the best distribution to fit the data set in question.

Table 5. Criteria for comparison based on Data Set I

Models	**-2*l***	**AIC**	**CAIC**	**BIC**	**HQIC**	**Ranks**
L	818.3085	822.3085	822.4323	827.5189	824.4173	**7**
TL	732.5721	738.5721	738.8221	746.3876	741.7351	**4**
BL	735.5881	743.5881	744.0091	754.0088	747.8055	**5**
GaL	757.6042	767.6042	768.2425	780.63	772.8759	**6**
MOL	720.1842	730.1842	730.8225	743.2101	735.4560	**3**
WL	561.8530	571.8530	572.4914	584.8789	577.1249	**2**
BTL	386.4596	396.4596	397.0979	409.4854	401.731	**1**

Table 6. Criteria for comparison based on Data Set II

Models	**-2*l***	**AIC**	**CAIC**	**BIC**	**HQIC**	**Ranks**
L	264.3777	274.3777	278.6634	279.3563	275.3496	**7**
TL	250.2585	260.2585	264.5442	265.2372	261.2304	**6**
BL	243.1503	253.1503	257.4359	258.1289	254.1222	**4**
GaL	247.2609	257.2609	261.5466	262.2396	258.2328	**5**
MOL	234.0875	244.0875	248.3732	249.0663	245.0594	**3**
WL	229.1455	239.1455	243.4312	244.1241	240.1174	**2**
BTL	222.134	232.134	236.4197	237.1127	233.1059	**1**

Table 7. Criteria for comparison based on data set III

Models	*-2l*	AIC	CAIC	BIC	HQIC	Ranks
L	669.5493	679.5493	680.0411	693.8095	685.3433	**7**
TL	658.2596	668.2596	668.7514	682.5197	674.0535	**6**
BL	642.6644	652.6644	653.1562	666.9246	658.4584	**4**
GaL	651.7443	661.7443	662.2361	676.0045	667.5383	**5**
MOL	636.3210	646.3210	646.8128	660.5812	652.1149	**3**
WL	602.9675	612.9675	613.4593	627.2277	618.7615	**2**
BTL	566.9087	576.9087	577.4005	591.1689	582.7027	**1**

Tables 5, 6 and 7 provide corresponding values of the -2l**,** AIC, CAIC, BIC and HQIC for each of the distributions. The values of the statistics in all tables (5, 6 and 7) are lower for the BTL distribution followed by the WL and MOL distributions, which is an indication that the BTL distribution performed better than the other distributions considered in the analysis and could be chosen as the best model compared to the other distributions. This also provides additional evidence to the fact that generalizing probability distributions provides compound distributions that are more flexible compared to the parent distributions.

We have also considered a goodness-of-fit test in order to know which distribution has a better fit given some data sets. Hence, we apply the Anderson-Darling ($\boldsymbol{A}^*$), Cramrvon Mises ($\boldsymbol{W}^*$) and Kolmogrov-Smirnov (K-S) statistics. Further information about this statistics can be obtained from Al-Zahrani (2012). These statistics can be computed as:

$$A_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[log\left(F_{BTL}(x_i,\hat{\gamma},\hat{\theta},\hat{\lambda},\hat{a},\hat{b})\right) + log\left(1-F_{BTL}(x_i,\hat{\gamma},\hat{\theta},\hat{\lambda},\hat{a},\hat{b})\right)\right]^2$$

$$W_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F_{BTL}(x_i,\hat{\gamma},\hat{\theta},\hat{\lambda},\hat{a},\hat{b}) - \frac{2i-1}{2n}\right]^2$$

$$K-S = \max_i\left[\frac{i}{n} - F_{BTL}(x_i,\hat{\gamma},\hat{\theta},\hat{\lambda},\hat{a},\hat{b}), F_{BTL}(x_i,\hat{\gamma},\hat{\theta},\hat{\lambda},\hat{a},\hat{b}) - \frac{i-1}{n}\right]$$

where $F_{BTL}(x_i,\hat{\gamma},\hat{\theta},\hat{\lambda},\hat{a},\hat{b})$ is the empirical distribution function and n is the sample size. The distribution with minimum $\boldsymbol{A}^*$**,** $\boldsymbol{W}^*$ and K-S values is chosen as the best distribution to fit the data sets.

Table 8. Goodness-of-fit statistics based on data sets I, II and III.

Models	Data set I			Data set II			Data set III		
	A^*	W^*	K-S	A^*	W^*	K-S	A^*	W^*	K-S
L	1.7758	0.5738	0.4354	1.9524	1.542	0.0794	2.6287	0.6648	0.4381
TL	1.7386	0.4528	0.3957	1.7445	1.4726	0.0759	2.4113	0.4458	0.2175
BL	1.4738	0.3739	0.3018	0.3351	0.0846	0.0563	1.6615	0.1619	0.1277
GaL	1.5949	0.3974	0.3375	0.2754	0.0883	0.0636	1.6523	0.1558	0.1283
MOL	1.4185	0.2585	0.1774	0.2527	0.0857	0.0539	1.4086	0.1382	0.1216
WL	1.2528	0.2263	0,1517	0.2166	0.0808	0.0478	1.0637	0.1254	0.1137
BTL	1.1853	0.1347	0.1013	0.1873	0.0725	0.0473	0.8467	0.1198	0.1039

From the table above, we can observe the A^*, W^* and K-S values of the distributions based on data sets I, II and III. From the table, it is clear and we confirmed that BTL has smaller or lower values of the the A^*, W^* and K-S statistics for all the data sets compared to the WL, MOL, GaL, PL, TL and L distributions, which is an indication that it has a better performance compared to the other distributions. Hence, we can confidently conclude that the BTL distribution is better than the others.

7. Conclusions

In this study, we have introduced the so-called beta transmuted Lomax (BTL) distribution. This is a generalization of the transmuted Lomax distribution using the genesis of the beta distribution. Many distributions including Lomax, beta Lomax and transmuted Lomax are embedded in this newly developed BTL distribution. The mathematical properties of the new family including explicit expansions for the ordinary moments, quantiles, generating functions and order statistics have been provided. The model parameters have been estimated by the maximum likelihood estimation method and the observed information matrix has been determined. It has been shown, by means of a real data set, that special cases of the BTL distribution can provide better fits than other families of distributions.

REFERENCES

ABDULLAHII, U. A., IEREN, T. G., (2018). On the inferences and applications of transmuted exponential Lomax distribution, *International Journal of Advanced Statistics and Probability*, Vol. 6(1), pp. 30–36.

ABDUL-MONIEM, I. B., ABDEL-HAMEED, H. F., (2012). On exponentiated Lomax distribution, *International Journal of Mathematical Archive*, Vol.3, pp. 144–2150.

ABD-ELFATTAH, A. M., ALABOUD, F. M., ALHARBEY, H. A., (2007). On Sample Size Estimation for Lomax Distribution, *Australian Journal of Basic and Applied Sciences*, Vol. 1(4), pp. 373–378.

ABD-ELFATTAH, A. M., ALHARBEY, H. A., (2010). Estimation of Lomax Parameters based on Generalized Probability Weighted Moment, *JKAU: Science*, Vol. 24(2), pp. 171–184.

AHSANULLAH, M., (1991). Record values of the Lomax Distribution, *Statistica Neerlandica*, Vol. 45, pp. 21–29.

AL-ZAHRANI, B., (2012). Goodness-of-fit for the Topp-Leone distribution with unknown parameters, *Applied Mathematical Sciences,* Vol. 6(128), pp. 6355–6363.

ASHOUR, S. K., ELTEHIWY, M. A., (2013). Transmuted Lomax Distribution, *American Journal of Applied Mathematics and Statistics*, Vol. 1(6), pp. 121–27.

AL-AWADHI, S. A., GHITANY, M. E., (2001). Statistical properties of Poisso-Lomax Distribution and its application to repeated accidents data, *Journal of Applied Statistical Science*, Vol. 10(4), pp. 365–372.

BALKEMA, A. A., DE HAAN, L., (1974). Residual life time at great age, *Annuals of Probability*, Vol. 2, pp. 792–804.

BALAKRISHNAN, N., AHSANULLAH, M., (1994). Relations for single and product moments of record values from Lomax distribution, *Sankhya B*, Vol. 56, pp. 140–146.

BINDU, P., SANGITA, K., (2015). Double Lomax Distribution and its Applications, *Statistics*, anno LXXV(3), pp. 331–342.

CHILDS, A., BALAKRISHNAN, N., MOSHREF, M., (2001). Order statistics from non-identical right truncated Lomax random variables with applications, *Statistics: Politics, arts, philosophy*, Vol. 42(2), pp.187–206.

CORDEIRO, G. M., NADARAJAH, S., (2011). Closed form expressions of moments of a class of a beta generalized distributions, *Brazilian Journal of Probability and Statistics,* Vol. 25, pp. 14–33.

CORDEIRO, G. M., ORTEGA, E. M. M., POPOVIC, B. V., (2013). The gamma-Lomax distribution, *Journal of Statistical computation and Simulation iFirst*, doi:10.1080/00949655. 822869.

EUGENE, N., LEE, C., FAMOYE, F., (2002). Beta-normal Distribution and its Applications, *Communications in Statistics – Theory and Methods*, Vol. 31, pp. 497–512.

GHITANY, M. E., AL-AWADHI, F. A., ALKHALFAN, L. A., (2007). Marshall-Olkin extended Lomax Distribution and its application to censored Data, *Communication in Statistics-Theory and Methods*, Vol. 36, pp. 1855–1866.

GHITANY, M. E., ATIEH, B., NADARAJAH, S., (2008). Lindley distribution and its application, *Mathematics and Computers in Simulation*, Vol. 78, pp. 493–506.

GRADSHTEYN, I. S., RYZHIK, I. M., (2000). Table of Integrals, Series, and Products, Academic Press, New York.

GROSS, A. J., CLARK, V. A., (1975). Survival Distributions: Reliability Applications in the Biomedical Sciences. John Wiley and Sons, New York.

GUPTA, GARG, M., MAHESH, G., (2016). The Lomax-Gumbel Distribution, *Palestine Journal of Mathematics,* Vol. 5(1), pp. 35–42.

HASSAN, A. S., AL-GHAMDI, A. S., (2009). Optimum step stress accelerated life testing for Lomax Distribution, *Journal of Application in Scientific Research*, Vol. 5, pp. 2153–2164.

HOWLADER, H. A., HOSSAIN, A. M., (2002). Bayesian Survival Estimation of Pareto Distribution of the second kind based on failure-censored data, *Computational Statistics and Data Analysis*, Vol. 38, pp. 301–314

JONES, MC., (2004). Family of Distributions Arising from Distribution of Order Statistics, *Test*, Vol. 13, pp. 1–43.

KAWSAR, F., UZMA, J., AHMAD, S. P., (2108). Statistical Properties of Rayleigh Lomax distribution with applications in Survival Analysis, *Journal of Data Science,* Vol. 16(3), pp. 531–548.

LEE, E. T., WANG, J. W., (2003). Statistical Methods for Survival Data Analysis (3rd ed.). New York: Wiley.

LEMONTE, A. J., CORDEIRO, G. M., (2013). An extended Lomax distribution, *Statistics*, Vol. 47, pp. 800–816.

LINGAPPAIAH, G. S., (1986). On the Pareto Distribution of the second kind (Lomax Distribution). *Revista de Mathemtica e Estatistica*, Vol. 4, pp. 63–68.

LOMAX, K. S., (1954). Business failures: Another example of the analysis of failure data, *Journal of American Statistical Association*, Vol. 45, pp. 21–29.

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*, Vol. 84, pp. 641–652.

MYHRE, J., SAUNDERS, S., (1982). Screen testing and conditional probability of survival, In: Crowley, J. and Johnson, R.A., eds. Survival Analysis. Lecture Notes-Monograph Series, *Institute of Mathematical Statistics*, Vol. 2, pp. 166–178.

NAYAK, K. T., (1987). Multivariate Lomax Distribution: Properties and Usefulness in Reliability Theory, *Journal of Applied Probability*, Vol. 24, pp. 170–177.

OGUNTUNDE, P. E., KHALEEL, M. A., AHMED, M. T., ADEJUMO, A. O., ODETUNMIBI, O. A., (2017). A New Generalization of the Lomax Distribution with Increasing, Decreasing, and Constant Failure Rate, *Hindawi, Modelling and Simulation in Engineering,* Vol. 2017, pp. 1–7.

TAHIR, M. H., CORDEIRO, G. M., MANSOOR, M., ZUBAIR, M., (2015). The Weibull Lomax distribution: properties and applications, *Hacet J Math Stat,* Vol. 44(2), pp. 461–480.

TAHIR, M. H., ADNAN HUSSAIN, M., CORDEIRO, G. M., HAMEDANI, G. G., MANSOOR, M., ZUBAIR, M., (2016). The Gumbel-Lomax distribution: properties and applications, *Journal of Statistical Theory and Applications*, Vol. 15(1), pp. 61–79.

TERNA, G. I., DAVID, A. K., (2018). On the Properties and Applications of Lomax-Exponential Distribution, *Asian Journal of Probability and Statistics,* Vol. 1(4), pp.1–13.

VIDONDO, B., PRAIRIE, Y. T., BLANCO, J. M., DUARTE, C. M., (1997). Some Aspects of the Analysis of Size Spectra in Aquatic Ecology, *Limnology and Oceanography*, Vol. 42, pp. 84–192.

ZEA, L. M., SILVA, R. B., M. BOURGUIGNON, M., A. M. SANTOS, M. A., CORDEIRO, G. M., (2012). The beta exponentiated Pareto distribution with application to bladder cancer susceptibility, *International Journal of Statistics and Probability*, Vol. 1, pp. 8–19.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 35–60, DOI 10.21307/stattrans-2020-013
Submitted – 22.03.2019; accepted for publishing – 10.03.2020

On the choice of the number of Monte Carlo iterations and bootstrap replicates in Empirical Best Prediction

Adam Chwila[1], **Tomasz Żądło**[2]

ABSTRACT

Empirical Best Predictors (EBPs) are widely used for small area estimation purposes. In the case of longitudinal surveys, this class of predictors can be used to predict any given population or subpopulation characteristic for any time period, including future periods. Generally, the value of an EBP is computed by means of Monte Carlo algorithms, while its MSE is usually estimated using the parametric bootstrap method. Model-based simulation studies of the properties of the predictors require numerous repetitions of the random generation of population data. This leads to a question about the dependence between the number of iterations in all the procedures and the stability of the results. The aim of the paper is to show this dependence and to propose methods of choosing the appropriate number of iterations in practice, using a set of real economic longitudinal data available at the United States Census Bureau website.

Key words: survey sampling, economic longitudinal data, prediction for future periods.

1. Introduction

Empirical Best Predictors have been used in small area estimation problems for a long time. In papers published by Jiang and Lahiri (2001) and Jiang (2003) prediction problems under generalized linear mixed models were studied. A large number of papers were published after a well-known Molina and Rao (2010) paper, where this class of predictors was used to predict poverty measures. What is more, they presented a special case of the predictor under normality of the transformed variable of interest together with the proposal of a very fast algorithm for a special case of the model called the nested error mixed model. Then, many authors generalized these results relaxing normality assumption (e.g. Elbers and vad der Weide (2014), Diallo (2014) and Diallo and Rao (2018)), considering nonlinear models and usually the prediction of small area

[1] University of Economics in Katowice, Katowice, Poland, E-mail: achwila@gmail.com.
ORCID: https://orcid.org/0000-0003-4671-4298.
[2] University of Economics in Katowice, Katowice, Poland, E-mail: tomasz.zadlo@ue.katowice.pl.
ORCID: https://orcid.org/0000-0003-0638-0748.

fractions (e.g. Berg and Chandra (2014), Boubeta, Lombardía and Morales (2016, 2017), Hobza and Morales (2016), Zimmermann and Münnich (2018)), analyzing the problem of back transformation of the variable of interest (Molina and Martin (2018)) and studying the semi-parametric EBP (Marino et al. (2019)).

In these papers, the authors assume a different number of iterations in the EBP procedure (which will be denoted by L), in the parametric bootstrap method used to estimate MSE (which will be denoted by B) and in Monte Carlo simulation studies (which will be denoted by K). The applications presented in these papers are usually supported by model-based simulation studies. It gives possibility to use additional methods to choose the appropriate number of iterations based on simulation results, such as stability of simulation results or the simulation bias of unbiased Best Predictor, which cannot be computed in practice (for real data).

It is clear that the appropriate choice of the number of iterations is different for different data, different models and different prediction problems and hence we would like to present some examples studied by different authors. Although in practice, as stated by Tzavidis et al. (2018), usually $L = 50$ or $L = 100$ is used, in the small area estimation literature different numbers of iterations L in the EBP procedure are studied:

- in applications: from 50 to 1000 in Molina and Rao (2010), 100 in Guadarrama, Molina and Rao (2018),
- in simulation studies: 50 in Molina and Rao (2010), 100 in Das and Haslett (2019), 500 in Boubeta, Lombardía, Morales (2017) and 2500 in Boubeta, Lombardía, Morales (2016).

Examples of numbers of iterations B taken into account by different authors are:

- in applications: 500 in Molina and Rao (2010), Hobza and Morales (2016), Boubeta, Lombardía and Morales (2017), Guadarrama, Molina and Rao (2018),
- in simulation studies: 500 in Molina and Rao (2010), Boubeta, Lombardía and Morales (2016), Guadarrama, Molina and Rao (2018); and 1000 in González-Manteiga, Lombardía, Molina, Morales and Santamaría (2008).

The numbers of iterations in Monte Carlo simulation studies assumed by different authors equal: 500 in Das and Haslett (2019), 500 and 10 000 and 50 000 for different purposes in Molina and Rao (2010); 500 and 1 000 and 10 000 for different purposes in Guadarrama, Molina and Rao (2018); 1 000 in González-Manteiga, Lombardía, Molina, Morales and Santamaría (2008), Guadarrama, Molina and Rao (2014), Boubeta, Lombardía, Morales (2016, 2017); 5 000 in Diallo and Rao (2018); 10 000 in Hobza and Morales (2016) and Molina and Martín (2018); 50 000 in Jiang and Lahiri (2006).

Based on a real economic longitudinal dataset we analyse three problems which, according to our best knowledge, are not presented in the literature:

- the dependence between the number of iterations L of the EBP procedure and the stability of EBP values,
- the dependence between the number of iterations B in the parametric bootstrap procedure and the stability of values of MSE estimators,
- the dependence between the number of Monte Carlo iterations K and the stability of ratios of MSEs of the predictors: the EBP and the BP.

We also propose:

- two criteria allowing the appropriate choice of L and B, which can be used in practice (based on real sample data),
- a criterion allowing to choose the appropriate number of iterations K in simulation studies.

2. Some remarks on bootstrap procedures

In this section we present the literature review on the convergence of bootstrap procedures taking into account two issues. Firstly, we are interested in analysing how bootstrap estimators under B replications approximate their values when B tends to infinity. Secondly, we show that based on some bootstrap procedures we can obtain asymptotically unbiased estimators of some unknown parameters. Although we are interested in the parametric bootstrap method, we discuss available results for different bootstrap procedures.

Davison and Hinkley (1997) pp. 34–37 study the problem of the decomposition of variances of different bootstrap estimators into the part resulting from data variation and simulation variation. They study nonparametric bootstrap procedure and derive variances and bootstrap variances of the following statistics: bootstrap estimator of the bias of the sample mean, bootstrap estimator of the variance of the sample mean and bootstrap estimator of the variance of the sample quantile. They present bootstrap variances of these statistics as functions of: (i) their unconditional variance and (ii) the simulation variance depending on the number of bootstrap iterations. It gives a direct tool to determine the number of nonparametric bootstrap replicates to obtain the required ratio of the simulation variance and the unconditional variance. Davison and Hinkley (1997) pp. 155–156 study also the problem of the convergence of the parametric bootstrap procedure but in the case of testing hypotheses. They derive powers of tests in two cases: for the given number of bootstrap iterations and when it tends to infinity. Their ratio is a function of bootstrap replicates, which allows one to determine the number of replicates to obtain the required level of the ratio.

Efron and Tibshirani (1986) p. 72 study the number nonparametric bootstrap replications in the case of estimation of the standard error showing that the CV of the bootstrap estimator of the standard error based on B replications is a function of: (i) the CV of the bootstrap estimator of the standard error based on infinite B replications, (ii) the number of bootstrap replications and (iii) the expected value (over the distribution of the variable of interest) of the kurtosis of the bootstrap distribution of the considered estimator. Because the formula is generally not estimable, it is not used to find a specific value of bootstrap replications but to determine a range of acceptable values.

An interesting procedure is proposed and studied by Andrews and Buchinsky (1997, 2000, 2001). They study two cases in bootstrap procedures – firstly, B iterations and, secondly, an infinite number of iterations. They determine the number of bootstrap iterations to obtain the value of the modulus of the percentage deviation between values of bootstrap estimators in these two cases not greater than the specified value with the declared probability close to 1. It can be used for different bootstrap techniques including parametric and nonparametric bootstrap and both for independent and dependent data. Estimation of the MSE is not considered by the authors – they consider estimation of the square root of variance, confidence intervals, test statistics and p-values. In simulation studies they consider properties of their method only for standard nonparametric bootstrap.

Even if a bootstrap estimator accurately approximates its value under infinite number of replications, it does not mean that it is a good estimator of the parameter. Usually bootstrap approximates the population distribution of certain sample statistics but the failure in convergence of the bootstrap distribution to the correct distribution may also occur (e.g. Beran (1997)). Hall and Martin (1988) prove that the nonparametric bootstrap quantile variance estimator converges with the increase of the sample size to the true variance (but slowly). Singh (1981) shows that the nonparametric bootstrap asymptotically (when the sample size tends to infinity) approximates the population distribution of the standardized sample mean and the distribution of the sample quantiles. The parametric bootstrap MSE estimator of the empirical best linear unbiased predictor proposed by Butar and Lahiri (2003) estimates the unknown MSE with the bias of order $o(D^{-1})$, where D is the number of small areas. Chatterjee, Lahiri and Li (2008) use parametric bootstrap to estimate the distribution of the centered and scaled empirical best linear unbiased predictor and show that it accurately approximates the true distribution (and derive the order of the approximation). Hall and Maiti (2006) propose a very accurately parametric bootstrap confidence intervals, that do not depend of the form of small area predictor, with the coverage error $O(D^{-3})$. Hall and Maiti (2006) present also results crucial for our analysis – they prove that the biases of parametric bootstrap MSE estimators (considered in this paper) of both the empirical best linear unbiased predictor and the

empirical best predictor are of order $O(D^{-1})$, where D is the number of small areas. What is more, the double bootstrap MSE estimator of the predictor, not considered in our paper due to very time-consuming computations, is of order $O(D^{-2})$.

3. Empirical Best Predictor

We consider the model-based approach in survey sampling assuming the following longitudinal mixed linear model for population data:

$$Q(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{1}$$

where $Q(\mathbf{Y})$ is the random vector of the variable of interest after transformation $Q(.)$ including random variables for future periods in the case of longitudinal data, $\mathbf{X}$ and $\mathbf{Z}$ are known matrices of full ranks of the auxiliary variables including known or assumed values for future periods, $\boldsymbol{\beta}$ is the unknown vector of fixed effects, $\mathbf{v}$ and $\mathbf{e}$ – called vectors of random effects and random components – are independent, $\mathbf{v} \sim (\mathbf{0}, \mathbf{G}(\boldsymbol{\delta}))$ and $\mathbf{e} \sim (\mathbf{0}, \mathbf{R}(\boldsymbol{\delta}))$, where $\boldsymbol{\delta}$ is a vector of unknown parameters called variance components. Without the loss of the generality, we assume that first elements in the population vector $Q(\mathbf{Y})$, are the random variables which realizations are known from the longitudinal survey, which can be written as $Q(\mathbf{Y}) = \left[Q(\mathbf{Y}_s^T) \quad Q(\mathbf{Y}_r^T) \right]^T$, where subscript "s" is used for the sample and "r" for non-sampled elements. What is more, a similar decomposition can be used for matrices of auxiliary variables: $\mathbf{X} = \left[\mathbf{X}_s^T \quad \mathbf{X}_r^T \right]^T$ and $\mathbf{Z} = \left[\mathbf{Z}_s^T \quad \mathbf{Z}_r^T \right]^T$, for the vector of random components $\mathbf{e} = \left[\mathbf{e}_s^T \quad \mathbf{e}_r^T \right]^T$ and for covariance matrix of random components $\mathbf{R}(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{R}_{ss}(\boldsymbol{\delta}) & \mathbf{R}_{sr}(\boldsymbol{\delta}) \\ \mathbf{R}_{rs}(\boldsymbol{\delta}) & \mathbf{R}_{rr}(\boldsymbol{\delta}) \end{bmatrix}$. Based on (1), the covariance matrix of $Q(\mathbf{Y})$, denoted by $\mathbf{V}(\boldsymbol{\delta})$, is given by $D^2(Q(\mathbf{Y})) = \mathbf{V}(\boldsymbol{\delta}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T + \mathbf{R}(\boldsymbol{\delta})$ and it can be decomposed as follows $\mathbf{V}(\boldsymbol{\delta}) = \begin{bmatrix} \mathbf{V}_{ss}(\boldsymbol{\delta}) & \mathbf{V}_{sr}(\boldsymbol{\delta}) \\ \mathbf{V}_{rs}(\boldsymbol{\delta}) & \mathbf{V}_{rr}(\boldsymbol{\delta}) \end{bmatrix}$, where $\mathbf{V}_{ss}(\boldsymbol{\delta}) = \mathbf{Z}_s\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}_s^T + \mathbf{R}_{ss}(\boldsymbol{\delta})$, $\mathbf{V}_{rr}(\boldsymbol{\delta}) = \mathbf{Z}_r\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}_r^T + \mathbf{R}_{rr}(\boldsymbol{\delta})$, $\mathbf{V}_{sr}(\boldsymbol{\delta}) = \mathbf{Z}_s\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}_r^T + \mathbf{R}_{sr}(\boldsymbol{\delta})$ and $\mathbf{V}_{rs}(\boldsymbol{\delta}) = \mathbf{V}_{sr}^T(\boldsymbol{\delta})$.

To estimate parameters of (1), i.e. vectors $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, different methods can be used including the restricted maximum likelihood method (REML) used in this paper (see e.g. Jiang (2007) pp. 12-15). In the method the value of the estimator of $\boldsymbol{\delta}$, denoted by $\hat{\boldsymbol{\delta}}$, is computed by maximization of the Gaussian likelihood function of $\mathbf{A}^T\mathbf{Y}_s$, where $\mathbf{A}$ is any matrix such that $\mathbf{A}^T\mathbf{X}_s = \mathbf{0}$. The method is robust on non-normality – it gives consistent estimators even if the distribution is not normal (Jiang (1996)). The

estimator of $\boldsymbol{\beta}$ is given by (e.g. Jiang (2007) p. 75): $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1}(\hat{\boldsymbol{\delta}}) \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1}(\hat{\boldsymbol{\delta}}) \mathbf{Y}_s$. The empirical best linear unbiased predictor of $\mathbf{v}$ is as follows (e.g. Jiang (2007) p. 76): $\hat{\mathbf{v}}(\hat{\boldsymbol{\delta}}) = \mathbf{G}(\hat{\boldsymbol{\delta}}) \mathbf{Z}_s^T \mathbf{V}_{ss}^{-1}(\hat{\boldsymbol{\delta}})(\mathbf{Y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}})$.

Under (1) the best predictor $\hat{\theta}$ of any function $\theta(Q^{-1}(\mathbf{Y}))$, or shortly θ, minimizing the mean squared error is given by (e.g. Molina and Rao (2010)):

$$\hat{\theta}_{BP} = E(\theta \mid Q(\mathbf{Y}_s)). \tag{2}$$

Special cases of $\theta(Q^{-1}(\mathbf{Y}))$ are population and subpopulation characteristics such as the mean or the median in the current or future period. The value of (2) can be computed if the shape and the parameters of the distribution $Q(\mathbf{Y}_r) \mid Q(\mathbf{Y}_s)$ are known. In practical applications the shape of the multivariate distribution of $Q(\mathbf{Y})$ is assumed, the parameters of the distribution (in the case of (1) - $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$) are estimated based on the known realization of $Q(\mathbf{Y}_s)$ (which gives $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$), and the distribution of $Q(\mathbf{Y}_r) \mid Q(\mathbf{Y}_s)$ is derived (or directly the conditional expectation given by (2)). The two-stage predictor obtained according to this idea is called the Empirical Best Predictor (EBP). Its value can be computed based on the following iterative algorithm, presented originally by Molina and Rao (2010):

- generate L vectors $Q(\mathbf{Y}_r)$ (denoted by $Q(\mathbf{Y}_r^{(l)})$, where $l = 1, 2, ..., L$) based on the empirical distribution of $Q(\mathbf{Y}_r) \mid Q(\mathbf{Y}_s)$ (the distribution of $Q(\mathbf{Y}_r) \mid Q(\mathbf{Y}_s)$ where $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are replaced by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$),
- construct L population vectors $Q(\mathbf{Y}^{(l)}) = \left[Q(\mathbf{Y}_s^T) \quad Q(\mathbf{Y}_r^{(l)T}) \right]^T$ $(l = 1, 2, ..., L)$, where one realization of $Q(\mathbf{Y}_s)$ available from the sample and different realizations of $Q(\mathbf{Y}_r)$ are used,
- compute the EBP as $\hat{\theta}_{EBP} = L^{-1} \sum_{l=1}^{L} \theta(Q^{-1}(\mathbf{Y}^{(l)}))$ (which means that the back transformation is needed).

If we assume (1) and multivariate normality of the transformed variable of interest, then the distribution $Q(\mathbf{Y}_r) \mid Q(\mathbf{Y}_s)$ is multivariate normal with the following vector of expected values $\mathbf{X}_r \boldsymbol{\beta} + \mathbf{V}_{rs}(\boldsymbol{\delta}) \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \left(Q(\mathbf{Y}_s) - \mathbf{X}_s \boldsymbol{\beta} \right)$ and the following variance-covariance matrix $\mathbf{V}_{rr}(\boldsymbol{\delta}) - \mathbf{V}_{rs}(\boldsymbol{\delta}) \mathbf{V}_{ss}^{-1}(\boldsymbol{\delta}) \mathbf{V}_{sr}(\boldsymbol{\delta})$. Molina and Rao (2010) also propose a very fast algorithm for EBP computation for the special case of (1) called the nested error mixed linear model, where the generation of population vectors from the multivariate normal conditional distribution is replaced by iid generation using the univariate normal distribution.

To estimate the mean squared error of the EBP the parametric bootstrap method can be used. The bootstrap model used to generate the data is given by (Chatterjee, Lahiri and Li (2008), González-Manteiga et al. (2008)):

$$Q(\mathbf{Y}^*) = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* + \mathbf{e}^*, \tag{3}$$

where $\mathbf{v}^* \sim \mathrm{N}(\mathbf{0}, \mathbf{G}(\hat{\boldsymbol{\delta}}))$, $\mathbf{e}_s^* \sim \mathrm{N}(\mathbf{0}, \mathbf{R}(\hat{\boldsymbol{\delta}}))$, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\beta}}$ are estimators of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$, respectively. We use the restricted maximum likelihood method to estimate the parameters. The MSE estimator is given by (González-Manteiga et al. (2008)):

$$\hat{MSE}(\hat{\theta}_{EBP}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_{EBP}(Q^{-1}(\mathbf{Y}_s^{*(b)})) - \theta(Q^{-1}(\mathbf{Y}^{*(b)}))\right)^2, \tag{4}$$

where $\hat{\theta}_{EBP}(Q^{-1}(\mathbf{Y}_s^{*(b)}))$ and $\theta(Q^{-1}(\mathbf{Y}^{*(b)}))$ are values of the predictor and the predicted characteristic, respectively, for the bth realization of the bootstrap model.

4. Data and model

Our considerations are based on whole population real economic longitudinal data available at the website of the United States Census Bureau (https://www.census.gov/library/ publications/ 2011/ compendia/ usa-counties-2011.html):

- the number of new private housing units of single-family houses authorized by building permits for years 2007-2009 (the variable of interest),
- the number of births for years 2006-2008 (the first auxiliary variable),
- the private nonfarm annual payroll in USD for years 2006-2008 (the second auxiliary variable)

for 177 counties from the following $D = 4$ states: Washington, Idaho, Oregon and California. We consider a relatively small population because of very time-consuming computations. Auxiliary variables are from the year preceding the construction of housing units. What is more, we assume that values of both auxiliary variables for 2009 are known and they are used to predict population and subpopulation characteristics of the variable of interest in 2010 (treated as the future period).

We mimic a real analysis. Because our further considerations are model-based and conditional (based on the given sample), we draw one sample. It is a stratified sample of counties, where states are strata, with proportional allocation (of size 20% of the population size) in the first period. Then, the same elements in periods 2 and 3 are in the sample, which gives a balanced panel sample. This gives us the division of the whole population dataset into the sample, where both the auxiliary information and the values of the variable of interest are available, and the non-sampled elements for which only auxiliary information is known.

A relatively large sample fraction is considered because: (i) the population size is small due to the complexity of computations and (ii) – at the same time – we must obtain enough sample observations for model parameters estimation purposes. Of course, this specific setting implies a limited generalization of our results for different datasets.

We consider the problem of prediction of the following population and subpopulations characteristics for the future period: means, medians, standard deviations, quartile deviations, moment and quartile skewness coefficients. For all of the variables the log transformation is used (after adding a constant), and hence the back transformation of the variable of interest is used to compute the EBP.

To find the best fitted linear mixed model we use the procedure presented by Verbeke and Molenberghs (2009) pp. 121-132, where firstly the fixed effects models are considered, then different random effects are added, to finally obtain the mixed model (in our case based on the AIC criterion). We have considered about 700 different models for both cases considered below.

The model we have chosen is given by (it will be called **model 1**):

$$Q(Y_{idt}) = \beta_1 + (\beta_2 + v_{1d})x_{1idt} + (\beta_3 x_{2idt} + v_{1i})\ln(t) + v_{2i} + v_{2d} + e_{1idt}, \qquad (5)$$

where $Q(Y_{idt}), x_{1idt}, x_{2idt}$ are log transformed variables (after adding a constant) $i = 1,2,...,N;$ $d = 1,2,...,D;$ $t = 1,2,...,M;$ $v_{1d} \sim (0, \sigma^2_{v1d}),$ $v_{2d} \sim (0, \sigma^2_{v2d}),$ $v_{1i} \sim (0, \sigma^2_{v1i}),$ $v_{2i} \sim (0, \sigma^2_{v2i}),$ $e_{1idt} \sim (0, \sigma^2_{e1}),$ random effects and random components are mutually independent, in our case the population size $N = 177$, the number of time periods (including the future one) $M = 4$ and the number of subpopulations $D = 4$.

Additionally, we have chosen the best fitted nested error model with the logarithmic trend (as in (5)) and only one random effect for the purpose of the comparative study (it will be called **model 2**):

$$Q(Y_{idt}) = \beta_4 + \beta_5 x_{1idt} + (\beta_6 x_{1idt} + \beta_7)\ln(t) + v_{3i} + e_{2idt}, \qquad (6)$$

where $v_{3i} \sim (0, \sigma^2_{v3i}),$ $e_{2idt} \sim (0, \sigma^2_{e2}),$ v_{3i} and e_{2idt} are mutually independent and other notations are as in (5). The choice of this class of models is due to the possibility of using the fast algorithm for EBP computation proposed in Molina and Rao (2010).

Based on permutation tests we can claim that parameters of both models are statistically significant. The normality assumption for both models is met for the considered longitudinal sample data (we have used Shapiro-Wilk test and residuals after the Cholesky transformation).

In the next sections we will consider EBPs under model 1 (given by (5)) denoted by EBP1 and under model 2 (given by (6)) denoted by EBP2, and their parametric bootstrap MSE estimators based on (4).

5. Number of iterations in EBP procedure

We consider stability of values of EBP1 and EBP2 computed under different numbers of iterations L (where $L = 100, 200, ..., 1000$) taken into account in the EBP iterative procedure presented in Section 3. Each boxplot in Figure 1 presents $M = 500$ values of EBP1 of one out of six population characteristics computed for different numbers of iterations L. For example, the first boxplot at the top left corner of Figure 1 presents 500 values of EBP1 used to predict the population mean, computed based on $L = 100$ iterations. In Figure 1, we see that results in each out of six considered cases tend to stabilize at around $L = 400$.

Similar plots are prepared for EBP2 and the prediction in the arbitrarily chosen third subpopulation (which gives 3 additional plots not presented in the paper). Then, based on values presented in each boxplot, we compute the value of the CV and present all of the results in Figure 2. The CV is given by:

$$CV_L = CV_{L,M}(\hat{\theta}^L_{EBP}) = \left(M^{-1} \sum_{i=1}^{M} \hat{\theta}^{L,i}_{EBP} \right)^{-1} \left(M^{-1} \sum_{i=1}^{M} \left(\hat{\theta}^{L,i}_{EBP} - M^{-1} \sum_{i=1}^{M} \hat{\theta}^{L,i}_{EBP} \right)^2 \right)^{0.5}, \quad (7)$$

where $CV_{L,M}(\hat{\theta}^L_{EBP})$ is the coefficient of variation computed based on M values of EBP and L is the number of iterations in the case of ith EBP estimation. For example, the star at the top left corner in Figure 2 is the value of the CV computed for values presented in the boxplot at the top left corner in Figure 1. Hence, in Figure 2 we can compare CVs of EBP1 and EBP2 of different population and subpopulation characteristics predicted for the future period. Coefficients of variation decrease from 5.74% ($L = 100$) for the standard deviation to 0.34% ($L = 1000$) for the median.

In six parts of Figure 2 we present the differences for different functions of random variables predicted for the future period. If we compare prediction methods (EBP1 and EBP2 of population characteristics; EBP1 and EBP2 of subpopulation characteristics), the results are similar – the differences are substantial only in the case of prediction of the standard deviation and the mean for small numbers of iterations. The differences between CVs for the third subpopulation and the whole population (EBP1 of population and subpopulation characteristics; EBP2 of population and subpopulation characteristics) are higher, especially for prediction of functions based on quantiles.

The results presented in Figure 2 are based on real sample data and they can be used to choose the number of iterations in practice assuming the maximum acceptable value of the CV (possibly different for different considered cases). For example, if – in the case of EBP computations – we accept values of the CV smaller or equal 3% in all of the considered cases, then $L = 400$ is sufficient. It can be noticed that for all of the considered prediction problems the linear growth in L causes a decrease in the CV

slower than the linear one. However, the exact path of the convergence is dependent on the predicted characteristic as well as on the choice of the population or the subpopulation (which is connected to the sample size). For the quantile measures the required number of iterations (for the certain CV goal) would be greater if the subpopulation was considered instead of the whole population. It can also be noticed that for certain characteristics (the standard deviation, the moment skewness coefficient, the quartile skewness coefficient) the absolute improvement of the results is more tangible, especially for the number of iterations around $L = 100$. Therefore, the incentive of enlarging the number of iterations would be dependent of the above aspects.

Alternatively, we can consider the assumed acceptable value of the change of the CV (see Figure 7 in Appendix), which is given by:

$$RCCV_L^{L+100} = 100 CV_L^{-1} (CV_{L+100} - CV_L), \tag{8}$$

where CV_L is given by (7). For example, if we accept the decrease (comparing L with $L-100$) of the CV smaller or equal 20%, then $L = 400$ is sufficient for all of the considered cases, too. It can be noticed that the relative change of CVs is a measure that, unlike the CV itself, behaves very similarly for all the considered characteristics. For example, the difference between $L = 100$ and $L = 200$ iterations causes the improvement around 30%. It can be also noticed that the relative improvements are independent on the choice of the population or the subpopulation. This means that the chosen measure (the CV or the relative change of CVs) may have an impact on the final conclusion. The difference in the observed convergence between EBP1 (based on the model with 4 random effects) and EBP2 (based on the model with 1 random effect) is negligible for all cases besides the standard deviation and the mean.

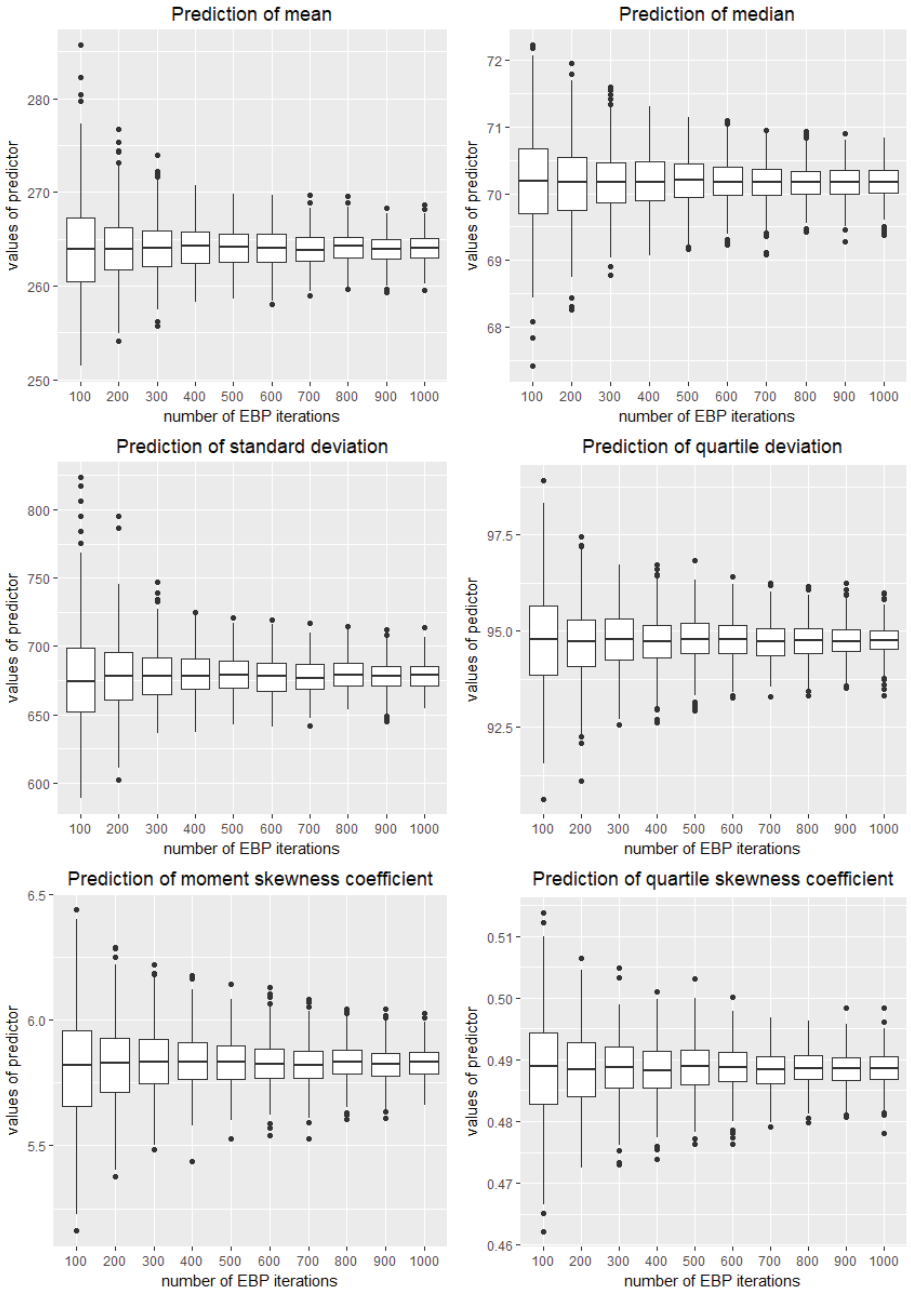


Figure 1. Variability of 500 values of EBP1 of different population characteristics in the future period computed for different numbers of iterations $L = 100, 200, ..., 1000$

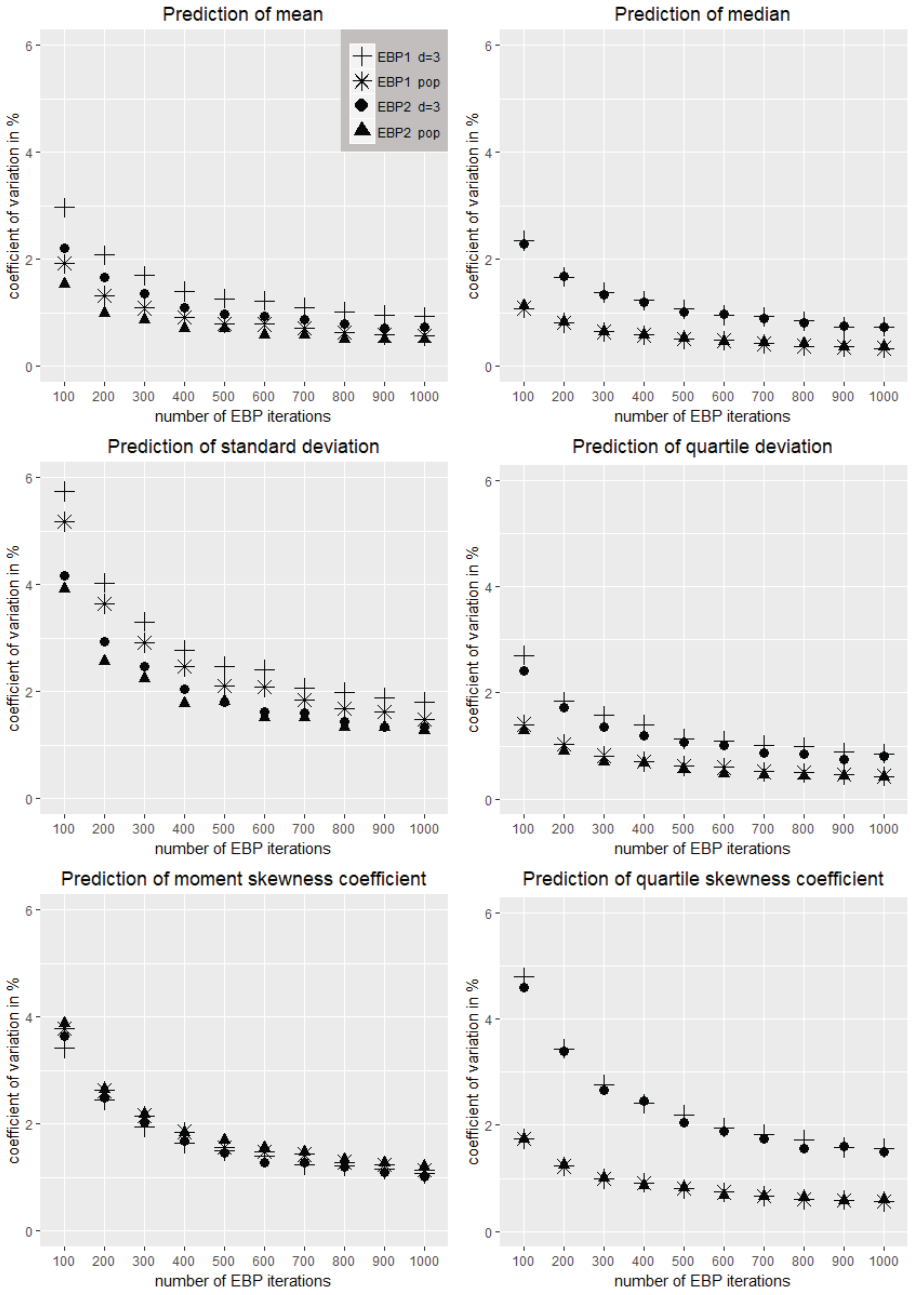


Figure 2. CVs computed based on 500 values of EBP1 and EBP2 of different population and subpopulation characteristics in the future period for different numbers of iterations $L = 100, 200, ..., 1000$

6. Number of iterations for parametric bootstrap MSE estimator

In all of the cases considered in this section EBPs are computed assuming $L = 500$ – higher than suggested in the previous section (i.e. $L = 400$) to obtain more stable results for MSE estimation. We consider MSE estimators of EBP1 under model 1 and EBP2 under model 2 computed for different numbers of iterations B (where $B = 100, 200, ..., 1000$) taken into account in the parametric bootstrap procedure presented in Section 3. Each boxplot in Figure 3 presents 100 values of MSE estimator of EBP1 computed for different numbers of iterations B for one out of six prediction problems. For example, the first boxplot at the top left corner presents 100 values of the MSE estimator of EBP1 of the population mean computed based on $B = 100$ iterations. The results presented in Figure 3 for $B = 100$ and $B = 200$ are generally unstable, at $B = 300$ they start to stabilize, for B from 500 to 1000 are quite similar. Similar figures are created for bootstrap MSE estimators of EBP2 and the third subpopulation (which gives three additional figures not presented in the paper).

Then, based on the values presented in each boxplot we compute the value of the CV and present all of the results in Figure 4. The coefficient of variation, similarly as in the case of (7), is given by:

$$CV_B = CV_{B,M}(\hat{MSE}^B(\hat{\theta}^L_{EBP})) =$$

$$= \left(M^{-1}\sum_{i=1}^{M} \hat{MSE}^{B,i}(\hat{\theta}^L_{EBP}) \right)^{-1} \left(M^{-1}\sum_{i=1}^{M} \left(\hat{MSE}^{B,i}(\hat{\theta}^L_{EBP}) - M^{-1}\sum_{i=1}^{M} \hat{MSE}^{B,i}(\hat{\theta}^L_{EBP}) \right)^2 \right)^{0.5}, \quad (9)$$

where $CV_{B,M}(\hat{MSE}^B(\hat{\theta}^L_{EBP}))$ is the coefficient of variation based on M values of the MSE estimator of EBP, L is the fixed number of EBP iterations, B is the number of bootstrap iterations in the case of ith MSE estimation.

For example, the star symbol at the top left corner in Figure 4 presents the value of the CV computed based on the values presented in the boxplot at the top left corner in Figure 3. Hence, we can compare CVs of the values of MSE estimators of EBP1 and EBP2 for six different prediction problems.

Values in Figure 4 decrease, although the decrease is not as smooth as in the case of the EBP (compare with Figure 2) – possibly due to the additional source of the variability resulting from the computation of EBP values and a smaller number of values per boxplot. Coefficients of variation decrease from 43.5% $(B = 100)$ for the standard deviation to 4.31% $(B = 1000)$ for the quantile skewness coefficient. The results for different models (MSEs estimators under model 1 and under model 2), the population and the third subpopulation and different prediction problems (except the prediction of the standard deviation in the future period) are similar, especially for larger numbers of iterations B.

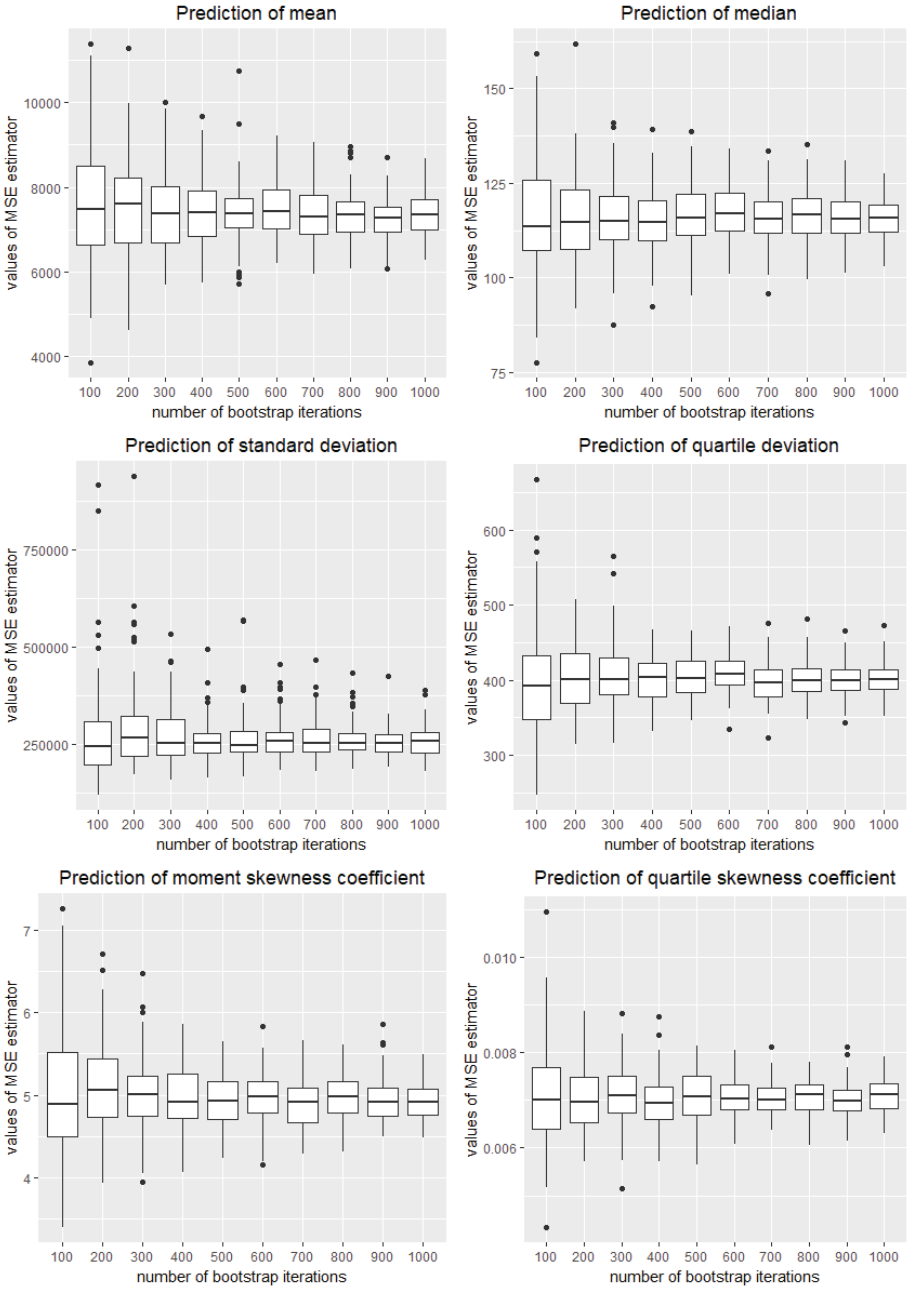


Figure 3. Variability of 100 values of parametric bootstrap MSE estimators of EBP1 of different population characteristics in the future period computed for different numbers of iterations $B = 100, 200, ..., 1000$

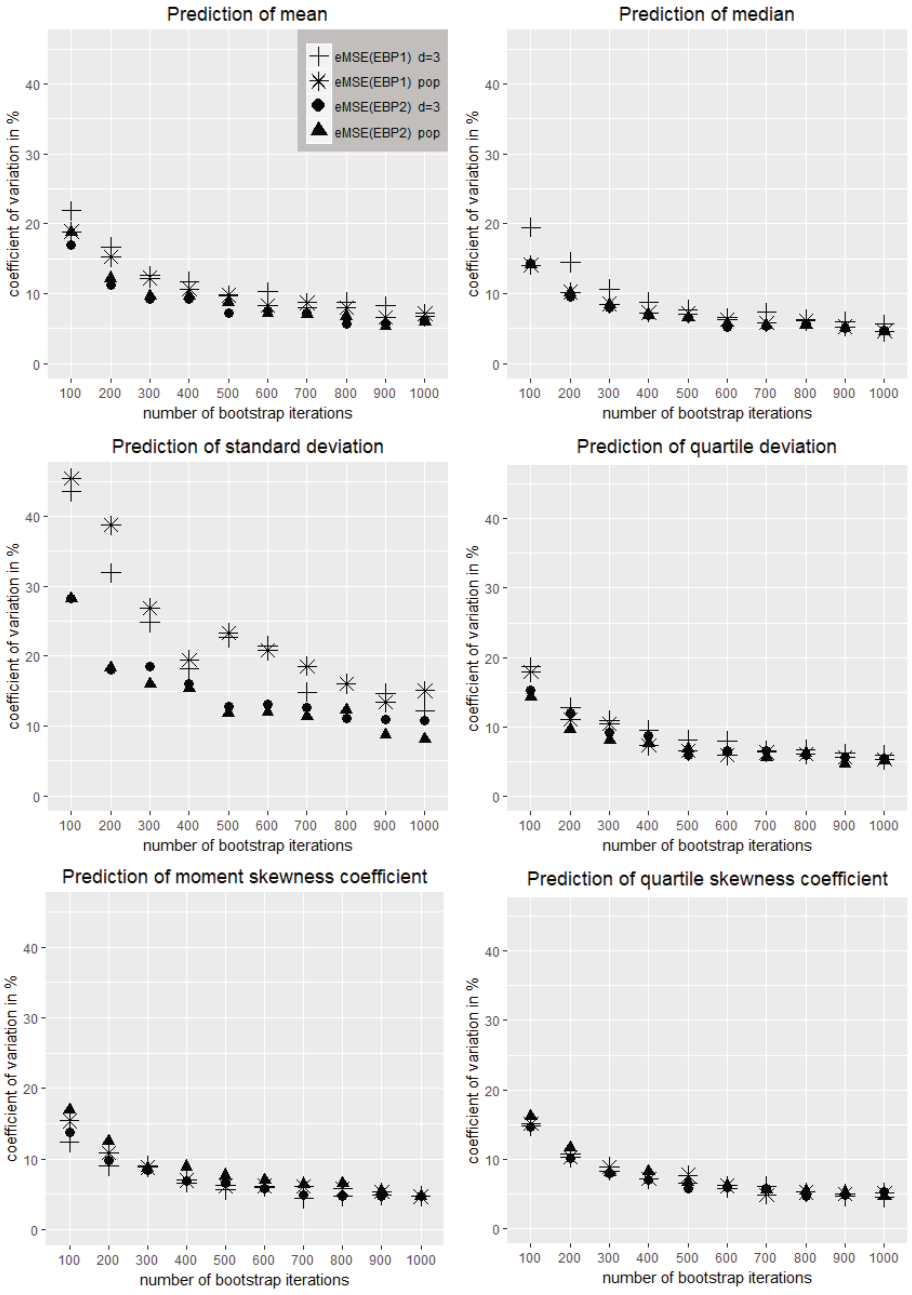


Figure 4. CVs computed based on 100 values of parametric bootstrap MSE estimators of EBP1 and EBP2 of different population and subpopulation characteristics in the future period for different numbers of iterations $B = 100, 200, ..., 1000$

Similarly to the previous section, the choice of the appropriate number of bootstrap iterations can be made based on the maximum acceptable value of the CV of values of MSE estimators. For example, if – in this case – we accept values of the CV smaller or equal 10%, then $B = 500$ is sufficient in most of the considered cases except the problem of prediction of the standard deviation (see Figure 4). Similarly to the results presented in the previous section, the linear growth in the number of bootstrap iterations causes a decrease in the CV slower than the linear one for all the considered prediction problems. The exact paths of convergence vary in dependence on the considered characteristic, the choice between the subpopulation and the population and the considered model (which can be noticed especially for the median).

Alternatively, we can consider the acceptable value of the change of the CV (see Figure 8 in Appendix), which is given by:

$$RCCV_B^{B+100} = 100 CV_B^{-1}(CV_{B+100} - CV_B), \tag{10}$$

where CV_B is given by (9). If we compare Figure 8 with Figure 7, we see that the changes are less stable because of the same reasons, as stated in the case of the comparison of Figure 4 with Figure 2 in the previous paragraph. If we accept the decrease (comparing B with $B-100$) of the CV smaller or equal 20%, then in most of the considered cases $B = 500$ is sufficient. The relative changes of CVs behave similarly for all the considered characteristics, although the results are quite unstable and difficult for more in-depth analysis. The difference in the convergence between EBP1 (based on the model with 4 random effects) and EBP2 (based on the model with 1 random effect) is negligible for all the cases beside standard deviation and mean, similarly as in the previous section.

7. Number of iterations in Monte Carlo simulation studies

Our considerations, proposals and conclusions in two previous sections were based on the real sample data. In this section we study the problem of model-based simulation studies of the properties of EBPs, where values of the variable of interest are generated based on model 1 (see (5)) for EBP1 and model 2 (see (6)) for EBP2. In simulation studies the appropriate number of Monte Carlo iterations is usually chosen based on the accepted value of the absolute simulation biases of unbiased statistics. For example, in design-based experiments Barbiero and Mecatti (2010) accept the relative values of modulus of simulation biases of the unbiased Horvitz-Thompson estimator and the unbiased estimator of its variance equal 1% and 3%, respectively. Similarly, in our case, we can assume the accepted relative value of modulus of simulation biases for the (unbiased) Best Predictors. But in the case of EBPs it is known that the ratio of MSEs of the EBP and the BP is greater than 1, while its simulation value may be lower than 1

even if the simulation bias of the unbiased BP is low. It means that the simulation ratio of MSEs of the EBP and BP may be of greater importance (as the measure of the quality of the Monte Carlo simulation study under the given number of iterations) than the value of the simulation bias of BP. Hence, we propose to check in simulation studies if (i) the simulation ratio of these MSEs is greater than 1 and (ii) to check the stability of values of these ratios. The value of the criterion is computed as:

$$K^{-1}\sum_{k=1}^{K}(\hat{\theta}_{EBP}^{k}-\theta^{k})^{2}\left(K^{-1}\sum_{k=1}^{K}(\hat{\theta}_{BP}^{k}-\theta^{k})^{2}\right)^{-1}, \tag{11}$$

where K is the number of iterations in the simulation study, $\hat{\theta}_{EBP}^{k}$, $\hat{\theta}_{BP}^{k}$ and θ^{k} are the values of the EBP, BP and the predicted characteristic, respectively, in the *k*th iteration of the simulation study.

All results in this section are computed for $L=500$. In Figures 5 and 6 we consider different numbers of iterations because model 1 is more complex than model 2. The results for the simpler model 2 presented in Figure 6 tend to stabilize at $K=5000$. In the case of a more complex model 1 (see Figure 5), results for $K=5000$ are unstable especially in the case of the prediction of the median and the quartile skewness coefficient. To explain these results we should take into account two issues. Firstly, we assume that the number of EBP iterations $L=500$ is acceptable as shown in Section 5. Secondly, the results presented in Figure 5 are obtained based on one simulation study for an assumed number of Monte Carlo iterations K, which can but does not have to show possible instability of one specific result. Hence, the observed peaks for these two cases should be interpreted as a result of too small number of Monte Carlo iterations leading to possible instability of the results. The results for the more complex model (Figure 6) tend to stabilize at $K=15000$ iterations. What is more, in many cases MSEs ratios for the third subpopulation are close to 1, which is an argument for higher K if it is possible. The paths of improvement of the results are different for different prediction problems, however the generalization of the results is quite difficult due to the single execution of the simulation for each K. Predictors of some characteristics (i.e. the standard and the quartile deviations) tend to behave more stable than others, which may indicate a different strategy of the optimal choice of K for the specific simulation conditions like the considered characteristics. The complexity of the model has a significant impact on the simulation stability, which is opposite to the results presented in the two previous sections.

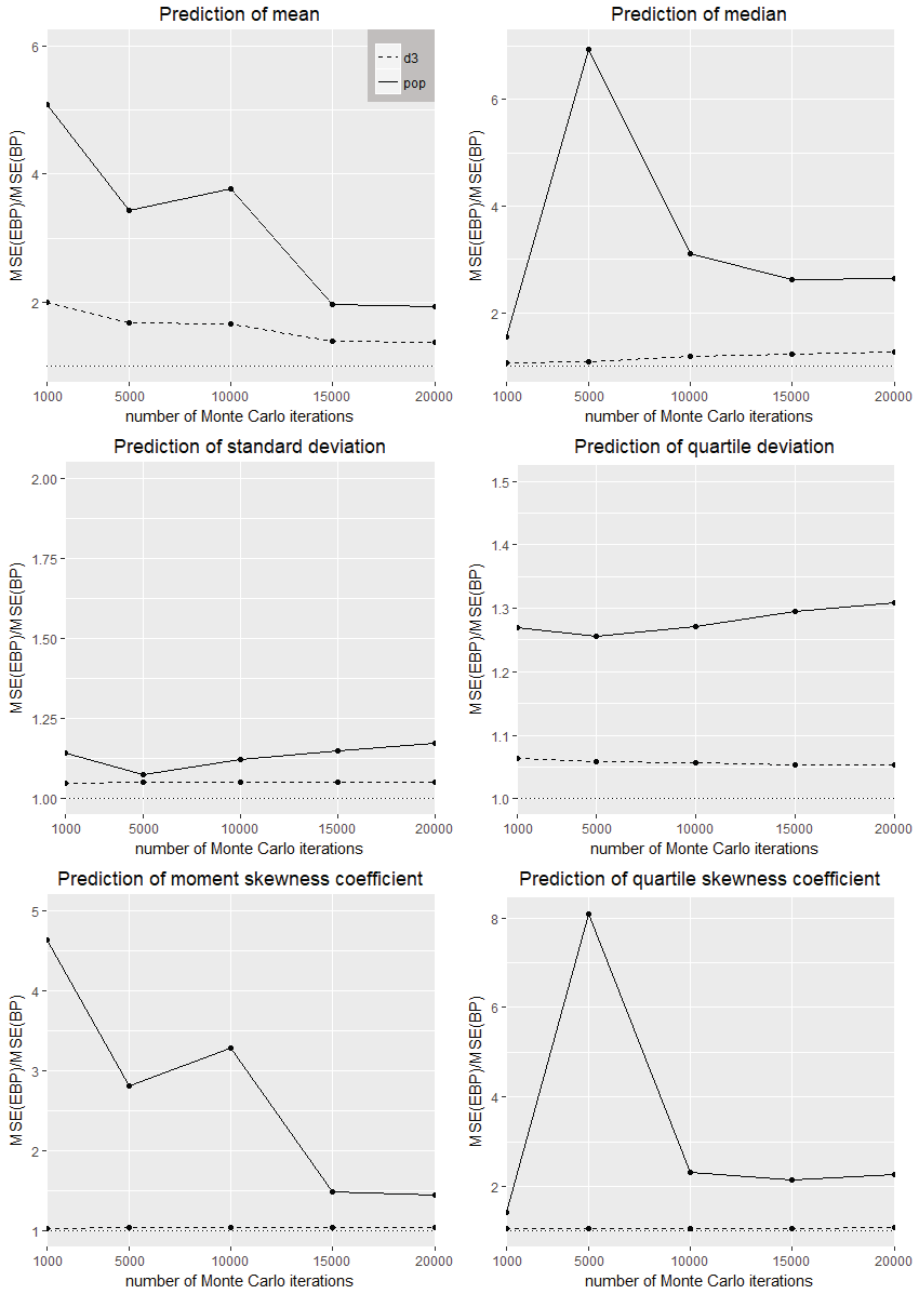


Figure 5. Ratios of MSE(EBP1) and MSE(BP1) of different population and subpopulation characteristics in the future period computed for different numbers of Monte Carlo iterations under model 1

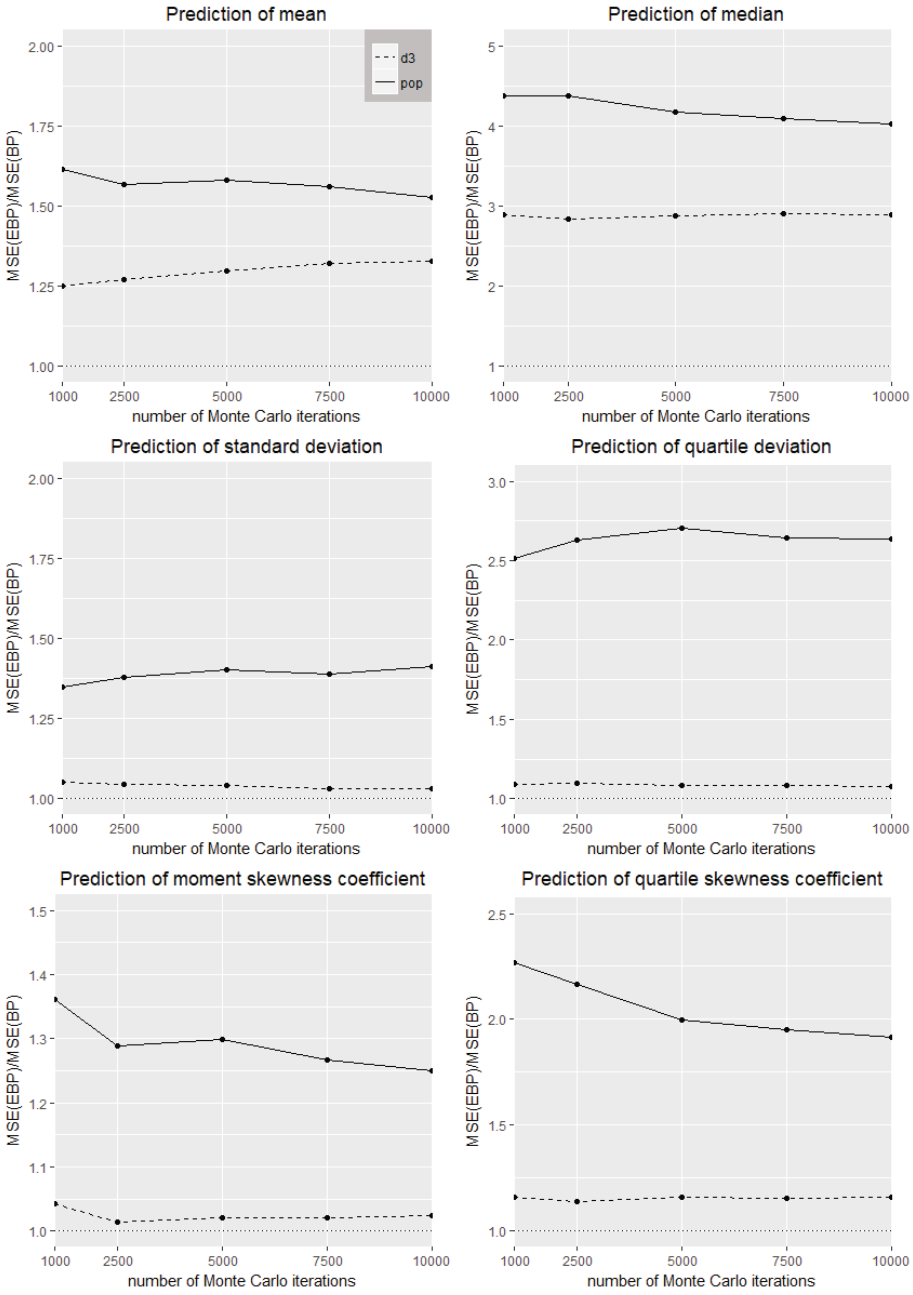


Figure 6. Ratios of MSE(EBP2) and MSE(BP2) of different population and subpopulation characteristics in the future period computed for different numbers of Monte Carlo iterations under model 2

8. Discussion

In this section we present possible generalizations of the proposed criteria, some alternatives and limitations of our procedure.

The CV, as well as the relative change of CVs, can be replaced by a more robust measure, e.g. based on quantiles (like the interquartile range) and the relative change of the chosen measure, respectively. It may be helpful especially in the case where two iterative algorithms are used at the same time as in the case of estimation of the MSE based on B bootstrap replications of the EBP approximated in L iterations (see Section 6), where the simulation variability resulting from the first procedure influences the results of the second procedure. What is more, in the case of consideration of highly volatile characteristics such as the standard deviation, more robust measure may be applied in practice, however in most cases the CV should be sufficient. The adequate measure can be determined by the researcher after the study of some boxplots presented in Figures 1 and 3.

The stopping role assumed in this paper to be the absolute or the relative difference of the appropriate measure of the simulation variability can also be changed. For example, we can assume that the required number of iterations is obtained (that the procedure should be stopped) if two distributions, represented by two adjacent boxplots in Figure 1 or in Figure 3, are the same, which is verified by the appropriate nonparametric test.

The drawback of our procedure results from the necessity of conducting the computations several times per one iteration number to obtain data represented by one boxplot. The alternative, to be developed and studied in further research, could be based on the idea of statistical quality control (e.g. control charts) where only one value is computed per one iteration number. In the statistical quality control it is checked when the monitored process becomes "not in control". In our case, we will have to check, based on the appropriate criteria, when the process becomes "in control" (becomes stable). Although in this approach the number of computations per one iteration will be one, we will have to increase the number of steps and replace, e.g. $L = 100, 200, 300, \ldots$ by $L = 10, 20, 30, \ldots$ but even though the total number of iterations will be smaller.

The methods considered in the paper are in practice highly dependent on the available time, overall complexity of the simulations and the available hardware. Furthermore, the sufficient improvement of the measures is a subjective case that is heavily dependent on the origin of the data (i.e. in the case of some medical simulations even small improvements can be very important). For the considered dataset the convergence of the CV computed for the EBP as well as the MSE estimator may vary, which provides additional difficulties in terms of generalization of the results.

9. Conclusion

We consider the problem of the stability of results in iterative procedures used for the computation of the empirical best predictor and its parametric bootstrap MSE estimator. We show the dependence between the number of iterations and the stability of iteratively obtained values of two predictors based on a simple and more complex model in the case of prediction of different future population and subpopulation characteristics. In the case of the EBP iterative algorithm and the parametric bootstrap procedure used to estimate the MSE we propose two methods of choosing the appropriate number of iterations. The first one is based on the maximum acceptable value of the CV of the results obtained several times for a given number of iterations. The second one is the stability criterion assessed based on the minimum relative decrease in the CV. In the case of Monte Carlo simulation studies we suggest two criteria based on the ratio of MSEs of the EBP and the BP. The number of Monte Carlo iterations should be controlled to obtain simulation ratios of the MSEs: stable and greater than one. All of the considerations are supported by real longitudinal economic data available at the United States Census Bureau website.

REFERENCES

ANDREWS, D. W. K., BUCHINSKY, M., (1997). On the number of bootstrap repetitions for bootstrap standard errors, confidence intervals, and tests, *Cowles Foundation Discussion Paper No. 1141R*, pp. 1–51.

ANDREWS, D. W. K., BUCHINSKY, M., (2000). A three-step method for choosing the number of bootstrap repetitions, *Econometrica*, Vol. 67, pp. 23–51.

ANDREWS, D. W. K., BUCHINSKY, M., (2001). Evaluation of a three-step method for choosing the number of bootstrap repetitions, *Journal of Econometrics,* Vol. 103, pp. 345–386.

BARBIERO, A., MECATTI, F., (2010). *Bootstrap algorithms for variance estimation in πPS* sampling, In: Mantovan, P., Secchi, P. (Eds.), Complex Data Modeling and Computationally Intensive Statistical Methods. Contributions to Statistics, Springer, Milano, pp. 57–69.

BERAN, R., (1997). Diagnosing Bootstrap Success, *Annals of the Institute of Statistical Mathematics*, Vol. 49, pp. 1–24.

BERG, E., CHANDRA, H., (2014). Small area prediction for a unit-level lognormal model, *Computational Statistics and Data Analysis*, Vol. 78, pp.159–175.

BOUBETA, M., LOMBARDÍA, M. J, MORALES, D., (2016). Empirical best prediction under area-level Poisson mixed models, *Test,* Vol. 25, pp. 548–569.

BOUBETA, M., LOMBARDÍA, M. J, MORALES, D., (2017). Poisson mixed models for studying the poverty in small areas, *Computational Statistics and Data Analysis*, Vol. 107, pp. 32–47.

BUTAR, B. F., LAHIRI, P., (2003). On measures of uncertainty of empirical Bayes small-area estimators, *Journal of Statistical Planning and Inference*, Vol. 112, pp. 63–76.

CHATTERJEE, S., LAHIRI, P. LI, H., (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models, *Annals of Statistics*, Vol. 36 (3), pp. 1221–1245.

DAS, S., HASLETT, S., (2019). A comparison of methods for poverty estimation in developing countries, *International Statistical Review*, DOI: 10.1111/insr.12314, available online: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12314.

DIALLO, M. S., (2014). *Small area estimation under skew-normal nested error models*, PhD diss., Carleton University.

DIALLO, M. S., RAO, J. N. K., (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors, *Scandinavian Journal of Statistics*, Vol. 2018, pp.1–25.

DAVISON, A. C., HINKLEY D. V., (1997). *Bootstrap Methods and their Application*, Cambridge University Press.

EFFRON, B., TIBSHIRANI, R., (1986), Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science*, Vol. 1(1), pp. 54–75.

ELBERS, CH., VAD DER WEIDE, R., (2014). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. *World Bank Group. Policy Research Working Paper* 6962, pp. 1–31.

GONZÁLEZ-MANTEIGA W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2008). Bootstrap mean squared error of small-area EBLUP, *Journal of Statistical Computation and Simulation*, Vol. 78(5), pp. 443–462.

GUADARRAMA, M., MOLINA, I., RAO, J. N. K., (2018). Small area estimation of general parameters under complex sampling designs, *Computational Statistics and Data Analysis,* Vol.121, pp. 20–40.

HALL, P., MAITI, T., (2006). On Parametric Bootstrap Methods for Small Area Prediction, *Journal of the Royal Statistical Society. Series B*, Vol. 68(2), pp. 221–238.

HALL, P., MARTIN, M. A., (1988). Exact convergence rate of bootstrap quantile variance estimator, *Probability Theory and Related Fields*, Vol. 80, pp. 261–268.

HOBZA, T., MORALES, D., (2016). Empirical best prediction under unit-level logit mixed models, *Journal of Official Statistics,* Vol. 32(3), pp. 661–692.

JIANG, J., (1996). REML estimation: asymptotic behavior and related topics, *Annals of Statistics*, Vol. 24 (1), pp. 255–286.

JIANG, J., (2003). Empirical best prediction for small-area inference based on generalized linear mixed models, *Journal of Statistical Planning and Inference,* Vol. 111, pp. 117–127.

JIANG, J., (2007). *Linear and Generalized Linear Mixed Models and Their Appliactions,* Springer, New York.

JIANG, J., LAHIRI, P., (2001). Empirical best prediction for small area inference with binary data, *Annals of the Institute of Statistical Mathematics*, Vol. 53(2), pp. 217–243.

JIANG, J., LAHIRI, P., (2006). Estimation of Finite Population Domain Means, *Journal of the American Statistical Association*, Vol. 101(473), pp. 301–311.

MARINO, M. F., RANALLI, M. G., SALVATI, N., ALFÒ, M., (2019). Semi-Parametric Empirical Best Prediction for small area estimation of unemployment indicators, *The Annals of Applied Statistics*, Vol. 13(2), pp. 1166–1197.

MOLINA, I., MARTIN, N., (2018). EBP under a nested error model with log transformation, *Annals of Statistics*, Vol. 46(5), pp. 1961–1993.

MOLINA, I., RAO, J. N. K., (2010). Small area estimation of poverty indicators, *The Canadian Journal of Statistics*, Vol. 38(3), pp. 369–385.

SINGH, K., (1981). On the Asymptotic Accuracy of Efron's Bootstrap, *The Annals of Statistics*, Vol. 9(6), pp. 1187–1195.

TZAVIDIS, N., ZHANG, L.-C., LUNA, A., SCHMID, T., ROJAS-PERILLA, N., (2018). From start to finish: a framework for the production of small area official statistics, *Journal of the Royal Statistical Society A*, Vol. 181(4), pp. 927–979.

VERBEKE, G., MOLENBERGHS, G., (2009). *Linear mixed models for longitudinal data,* Springer-Verlag, New York.

ZIMMERMANN, T., MÜNNICH, R., (2018). Small area estimation with a lognormal mixed model under informative sampling, *Journal of Official Statistics*, Vol. 34(2), pp. 523–542.

APPENDIX

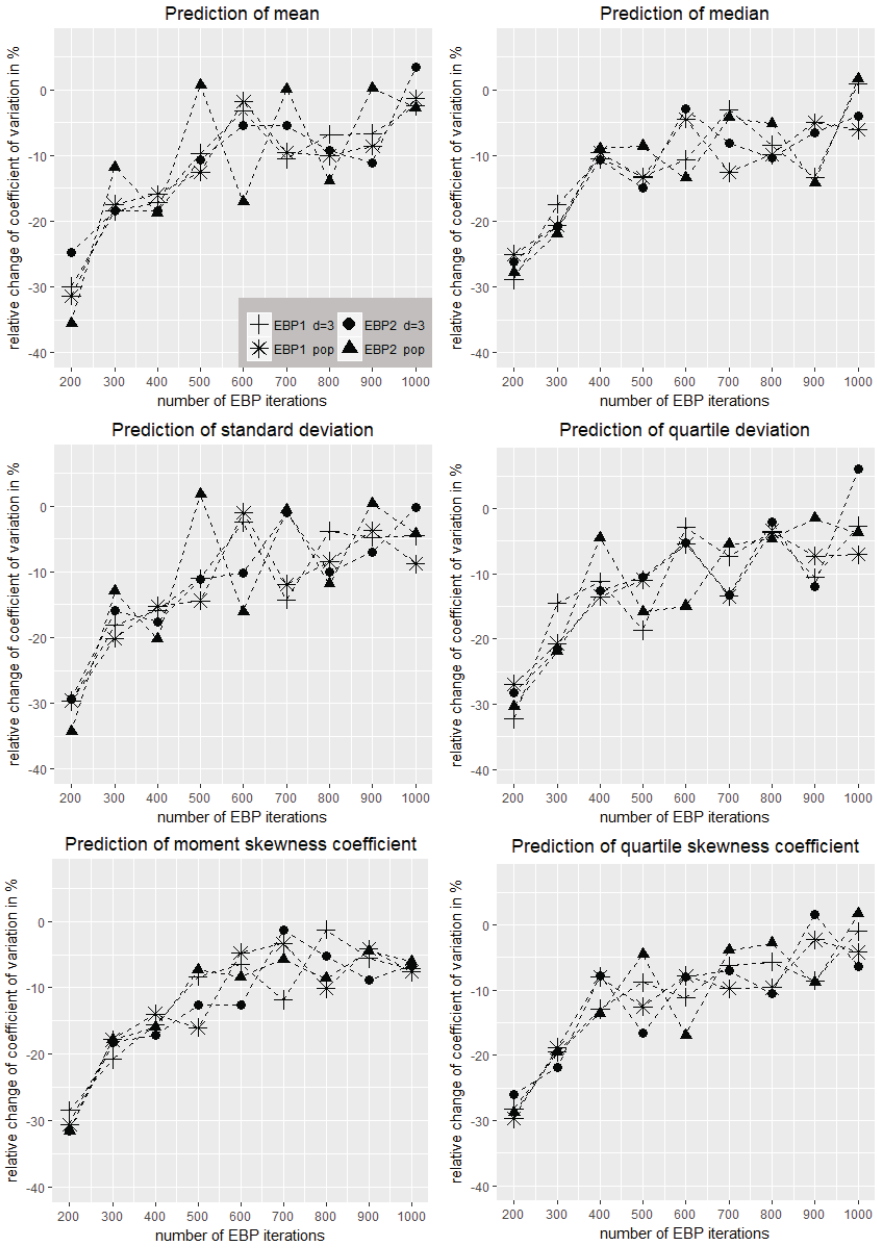


Figure 7. Relative changes of CVs of 500 values of EBP1 and EBP2 of different population and subpopulation characteristics computed as $100(CV_L - CV_{L-100}) / CV_{L-100}$ for $L = 200, 300, ..., 1000$

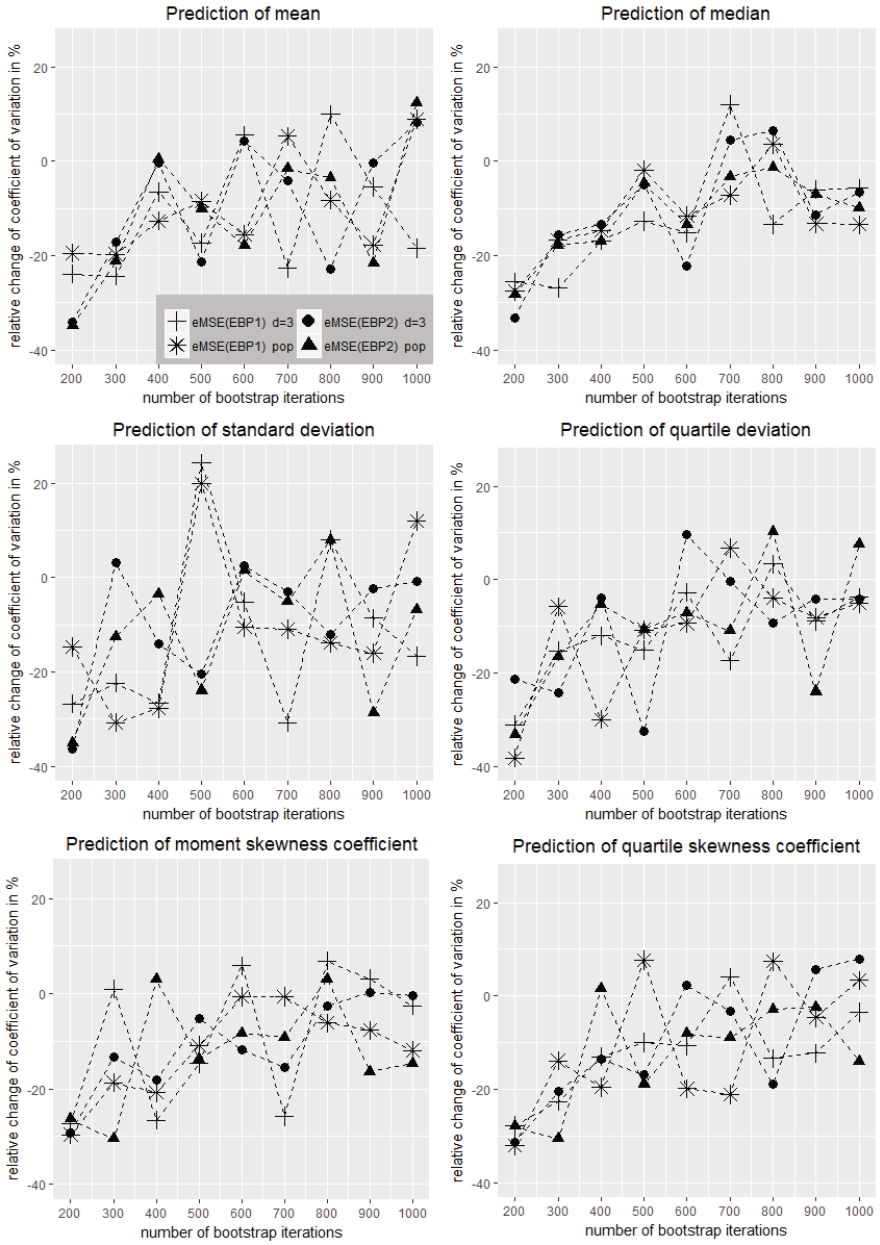


Figure 8. Relative changes of CVs of 100 values of parametric bootstrap MSE estimates of EBP1 and EBP2 of different population and subpopulation characteristics computed as $100(CV_B - CV_{B-100}) / CV_{B-100}$ for $B = 200, 300, ..., 1000$

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 61–87, DOI 10.21307/stattrans-2020-014
Submitted – 16.04.2019; accepted for publishing – 14.04.2020

How privacy may be protected in optional randomized response surveys

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

ABSTRACT

There are materials in literature about how privacy on stigmatizing features like alcoholism, history of tax-evasion, or testing positive in AIDS-related testing may be partially protected by a proper application of randomized response techniques (RRT). The paper demonstrates what amendments are necessary for this approach while applying optional RRTs covering qualitative characteristics, permitting a sampled respondent either to directly reveal sensitive data or choose a randomized response respectively with complementary probabilities. Only a few standard RRTs are illustrated in the text.

AMS subject classification: 62D05

Key words: protection of privacy, randomized response, sensitive issues, Warner and other techniques.

1. Introduction

Chaudhuri (2011) and Chaudhuri and Christofides (2013) in their books and Chaudhuri and Dihidar (2009), Chaudhuri and Saha (2005) and Chaudhuri, Christofides and Saha (2009) in their published papers have recounted details about how to protect privacy in randomized responses (RR) given out by the respondents following various RR devices.

We have reservations about only a few RR techniques because in a couple of text books and several authentic published review papers, only a few RR techniques are illustrated as we have done with no prejudice against the ones we omit to save space.

Here, we intend to investigate possibilities of protecting privacy in generating optional RR's covering qualitative stigmatizing issues. The optional RR (ORR)

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

technique was introduced by Chaudhuri and Mukerjee (1985). A large number of developments following Chaudhuri and Mukerjee (1985) approach were proposed by Gupta (2001), Gupta et al. (2002), Pal (2008) and many others. Subsequent developments are due to Arnab (2004), Chaudhuri and Saha (2005), Saha (2007), Huang (2008), Arnab and Rueda (2016) among others with slight differences in approaches. As we see, in ORR a sampled person is offered an option either to (i) report directly whether he/she bears a stigmatizing feature, say A (which may mean alcoholism or testing HIV positive, etc.) or (ii) give out an RR adopting a device offered and explained to him/her. The option (i) may be implemented with an unknowable probability and (ii) with the complementary probability. How to implement (i) or (ii) may be clearly explained to the respondent who may or may not divulge which of these options is actually applied. Different ORR techniques are described in this article.

In the cases of RR's, it is observed that privacy is protected only for specific parametric combinations in the RR devices and protection leads to loss of control in achieving accuracy in estimation of the population proportion of people bearing sensitive features. Such features will be seen in what follows with optional RR situations as well. But certain other striking possibilities are revealed below with optional RR's (ORR) rather than with compulsory RR's (CRR). Details are shown in Sections 2 and 3 below. Section 4 presents some numerical findings, through simulation.

2. Certain basics for protection of privacy in general sampling

Let $U = (1, 2, \ldots, N)$ denote a finite population of units. On drawing a sample according to a general sampling design P, the selected units are approached with a request to provide ORR's in order to estimate the proportion of the population units bearing a sensitive characteristic A, say.

Let, for a person labelled i, L_i be the unknowable prior probability that i bears A and $L_i(R)$ denote the posterior probability that given the RR or DR denoted R, the respondent bears A. Following Chaudhuri, Christofides and Saha (2009), the literature considers for a measure of jeopardy inherent in the response R the quantity

$$J_i(R) = \frac{L_i(R)/L_i}{(1-L_i(R))/(1-L_i)}$$

assuming the denominator is non-zero, with the RR device parameters rightly chosen.

Let, for a general ORR device, $c_i (0 < c_i < 1 \forall i \in U)$ be an unknowable probability that the i^{th} person chooses to answer directly without divulging this secret to the enquirer. Further, let for i,

I_i = the DR, with probability c_i

= the RR for a specified device with probability (1- c_i) of course.

The investigator is to explain to a respondent a formal way to implement choosing such an undisclosed c_i and 1- c_i to be the probability of giving a DR and respectively an RR with no option to change it for an alternative RR device. For example, c_i may be fixed (without disclosing to the investigator) as $\frac{13}{100}$ on choosing a 2-digited random number from 01, ..., 13 leaving the rest namely 14, ...,99,00 for giving out an RR.

Warner's RRT demands from a chosen person i a response

$$R_i = y_i \text{ with probability } p\,(0 < p < 1,\ p \neq \frac{1}{2}),$$
$$= 1 - y_i \text{ with probability } 1\text{-}p$$

if i chooses a card marked A and bears the stigmatizing feature A or chooses a card marked the complement of A (A^c) from a pack of cards with a proportion marked A and the rest marked the complement of $A(A^c)$ and

$$y_i = 1 \text{ if } i \text{ bears } A$$
$$= 0, \text{if } i \text{ bears } A^c.$$

Then, for this RRT due to Warner (1965) the expected value of R_i is

$$E(R_i) = py_i + (1-p)(1-y_i) = (1-p) + (2p-1)y_i \qquad \text{(I)}$$

for every i in U.

For the ORR technique instead for Warner's RRT the ORR is

$$OR_i = y_i \text{ with probability } c_i \text{ in the closed interval from 0 to 1}$$
$$= R_i \text{ with probability } (1-c_i)$$

$$\text{Then, } E(OR_i) = c_i y_i + (1-c_i)[(1-p) + (2p-1)y_i]$$
$$= (1-c_i)(1-p) + [c_i + (1-c_i)(2p-1)]\,y_i \qquad \text{(II)}$$

Clearly, if c_i equals zero, (II) matches (I) and if c_i differs from 0, (II) differs from (I) as well.

For simplicity let the response be either 'Yes' or 'No' only. We may write, applying Bayes' theorem, writing A^c as the complement of A,

$$\text{Prob}(A \mid Yes) = \frac{L_i \,\text{Pr ob}(Yes \mid A)}{L_i \,\text{Pr ob}(Yes \mid A) + (1-L_i)\,\text{Pr ob}(Yes \mid A^c)}$$

on supposing that Warner's RR device in Chaudhuri's (2001) form is employed.

Defining $y_i = 1$ if i bears A and 0 if i does not bear A, we may work out

$$\text{Prob}(Yes \mid A) = c_i y_i + (1-c_i)py_i$$

$$= p + c_i(1-p) \text{ since } y_i = 1$$

and $\text{Prob}(Yes \mid A^c) = (1-c_i)(1-p)(1-y_i)$

$$= (1-p)(1-c_i) \text{ since } y_i = 0$$

Hence, it follows that

$$\text{Prob}(A \mid Yes) = \frac{L_i[p + c_i(1-p)]}{L_i[p + c_i(1-p)] + (1-L_i)(1-p)(1-c_i)}$$

$$= \frac{L_i[p + (1-p)c_i]}{pL_i + (1-p)[1 - L_i(1-c_i)]}$$

and $J_i(1) = \dfrac{L_i(1)/L_i}{(1-L_i(1))/(1-L_i)}$.

With a little algebra,

$$1 - L_i(1) = \frac{(1-p)(1-L_i)}{pL_i + (1-p)[1-L_i(1-c_i)]};$$

so,

$$J_i(1) = \frac{L_i(1)}{1-L_i(1)} \frac{1-L_i}{L_i}$$

$$= \frac{p + c_i(1-p)}{(1-p)(1-c_i)}. \qquad (1)$$

Again,

$$J_i(0) = \frac{L_i(0)/L_i}{(1-L_i(0))/(1-L_i)}.$$

Now,

$$\text{Prob}(A \mid No) = \frac{L_i \text{Prob}(No \mid A)}{L_i \text{Prob}(No \mid A) + (1-L_i)\text{Prob}(No \mid A^c)}$$

$$\text{Prob}(No \mid A) = c_i(1-y_i) + (1-c_i)(1-p)$$

$$= (1-c_i)(1-p) \quad \text{since } y_i = 1;$$

$$\text{Prob}(No \mid A^c) = c_i(1-y_i) + (1-c_i)p$$

$$= c_i + p(1-c_i) \quad \text{since } y_i = 0.$$

So,

$$L_i(0) = \frac{L_i(1-c_i)(1-p)}{L_i(1-c_i)(1-p)+(1-L_i)[c_i+(1-c_i)p]}$$

$$1-L_i(0) = \frac{(1-L_i)[c_i+(1-c_i)p]}{(1-L_i)[c_i+(1-c_i)p]+L_i(1-c_i)(1-p)};$$

since $L_i(0) = \dfrac{L_i \Pr ob(No/\mathrm{A})}{L_i \Pr ob(No/\mathrm{A})+(1-L_i)\Pr ob(No/\mathrm{A}^c)}$,

so,

$$J_i(0) = \frac{L_i(0)/L_i}{(1-L_i(0))/(1-L_i)} = \frac{(1-c_i)(1-p)}{c_i+(1-c_i)p} = \frac{(1-c_i)(1-p)}{p+c_i(1-p)}. \quad (2)$$

Hence,

$$J_i(1)\times J_i(0) = (1)\times(2) = 1$$

Thus, our proposed measure of jeopardy is $\overline{J}_i \equiv$ the G.M. of $J_i(1)$ and $J_i(0)$ and this carries over for every *i* as $\overline{J}_i = 1$.

Ensuring privacy protection is not enough. The estimation of the variance of the estimator employed is also a crucial requirement. So, adjustments in the RRT's are needed. Thus, in employing Warner's RRT not just one RR is adequate; two independent RR's are needed when the ORR technique is to be employed by Warner's RRT allowing options for DR's. This is elaborated in Section 3.

3. Optional Randomized Response Technique with two independent randomized responses

The person labelled i is requested to give out two ORR's independently with different known RR device probabilities. Denoting the responses as R and R', the posterior probability and the measure of jeopardy may be written as $L_i(R,R')$ and $J_i(R,R')$ respectively, corresponding to the i^{th} person's response (R,R').

Now, applying Bayes' theorem,

$$\Pr ob(A|(R,R')) = \frac{L_i \Pr ob(R|A)\Pr ob(R'|A)}{L_i \Pr ob(R|A)\Pr ob(R'|A)+(1-L_i)\Pr ob(R|A^c)\Pr ob(R'|A^c)} \quad (3)$$

as the responses are independent for every person,

and the response specific jeopardy measure for the ith person

$$J_i(R,R') = \frac{L_i(R,R')/L_i}{(1-L_i(R,R'))/(1-L_i)} \tag{4}$$

indicates the risk of divulging the respondent's status due to his/her specific response (R,R'). Chaudhuri et al. (2009) preferred an average measure. Here, we propose geometric mean as an average measure instead of arithmetic mean, earlier suggested by Chaudhuri Christofides and Saha (2009). A Geometric Mean (GM) in lieu of Arithmetic Mean (AM) is proposed to achieve an algebraic simplicity. Thus, the measure of jeopardy for the ith person is

$$\overline{J}_i = \text{G.M of } J_i(R,R') \ \forall R,R' \ . \tag{5}$$

In this section, we discuss the response specific measure of jeopardy in ORR technique and our proposed measure combining all the response specific jeopardy measures for qualitative characteristics.

Although $J_i(R,R')$ depends on unknown probability c_i, it can be shown in the later sections that the measure of jeopardy $\overline{J}_i$ is free from c_i.

3.1. ORR using Warner's (1965) RR model

Suppose the sampled person labelled i is directed to respond his/her true value of the specific stigmatizing attribute or by Warner's RR device. A box with identical cards with A or A^c in proportions $p_1 : 1-p_1$ $(0<p_1<1,\ p_1 \neq \frac{1}{2})$ is given to the respondent. He/she is requested to draw a card and without divulging the card-type drawn he/she is to truthfully say his/her outcome if the card type drawn matches or not his/her characteristic. The whole process is repeated one more time independently but with different Warner's RR device with another similar box - cards marked by A or A^c, which are in proportions $p_2 : 1-p_2$ $(0<p_2<1,\ p_2 \neq \frac{1}{2})$.

Thus, the independent Optional randomized responses for i th $(i = 1,2,...,N)$ person are Z_i and Z_i'.

Here, $Z_i = y_i$, with unknown probability c_i

= the Warner's RR, with unknown probability $1-c_i$, using first box

and

$Z_i' = y_i$, with unknown probability c_i

= the Warner's RR, with unknown probability $1-c_i$, using another box.

Here,

$y_i = 1$ if the person bears the sensitive characteristic

= 0, else.

Note that the investigator's instruction is to keep the same c_i for both z and z'.

Then, denoting RR based expectations and variances as E_R and V_R we may write

$$E_R(Z_i) = c_i y_i + (1-c_i)[p_1 y_i + (1-p_1)(1-y_i)]$$

$$E_R(Z_i') = c_i y_i + (1-c_i)[p_2 y_i + (1-p_2)(1-y_i)]$$

Thus, an unbiased estimator of y_i is $r_i = \dfrac{(1-p_2)Z_i - (1-p_1)Z_i'}{p_1 - p_2}$, $p_1 \neq p_2$ and an unbiased estimator of the variance $V_R(r_i)$ is $v_i = \dfrac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(Z_i - Z_i')^2$. The details of the proof is given in Appendix 1.This variance estimator form is slightly different from Chaudhuri and Dihidar's (2009).We prefer this form as it is a function of two independent responses.

The possible responses for each individual in the above method are (1, 1) (0, 0) (1, 0) and (0, 1).

Note that a different response 1 or 0 may come from the same person for the first and second trials; of course it does not matter because it may reveal that a person may have opted for an RR rather than a DR; this does not reveal the person's sensitive feature.

Suppose the response of i^{th} labelled person is (1, 1). Then, using the equation (3) we get

$$\frac{L_i(1,1)}{L_i} = \frac{P(Z_i = 1 \mid y_i = 1)P(Z_i' = 1 \mid y_i = 1)}{L_i P(Z_i = 1 \mid y_i = 1)P(Z_i' = 1 \mid y_i = 1) + (1-L_i)P(Z_i = 1 \mid y_i = 0)P(Z_i' = 1 \mid y_i = 0)}$$

$$= \frac{\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\}}{L_i\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\} + (1-L_i)\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\}}$$

$$\frac{1-L_i(1,1)}{1-L_i} = \frac{\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\}}{L_i\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\} + (1-L_i)\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\}}$$

So, from equation (4) the response (1,1) - specific jeopardy measure is

$$J_i(1,1) = \frac{L_i(1,1)/L_i}{(1-L_i(1,1))/(1-L_i)} = \frac{\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\}}{\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\}} \tag{6}$$

If the i^{th} person's response is (0,0) then we may write

$$\frac{L_i(0,0)}{L_i} = \frac{P(Z_i = 0 \mid y_i = 1)P(Z_i = 0 \mid y_i = 1)}{L_i P(Z_i = 0 \mid y_i = 1)P(Z_i = 0 \mid y_i = 1) + (1 - L_i)P(Z_i = 0 \mid y_i = 0)P(Z_i = 0 \mid y_i = 0)}$$

$$= \frac{\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\}}{L_i\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\} + (1-L_i)\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\}}$$

$$\frac{1 - L_i(0,0)}{1 - L_i} = \frac{\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\}}{L_i\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\} + (1-L_i)\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\}}.$$

So, the response (0,0) - specific jeopardy measure is

$$J_i(0,0) = \frac{L_i(0,0)/L_i}{(1 - L_i(0,0))/(1 - L_i)} = \frac{\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\}}{\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\}}. \tag{7}$$

For the response (1,0), the corresponding posterior probability $L_i(1,0)$ and Jeopardy measure $J_i(1,0)$ may be expressed as

$$L_i(1,0) = \frac{L_i\{c_i + (1-c_i)p_1\}\{(1-c_i)(1-p_2)\}}{L_i\{c_i + (1-c_i)p_1\}\{(1-c_i)(1-p_2)\} + (1-L_i)\{(1-c_i)(1-p_1)\}\{c_i + (1-c_i)p_2\}}$$

$$J_i(1,0) = \frac{\{c_i + (1-c_i)p_1\}\{(1-c_i)(1-p_2)\}}{\{(1-c_i)(1-p_1)\}\{c_i + (1-c_i)p_2\}}. \tag{8}$$

Similarly, for the response (0,1), the posterior probability is

$$L_i(0,1) = \frac{L_i\{(1-c_i)(1-p_1)\}\{c_i + (1-c_i)p_2\}}{L_i\{(1-c_i)(1-p_1)\}\{c_i + (1-c_i)p_2\} + (1-L_i)\{c_i + (1-c_i)p_1\}\{(1-c_i)(1-p_2)\}}$$

and the response specific measure of jeopardy is

$$J_i(0,1) = \frac{\{(1-c_i)(1-p_1)\}\{c_i + (1-c_i)p_2\}}{\{c_i + (1-c_i)p_1\}\{(1-c_i)(1-p_2)\}}. \tag{9}$$

Thus, our proposed measure of jeopardy is the geometric mean of all response specific jeopardy measures (6), (7), (8) and (9), which is exactly 1 for each and every individual. If $p_1 \to p_2$, responses of every individual are well protected but estimate of the variance tends to be infinite. Yet the overall measure does not reveal the status of the respondent.

3.2. ORR using Greenberg et al.'s (1969) unrelated question RR model

The ORR technique with an unrelated question model is same as the above discussed technique except the RR device. Here, the RR device is Greenberg et al.'s unrelated question model (1969) instead of Warner's model. In this RR, two boxes contain cards marked as A, the stigmatizing attribute or B, the innocuous attribute. The attribute A, is unrelated to the attribute B . The two types of cards are mixed with different known proportions say p_1 and $(1-p_1)$ and p_2 and $(1-p_2)$ in box 1 and box 2 respectively. Each respondent is requested to draw two cards independently from box 1 and box 2 respectively and report according to the above device.

So, the Optional randomized response for i^{th} person is

$Z_i = y_i$, with unknown probability c_i

= the Greenberg et al.'s RR, with unknown probability $1-c_i$, using box 1

and

$Z_i' = y_i$, with unknown probability c_i

= the Greenberg et al.'s RR, with unknown probability $1-c_i$, using box 2.

Defining

$x_i = 1$ if the person bears the innocuous character *B*

$= 0$ if the person bears the innocuous character $B^{C,}$ the complement of B,

we may write,

$$P(Z_i = 1) = c_i y_i + (1-c_i)[p_1 y_i + (1-p_1)x_i] \text{ and}$$

$$P(Z_i = 0) = c_i(1-y_i) + (1-c_i)[p_1(1-y_i) + (1-p_1)(1-x_i)].$$

Hence, it follows that

$$E_R(Z_i) = c_i y_i + (1-c_i)[p_1 y_i + (1-p_1)x_i].$$

Similarly,

$$E_R(Z_i') = c_i y_i + (1-c_i)[p_2 y_i + (1-p_2)x_i].$$

Thus, an unbiased estimator of y_i under the above model is $r_i = \dfrac{(1-p_2)Z_i - (1-p_1)Z_i'}{p_1 - p_2}$ taking $p_1 \neq p_2$ and an unbiased estimator of the variance $V_R(r_i)$ is $v_i = \dfrac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(Z_i - Z_i')^2$ since $p_1 \neq p_2$. The proof is given in Appendix 2.

The necessary conditional probabilities are shown below to calculate posterior probabilities and response specific jeopardy measures defined as in the equation (3).

Now, $P(Z_i = 1 | y_i = 0) = (1-c_i)(1-p_1)x_i = (1-c_i)(1-p_1)$, as the situation arises if the response of the i^{th} individual is 1 but the true value of the sensitive characteristic is zero. This is possible only if the respondent chooses RR device and responds to the question regarding the innocuous attribute *B* due to the assumption that the respondents provide true response. So, $x_i = 1$ is obvious.

With the same line of reasoning, we get $P(Z_i = 0 | y_i = 1) = (1-c_i)(1-p_1)$.

As we know, $P(A|B) + P(A^C | B) = 1$.

Clearly, $P(Z_i = 1 | y_i = 1) = 1 - P(Z_i = 0 | y_i = 1) = c_i + (1-c_i)p_1$.

Similarly, $P(Z_i = 0 | y_i = 0) = 1 - P(Z_i = 1 | y_i = 0) = c_i + (1-c_i)p_1$.

Proceeding as described in 3.1, their response specific jeopardy measures are

$$J_i(1,1) = \frac{L_i(1,1)/L_i}{(1-L_i(1,1))/(1-L_i)} = \frac{\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\}}{\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\}} \tag{10}$$

$$J_i(0,0) = \frac{\{(1-c_i)(1-p_1)\}\{(1-c_i)(1-p_2)\}}{\{c_i + (1-c_i)p_1\}\{c_i + (1-c_i)p_2\}} \tag{11}$$

$$J_i(1,0) = \frac{\{c_i + (1-c_i)p_1\}\{(1-c_i)(1-p_2)\}}{\{(1-c_i)(1-p_1)\}\{c_i + (1-c_i)p_2\}} \tag{12}$$

$$J_i(0,1) = \frac{\{(1-c_i)(1-p_1)\}\{c_i + (1-c_i)p_2\}}{\{c_i + (1-c_i)p_1\}\{(1-c_i)(1-p_2)\}}. \tag{13}$$

Now, our proposed measure of jeopardy by the equation (5) is the geometric mean (G.M) of the above response specific jeopardy measures. Here, the GM is $\bar{J}_i = \{J_i(1,1) \times J_i(0,0) \times J_i(1,0) \times J_i(0,1)\}^{1/4} = \{(10) \times (11) \times (12) \times (13)\}^{1/4} = 1$, whatever be the value of the selection probabilities of a card from RR devices. Here p_1 cannot tend to p_2, otherwise variance estimate will be infinite.

3.3. ORR using Forced response model

In ORR with forced response model the sampled person labelled i is requested to give out the truthful response y_i with unknown probability c_i or the forced RR response with probability $1-c_i$. In forced RR device, the person is offered two boxes with three types of cards marked as “Yes”, “No” and “Honest Response” but they are in different proportions. For the first box, “Yes”, “No” and “Honest Response” are

in proportions p_1, p_2 and $1-p_1-p_2$ $(0<p_1,\ p_2<1)$ respectively. For the second box, they are in proportions p_3, p_4 and $1-p_3-p_4$ $(0<p_3,\ p_4<1)$ respectively. But we should add the restriction $p_1p_4=p_2p_3$ on the known probabilities p_1, p_2, p_3, p_4 to derive an unbiased estimator for the proportion with stigmatizing attribute A.

So, the Optional randomized response for i^{th} person is

$Z_i = y_i$, with unknown probability c_i

= the Forced RR, with unknown probability $1-c_i$, using the first box

and

$Z_i' = y_i$, with unknown probability c_i

= the Forced RR, with unknown probability $1-c_i$, using the second box.

Then,

$$P(Z_i=1)=c_iy_i+(1-c_i)[(1-p_1-p_2)y_i+p_1]$$
$$P(Z_i=0)=c_i(1-y_i)+(1-c_i)[(1-p_1-p_2)(1-y_i)+p_2]$$
$$P(Z_i'=1)=c_iy_i+(1-c_i)[(1-p_3-p_4)y_i+p_3]$$
$$P(Z_i'=0)=c_i(1-y_i)+(1-c_i)[(1-p_3-p_4)(1-y_i)+p_4].$$

The unbiased estimator of y_i is $r_i=\dfrac{p_3Z_i-p_1Z_i'}{p_3-p_1}$, $(p_3\neq p_1)$ and the unbiased estimator of the variance $V_R(r_i)$ is $v_i=\dfrac{p_1p_3(Z_i-Z_i')^2}{(p_3-p_1)^2}$. It is proved in Appendix 2.

Then, the posterior probabilities and the response specific jeopardy measures for different responses are shown below.

$$L_i(1,1)=\frac{L_i\{c_i+(1-c_i)(1-p_2)\}\{c_i+(1-c_i)(1-p_4)\}}{L_i\{c_i+(1-c_i)(1-p_2)\}\{c_i+(1-c_i)(1-p_4)\}+(1-L_i)\{(1-c_i)p_1\}\{(1-c_i)p_3\}}$$

$$L_i(0,0)=\frac{L_i\{(1-c_i)p_2\}\{(1-c_i)p_4\}}{L_i\{(1-c_i)p_2\}\{(1-c_i)p_4\}+(1-L_i)\{c_i+(1-c_i)(1-p_1)\}\{c_i+(1-c_i)(1-p_3)\}}$$

$$L_i(1,0)=\frac{L_i\{c_i+(1-c_i)(1-p_2)\}\{(1-c_i)p_4\}}{L_i\{c_i+(1-c_i)(1-p_2)\}\{(1-c_i)p_4\}+(1-L_i)\{(1-c_i)p_1\}\{c_i+(1-c_i)(1-p_3)\}}$$

$$L_i(0,1)=\frac{L_i\{(1-c_i)p_2\}\{c_i+(1-c_i)(1-p_4)\}}{L_i\{(1-c_i)p_2\}\{c_i+(1-c_i)(1-p_4)\}+(1-L_i)\{c_i+(1-c_i)(1-p_1)\}\{(1-c_i)p_3\}}$$

$$J_i(1,1)=\frac{\{c_i+(1-c_i)(1-p_2)\}\{c_i+(1-c_i)(1-p_4)\}}{\{(1-c_i)p_1\}\{(1-c_i)p_3\}} \tag{14}$$

$$J_i(0,0)=\frac{\{(1-c_i)p_2\}\{(1-c_i)p_4\}}{\{c_i+(1-c_i)(1-p_1)\}\{c_i+(1-c_i)(1-p_3)\}} \tag{15}$$

$$J_i(1,0)=\frac{\{c_i+(1-c_i)(1-p_2)\}\{(1-c_i)p_4\}}{\{(1-c_i)p_1\}\{c_i+(1-c_i)(1-p_3)\}} \tag{16}$$

$$J_i(0,1)=\frac{\{(1-c_i)p_2\}\{c_i+(1-c_i)(1-p_4)\}}{\{c_i+(1-c_i)(1-p_1)\}\{(1-c_i)p_3\}}. \tag{17}$$

Here, the proposed jeopardy measure for the i th person is the G.M of $J_i(1,1), J_i(0,0), J_i(1,0), J_i(0,1)$.

It is $$\bar{J}_i=[\frac{p_2^2p_4^2\{c_i+(1-c_i)(1-p_2)\}^2\{c_i+(1-c_i)(1-p_4)\}^2}{p_1^2p_3^2\{c_i+(1-c_i)(1-p_1)\}^2\{c_i+(1-c_i)(1-p_3)\}^2}]^{1/4}$$

$$=\frac{p_2}{p_1}[\frac{\{c_i+(1-c_i)(1-p_2)\}\{c_i+(1-c_i)(1-p_4)\}}{\{c_i+(1-c_i)(1-p_1)\}\{c_i+(1-c_i)(1-p_3)\}}]^{1/2} \tag{17.1}$$

Thus, the GM need not always be unity as is also the case in (18.1) and later and also in (19) below.

Thus, the measure of jeopardy depends on the selection of p_1, p_2, p_3 and p_4 Here $\bar{J}_i\to 1$ if $p_1\to p_2$ and $p_3\to p_4$.

3.4. ORR using Kuk's (1990) RR model

Let the sampled person be instructed to record his/her true value of bearing the sensitive attribute A using the ORR device adopting the RR device or direct response. The respondent is directed to draw k (with replacement) number of cards from one of two boxes having red and black cards in different proportions $(\theta_1:1-\theta_1$ and $\theta_2:1-\theta_2)$ with $0<\theta_1,\theta_2<1$ and requested to report the number $\frac{(\frac{f_i}{k}-\theta_2)}{\theta_1-\theta_2}$, f_i being the number of red cards out of k cards if the sampled person i decides to adopt Kuk's RR device. Cards should be drawn from the first box if the respondent bears the sensitive attribute, otherwise the cards are drawn from the second box having the

proportion of red and black cards in proportions $(\theta_2 : 1-\theta_2)$ without disclosing which box is used to draw the cards.

So, the ORR response for i^{th} person is

$Z_i = y_i$ with the unknown probability c_i

$= \dfrac{\dfrac{f_i}{k} - \theta_2}{\theta_1 - \theta_2}$ with the unknown probability $1-c_i$, and

$E_R(f_i) = k[y_i\theta_1 + (1-y_i)\theta_2]$

leading to $E_R(Z_i) = c_i y_i + (1-c_i)E_R\left(\dfrac{\dfrac{f_i}{k} - \theta_2}{\theta_1 - \theta_2}\right) = y_i$.

To estimate the variance, the process is repeated one more time and the response variable Z_i' is the same as above but the number of red cards is denoted by f_i'. So, the final unbiased estimator of y_i is $\dfrac{Z_i + Z_i'}{2}$ and the related unbiased variance estimator is $v_i = \dfrac{1}{4}(Z_i - Z_i')^2$ following Chaudhuri et al. (2013, 2016).

The posterior probability can be defined as

$$L_i(f_i, f_i') = \frac{L_i P(Z_i = f_i \mid y_i = 1)P(Z_i' = f_i' \mid y_i = 1)}{L_i P(Z_i = f_i \mid y_i = 1)P(Z_i' = f_i' \mid y_i = 1) + (1-L_i)P(Z_i = f_i \mid y_i = 0)P(Z_i' = f_i' \mid y_i = 0)}$$

$$= \frac{L_i \psi_{1i}\psi_{1i}'}{L_i \psi_{1i}\psi_{1i}' + (1-L_i)\psi_{2i}\psi_{2i}'}$$

where $\psi_{1i} = c_i I_i + (1-c_i)\theta_1^{f_i}(1-\theta_1)^{k-f_i}$ with the indicator function I_i defining as $I_i = 1$ if $f_i = 1$ and 0 otherwise and $\psi_{2i} = c_i I_i' + (1-c_i)\theta_2^{f_i}(1-\theta_2)^{k-f_i}$ with another indicator function I_i' defined as $I_i' = 1$ if $f_i = 1$ and 0 otherwise.

Similarly, $\psi_{1i}' = c_i I_i + (1-c_i)\theta_1^{f_i'}(1-\theta_1)^{k-f_i'}$ and $\psi_{2i}' = c_i I_i' + (1-c_i)\theta_2^{f_i'}(1-\theta_2)^{k-f_i'}$ with two indicator functions defined as just above, and the response specific jeopardy measure is

$$J_i(f_i, f_i') = \frac{\psi_{1i}\psi_{1i}'}{\psi_{2i}\psi_{2i}'} = J_i(f_i).J_i(f_i') \tag{18}$$

where $J_i(f_i) = \frac{\psi_{1i}}{\psi_{2i}}$; $J_i(f_i') = \frac{\psi'_{1i}}{\psi'_{2i}}$ for all $f_i, f_i' = 0,1,2,...k$

and

$$\bar{J}_i = (\prod_{\forall f_i, f_i'} J_i(f_i, f_i'))^{1/k+1} = (\prod_{\forall f_i, f_i'} J_i(f_i) J_i(f_i'))^{1/k+1} = (\prod_{\forall f_i} J_i(f_i))^{1/(k+1)}. \quad (18.1)$$

Consequently, $J_i(0) = \frac{(1-c_i)(1-\theta_1)^k}{c_i + (1-c_i)(1-\theta_2)^k}$ and

$J_i(1) = \frac{c_i + (1-c_i)\theta_1(1-\theta_1)^{k-1}}{(1-c_i)\theta_2(1-\theta_2)^{k-1}}$ do not tend to 1 whatever the choice of θ_1, θ_2

But $J_i(f_i) = \frac{(1-c_i)\theta_1^{f_i}(1-\theta_1)^{k-f_i}}{(1-c_i)\theta_2^{f_i}(1-\theta_2)^{k-f_i}} = (\frac{\theta_1}{\theta_2})^{f_i}(\frac{1-\theta_1}{1-\theta_2})^{k-f_i}$, for all $f_i, f_i' = 2,3..k$.

And it tends to 1 if $\theta_1 \to \theta_2$

$$\bar{J}_i = [J_i(0).J_i(1).J_i(2).....J_i(k-1).J_i(k)]^{1/k+1}$$

$$= [\frac{(1-c_i)(1-\theta_1)^k}{c_i + (1-c_i)(1-\theta_2)^k} \frac{c_i + (1-c_i)\theta_1(1-\theta_1)^{k-1}}{(1-c_i)\theta_2(1-\theta_2)^{k-1}} . \frac{\theta_1^2(1-\theta_1)^{k-2}}{\theta_2^2(1-\theta_2)^{k-2}} \frac{\theta_1^3(1-\theta_1)^{k-3}}{\theta_2^3(1-\theta_2)^{k-3}} \cdots \frac{\theta_1^k}{\theta_2^2}]^{1/k+1}$$

$$= [\frac{c_i + (1-c_i)\theta_1(1-\theta_1)^{k-1}}{c_i + (1-c_i)(1-\theta_2)^k} \frac{(1-c_i)(1-\theta_1)^k}{(1-c_i)\theta_2(1-\theta_2)^{k-1}} . \frac{\theta_1^2(1-\theta_1)^{k-2}}{\theta_2^2(1-\theta_2)^{k-2}} \frac{\theta_1^3(1-\theta_1)^{k-3}}{\theta_2^3(1-\theta_2)^{k-3}} \cdots \frac{\theta_1^k}{\theta_2^2}]^{1/k+1}. \quad (19)$$

It is observed that $\bar{J}_i$ tends to 1 if $\theta_1, \theta_2 \to \frac{1}{2}$.

4. Simulation study

In this section, we present some numerical illustrations. The tables along the figures provide how our proposed method works for different prior probabilities L_i with the probability of direct response c_i , which is actually unknown, but here, for calculating posterior probabilities with their response specific jeopardy measures, we assume c_i artificially, say, 0.06,0.12,0.63,0.57,0.91, etc. In Figures 1 , 2 , 3 , taking L_i along horizontal axis (in graph "L_i") and c_i along vertical directions (in graph "C_i"), the plotted points represent the geometric mean of response specific jeopardy measures along with relevant "p" values or "θ" values (denoted by (**GM_J;** p_1, p_2) or (**GM_J;** p_1, p_2, p_3, p_4) or (**GM_J;** θ_1, θ_2)in graphs). Table 1 shows the calculations for ORR using

Warner's model and Greenberg et al.'s unrelated question model with the overall measure of jeopardy $\bar{J}_i$ in the last column, which is exactly 1, whatever the values of L_i, c_i, p_1 and p_2 as discussed in Section 3.1. Here, it is slightly different from 1 due to approximations in posterior probabilities and response specific jeopardy measures. Figure 1 is a representation of the measure of jeopardy $\bar{J}_i$ for all the combinations of (L_i, c_i) as mentioned in Table 1. Table 2 represents the calculation for ORR using Forced response model imposing the restriction $p_1 p_4 = p_2 p_3$ as pointed out in Section 3.3. Figure 2 is a diagrammatic representation of the measure of jeopardy $\bar{J}_i$ while the ORR survey is performed by using the forced model. The numerical study for ORR using Kuk's model is shown in Table 3 along with Figure 3. If the number of cards (k) drawn for RR devices is 2, an artificial data set is used for the simulation study and the results are shown in Table 4. The data consist of an imaginary set of 116 undergraduate students aged below 20 and their reckless driving with weekly expenditures. We are interested to estimate the proportion of the students who broke the traffic rules last year. An unrelated auxiliary variate, whether they are interested in painting, takes for the numerical illustration of optional randomized techniques with Greenberg et al.'s (1969) RR device, as mentioned in Section 3.2. Let U= (1,2,...*i*...,N) be a finite labelled population with N units and the proportion π may be defined as $\pi = \frac{1}{N}\sum_{i=1}^{N} y_i$ treating y as a "study qualitative stigmatizing variable", as mentioned in Section 3.

Samples are taken from the population with unequal probability sampling scheme of Lahiri (1951) – Midzuno (1952) – Sen (1953) used for the selection of a sample of 39 units to estimate the population proportion. Here, the first unit is selected with the probability $p_i^* = \frac{z_i}{Z}$ $(\text{where } Z = \sum_{1}^{N} z_i)$, the normed size measure and the remaining ones are selected by simple random sampling without replacement (SRSWOR) from the remaining units in the population after the first draw. The variable "Have you ever been fined for breaking traffic rules" is our study qualitative characteristic with "Weekly expenditure" as the size measure. In this design, the inclusion probability π_i of the i^{th} unit in the sample of size n from the population of size N is $p_i^* + (1 - p_i^*)\frac{n-1}{N-1}$ as the i^{th} unit may be selected in first position with probability p_i^* or in any other position with probability $(1 - p_i^*)$ through SRSWOR with probability $\frac{n-1}{N-1}$. Clearly,

the second order inclusion probability of the unit (i, j) may be obtained by the following formula $\pi_{ij} = \frac{(n-1)(N-n)(p_i^* + p_j^*) + (n-1)(n-2)}{(N-1)(N-2)}$. We employ Horvitz-Thompson estimator (HTE) to estimate the proportion $\pi = \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i}$ in the case of qualitative character. Since y_i is not directly assessable, an unbiased estimator r_i of y_i is assigned here. Hence, $e = \frac{1}{N}\sum_{i\in s}\frac{r_i}{\pi_i}$ is our the final unbiased estimator of the population proportion with an unbiased estimator of variance $\mathrm{v}(e) = \frac{1}{N^2}[\sum_{i<j\in s}\sum\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j})^2 + \sum_{i\in s}\frac{v_i}{\pi_i}]$ where v_i is an unbiased estimator of variance of r_i. The HT estimator e for the proportion need not be a proper fraction and this anomaly arises not because of ORR as it is natural even for DR's. Proportion estimation is a big challenge in statistics. In Randomized response surveys with unequal probabilities, we usually do not face the problem of getting e values outside the range [0,1].

To judge the efficacy of our results, average coverage probabilities (ACP), average coefficient of variation (ACV) and the average Length (AL) of the 95% confidence intervals based on $e \pm 1.96\sqrt{v(e)}$ have been used. To calculate, we draw $T = 1000$ samples from the population by Lahiri (1951) – Midzuno (1952) – Sen (1953) sampling scheme. For each sample we perform ORR methods to calculate the estimates and variance estimates.

The point estimator will be judged good if the estimated coefficient of variation, namely $\mathrm{CV} = 100\frac{\sqrt{v(e)}}{e}$, has a small magnitude, preferably less than 10% or at most 30%. A confidence interval (CI) will be judged good if on drawing a large number of simulated samples, say B in the number taken as 1000, from a population at hand, the (1) CI's happen to cover the known value of the parameter, a percentage of times close to 95% -this percentage is called the ACP, the Average Coverage Percentage and (2) if the average value of the length, AL, say, of a CI is small enough. Between two CI's the one with a lower value of AL will be preferred unless its ACP is too far from 95% compared to that for the other. Tables 4.1., 4.2., 4.3. and 4.4. represent the ACV (in %), ACP (in %) and AL for four different ORR techniques. Figures 4.1., 4.2., 4.3. represent the ACV and ACP values denoted as (ACV, ACP) taking paired "p" ("θ" for optional Kuk model) values along horizontal and vertical axes.

Table 1. ORR with Warner and Unrelated - measure of jeopardy.

L_i	c_i	p_1	p_2	$J_i(0,1)$	$J_i(1,0)$	$J_i(0,0)$	$J_i(1,1)$	**Warner** $\overline{J}_i$	**Unrelated** $\overline{J}_i$
0.1	0.06	0.44	0.49	1.2216	0.8186	1.0409	0.9607	1	1
0.3	0.63	0.3	0.73	3.1622	0.3162	0.039	25.6154	0.9989	0.9989
0.4	0.42	0.95	0.11	0.0285	35.028	0.0335	29.8462	0.9981	0.9981
0.5	0.91	0.42	0.28	0.8246	1.2128	0.0034	297.6667	1.0121	1.0121
0.6	0.37	0.07	0.98	142.46	0.007	0.0145	68.7966	0.9948	0.9948
0.8	0.55	0.43	0.62	1.7154	0.5829	0.072	13.8959	1.0004	1.0004
0.9	0.53	0.57	0.73	1.6731	0.5977	0.0374	26.7692	1.0012	1.0012

Table 2. ORR with Forced model - measure of jeopardy.

L_i	c_i	p_1	p_2	p_3	p_4	$J_i(1,0)$	$J_i(0,1)$	$J_i(0,0)$	$J_i(1,1)$	$\overline{J}_i$
0.1	0.42	0.64	0.23	0.24	0.0863	0.1367	1.4002	0.012	15.9556	0.4375
0.2	0.43	0.45	0.4	0.52	0.4622	1.1	0.7667	0.1154	7.3051	0.9183
0.6	0.18	0.61	0.25	0.47	0.1926	0.4195	0.8615	0.1049	3.4467	0.6013
0.6	0.26	0.39	0.31	0.43	0.3418	0.9762	0.7592	0.1191	6.2231	0.8609
0.7	0.3	0.25	0.4	0.37	0.5920	2.2162	0.7749	0.1892	9.0769	1.3105
0.9	0.1	0.69	0.21	0.6	0.1826	0.4544	0.7778	0.1739	2.0323	0.5945
0.9	0.12	0.58	0.35	0.37	0.2233	0.4039	1.5337	0.1889	3.2799	0.7871

Table 3. ORR with Kuk model (k=2) - measure of jeopardy.

L_i	c_i	θ_1	θ_2	$J_i(0)$	$J_i(1)$	$J_i(2)$	$\overline{J}_i$
0.1	0.18	0.72	0.43	0.1333	1.75	2.867	0.874
0.2	0.36	0.78	0.13	0.0357	6.7143	39	2.107
0.3	0.55	0.54	0.34	0.1333	6.6	2.6	1.318
0.4	0.72	0.49	0.5	0.0886	11.286	1	1
0.5	0.8	0.58	0.53	0.0476	17	1.167	0.981
0.6	0.74	0.55	0.2	0.0549	20	8	2.063
0.7	0.78	0.75	0.76	0.0127	20.5	0.923	0.622
0.9	0.49	0.77	0.81	0.0588	7.25	0.909	0.729

Table 4.1. ACV, ACP, AL for Optional Warner Model

p_1	p_2	ACV	ACP	AL
0.28	0.19	48.1865	94.6	3.8623
0.49	0.36	37.5586	99.1	2.0169
0.54	0.29	26.3733	98	1.0650
0.63	0.56	42.3268	97.3	2.6278
0.66	0.45	24.7465	97.9	0.9712
0.77	0.65	26.8792	99.4	1.0953
0.81	0.63	20.2835	92.5	0.7143

Table 4.2. ACV, ACP, AL for Optional Unrelated Model

p_1	p_2	ACV	ACP	AL
0.36	0.23	42.2166	95.2	2.4704
0.51	0.39	38.2907	86.8	2.0682
0.56	0.49	45.1583	86.5	3.0992
0.69	0.54	28.0311	98.3	1.1740
0.72	0.55	24.9143	93.6	0.9746
0.88	0.34	11.7061	96.8	0.3454
0.92	0.61	12.1766	93.5	0.3634

Table 4.3. ACV, ACP, AL for Optional Forced Model

p_1	p_2	p_3	p_4	ACV	ACP	AL
0.64	0.23	0.24	0.0863	33.7349	90.1	0.5066
0.45	0.4	0.52	0.4622	46.4117	79.4	3.1305
0.39	0.31	0.43	0.3418	55.3157	76.4	4.6584
0.25	0.4	0.37	0.5920	27.8805	91.2	1.1824
0.32	0.38	0.4	0.4750	37.776	84.4	2.0412
0.35	0.23	0.47	0.3089	32.5996	87.7	1.5525
0.15	0.13	0.22	0.1907	28.2077	98.1	1.1993

Table 4.4. ACV, ACP, AL for Optional Kuk Model if k=2

θ_1	θ_2	ACV	ACP	AL
0.6	0.2	5.4740	94.5	0.1749
0.8	0.6	11.3493	95.4	0.3933
0.56	0.4	12.1329	96	0.3856

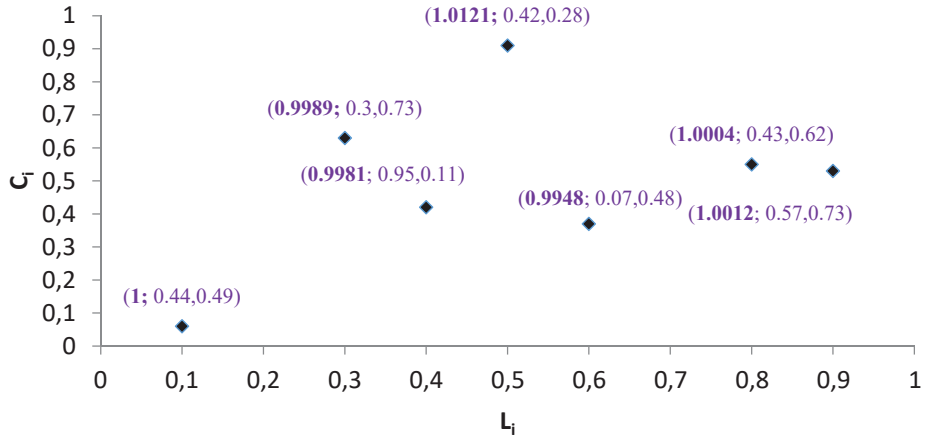


Figure 1. Measure of Jeopardy for Warner and Unrelated ORR

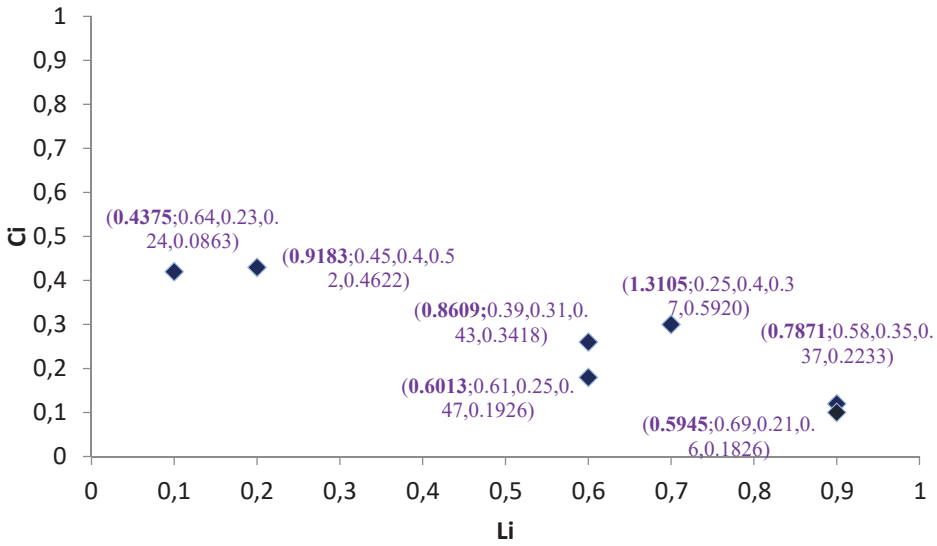


Figure 2. Measure of Jeopardy for Forced ORR

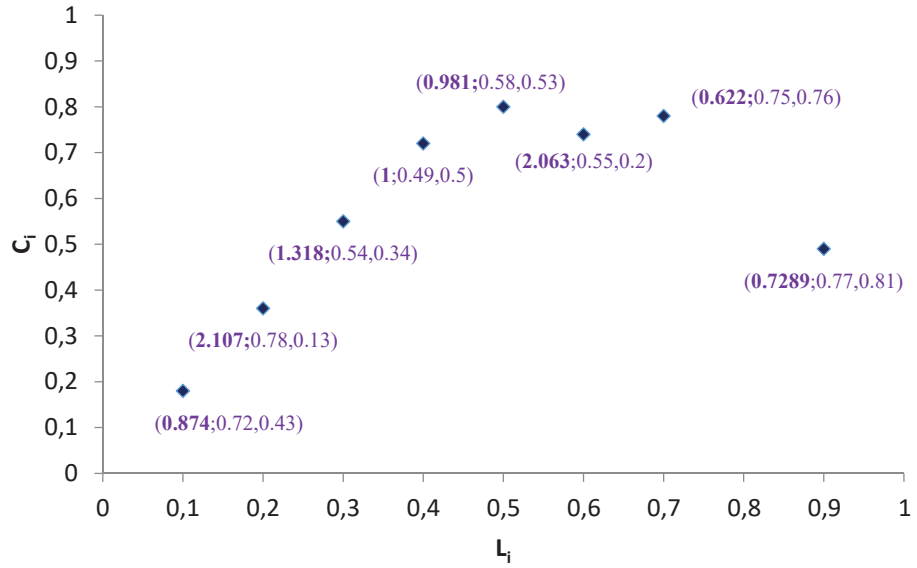


Figure 3. Measure of Jeopardy for Kuk ORR

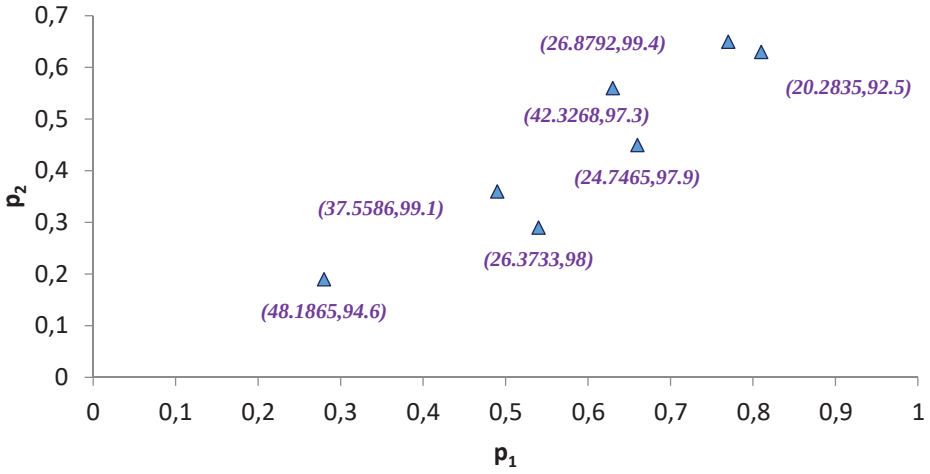


Figure 4.1. Representation of ACP , ACV for Warner ORR

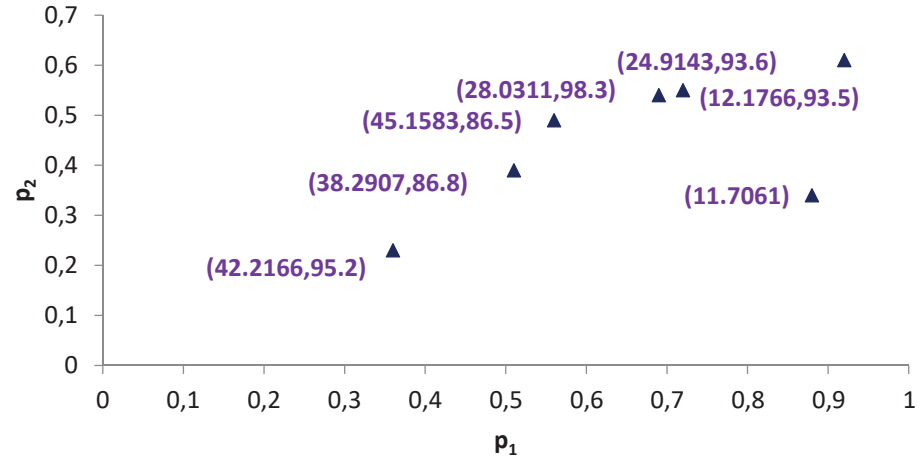


Figure 4.2. Representation of ACP , ACV for Unrelated ORR

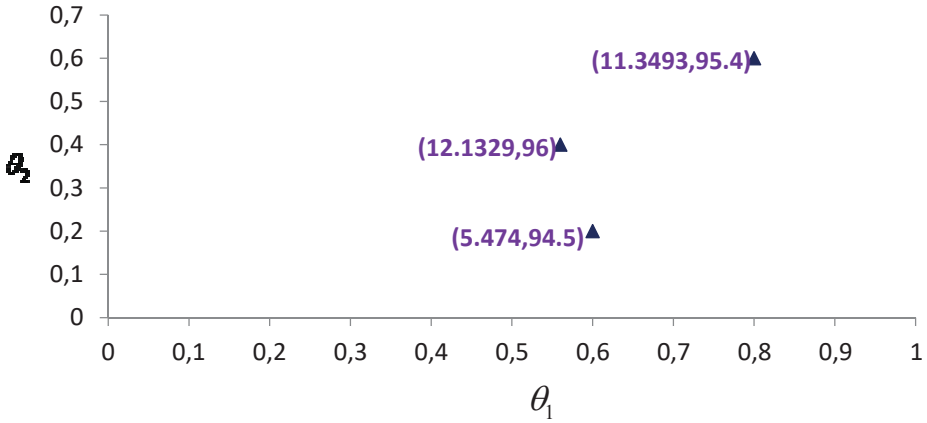


Figure 4.3. Representation of ACP , ACV for Kuk ORR

5. Concluding remarks

Most of the literature on the theory of RR is restricted to simple random sampling (SRS) with replacement (SRSWR). We strongly believe that extension of the theory of RR to varying probability sampling is necessary.

In our proposed ORR method, the probability of choosing between a 'direct' and an 'RR' should vary across individuals rather than be a constant and that is unknown. To get an unbiased variance estimator, two responses from each individual are

necessarily required. Regarding the privacy protection of each individual, we can proceed with the ORR method. As a measure of jeopardy, the average jeopardy measure with geometric mean is successfully carried out.

From our results we observe that all of the competing ORR methods show satisfactory results in terms of ACP and ACV values.

Acknowledgement

We are greatly appreciative of the kind comments from the referees as a means to our fruitful efforts in implementation of a possible improvement on our initial draft.

REFERENCES

ARNAB, R., (2004). Optional randomized response techniques for complex survey designs. Biom, J. 46, pp. 114–124.

ARNAB, R., RUEDA, M., (2016). Optional randomized response: A critical review. Handbook of statistics, Elsevier, 34, pp. 253–271

CHAUDHURI, A., (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population, *Journal of Statistical Planning and Inference*, 94, pp. 37–42.

CHAUDHURI, A., (2011). Randomized response and indirect questioning techniques in surveys. CRC Press, Fl. USA.

CHAUDHURI, A., CHRISTOFIDES, T. C., (2013). Indirect questioning in sample surveys. Springer-Verlag, Berlin, Heidelberg.

CHAUDHURI, A., DIHIDAR, K., (2009). Estimating means of stigmatizing qualitative and quantitative variables from discretionary responses randomized or direct. *Sankhya B*, 71, pp. 123–136.

CHAUDHURI, A., MUKERJEE, R., (1985). Optionally randomized responses techniques. *Calcutta Statistical Association Bulletin,* 34, pp. 225–229.

CHAUDHURI, A., SAHA, A., (2005). Optional versus compulsory randomized response techniques in complex surveys, *Journal of Statistical Planning and Inference*, 135, pp. 516–527.

CHAUDHURI, A., CHRISTOFIDES, T. C., RAO, C. R., (2016). Handbook of statistics, Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits. Elsevier, NL, 34, pp. 2–525.

CHAUDHURI, A., CHRISTOFIDES, T. C., SAHA, A., (2009). Protection of privacy in efficient application of randomized response techniques, *Statistical Methods and Applications,* 18, pp. 389–418.

GREENBERG, B. G., ABUL-ELA, A.-L., SIMMONS, W. R., HORVITZ, D. G., (1969). The unrelated question RR model: Theoretical framework, *Journal of American Statistical Association,* 64, pp. 520–539.

GUPTA, S., (2001). Qualifying the sensitivity level of binary response personal interview survey questions. *Journal of Combinatorics, Information and System Sciences,* 26 (1- 4), pp. 101−109.

GUPTA, S., GUPTA, B., SINGH, S., (2002). Estimation of sensitivity level of personal interview survey question, *Journal of Statistical Planning and Inference,* 100, pp. 239–247.

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

HUANG, K. C., (2008). Estimation of sensitive characteristics using optional randomized techniques, Qual. Quant. 42, pp. 679–686.

KUK, A. Y. C., (1990). Asking sensitive questions indirectly. *Biometrika*, 77(2), pp. 436–438.

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

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

PAL, S., (2008). Unbiasedly estimating the total of a stigmatizing variable from a complex survey on permitting options for direct or randomized responses, *Statistical Papers,* 49, pp. 157−164

SAHA, A., (2007). Optional randomized response in stratified unequal probability sampling. A simulation based numerical study with Kuk's method, Test 16, pp. 346–354.

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

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

APPENDICES

Appendix 1. variance estimation in ORR using Warner's (1965) RR model (under Section 3.1.)

Estimator $r_i = \dfrac{(1-p_2)Z_i - (1-p_1)Z_i'}{p_1 - p_2}$ (from Section 3.1)

$$V_R(r_i) = \frac{(1-p_2)^2 V_R(Z_i) + (1-p_1)^2 V_R(Z'_i)}{(p_1 - p_2)^2} \quad \text{and}$$

$$\begin{aligned}
V_R(Z_i) &= E_R(Z_i^2) - [E_R(Z_i)]^2 \\
&= E_R(Z_i) - [E_R(Z_i)]^2 \\
&= E_R(Z_i)[1 - E_R(Z_i)] \\
&= [c_i y_i + (1-c_i)\{(1-p_1) + (2p_1 - 1)y_i)\}][1 - c_i y_i - (1-c_i)\{(1-p_1) + (2p_1 - 1)y_i)\}] \\
&= c_i y_i + (1-c_i)(1-p_1) + (1-c_i)(2p_1 - 1)y_i - c_i^2 y_i - c_i(1-c_i)(1-p_1)y_i - c_i(1-c_i)(2p_1 - 1)y_i \\
&\quad - c_i(1-c_i)(1-p_1)y_i - (1-c_i)^2(1-p_1)^2 - (1-c_i)^2(1-p_1)(2p_1 - 1)y_i \\
&\quad - c_i(1-c_i)(2p_1 - 1)y_i - (1-c_i)^2(2p_1 - 1)^2 y_i \\
&= c_i(1-c_i)y_i + (1-c_i)(1-p_1) + (1-c_i)(2p_1 - 1)y_i - c_i(1-c_i)(1 - p_1 + 2p_1 - 1)y_i \\
&\quad - c_i(1-c_i)p_1 y_i - (1-c_i)^2\{(1-p_1)^2 + 2(1-p_1)(2p_1 - 1)y_i + (2p_1 - 1)^2 y_i\} \\
&= c_i(1-c_i)y_i + (1-c_i)(1-p_1) + (1-c_i)(2p_1 - 1)y_i - 2c_i(1-c_i)p_1 y_i \\
&\quad - (1-c_i)^2(1-p_1)^2 - (1-c_i)^2(2p_1 - 1)(2 - 2p_1 + 2p_1 - 1)y_i \\
&= (1-c_i)[c_i y_i + (1-p_1) + (2p_1 - 1)y_i - 2c_i p_1 y_i - (1-c_i)(1-p_1)^2 - (1-c_i)(2p_1 - 1)y_i] \\
&= (1-c_i)[(1-p_1) + c_i y_i + (2p_1 - 1)(1 - 1 + c_i)y_i - 2c_i p_1 y_i - (1-c_i)(1-p_1)^2] \\
&= (1-c_i)[(1-p_1) + c_i y_i + (2p_1 - 1)c_i y_i - 2p_1 c_i y_i - (1-c_i)(1-p_1)^2] \\
&= (1-c_i)(1-p_1)[1 - (1-c_i)(1-p_1)] \\
&= (1-c_i)(1-p_1)[c_i + (1-c_i)p_1]
\end{aligned}$$

Similarly, $V_R(Z_i') = (1-c_i)(1-p_2)[c_i + (1-c_i)p_2]$.

So,

$$V_R(r_i) = \frac{(1-p_2)^2(1-c_i)(1-p_1)(c_i+(1-c_i)p_1)+(1-p_1)^2(1-c_i)(1-p_2)(c_i+(1-c_i)p_2)}{(p_1-p_2)^2}$$

$$= \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(1-c_i)[(1-p_2)(c_i+(1-c_i)p_1)+(1-p_1)(c_i+(1-c_i)p_2)]$$

$$= \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(1-c_i)[(2-p_1-p_2)c_i+(1-c_i)\{p_1(1-p_2)+p_2(1-p_1)\}]$$

$$= \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(1-c_i)[(2-p_1-p_2)c_i+(1-c_i)\{p_1+p_2-2p_1p_2\}]$$

$$= \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(1-c_i)[(2-p_1-p_2)\{1-(1-c_i)\}+(1-c_i)\{p_1+p_2-2p_1p_2\}]$$

$$= \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(1-c_i)[(2-p_1-p_2)-(1-c_i)\{2-2p_1-2p_2+2p_1p_2\}]$$

$$= \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(1-c_i)[(2-p_1-p_2)-2(1-c_i)(1-p_1)(1-p_2)]$$

Now,

$$E(Z_i-Z_i')^2 = E(Z_i+Z_i'-2Z_iZ_i')$$

$$= E(Z_i)+E(Z_i')-2E(Z_iZ_i')$$

$$= 2c_iy_i+(1-c_i)[(p_1+p_2)y_i+(2-p_1-p_2)(1-y_i)]-2\{c_iy_i+(1-c_i)(1-p_1+(2p_1-1)y_i\}$$

$$\{c_iy_i+(1-c_i)(1-p_2+(2p_2-1)y_i\}$$

$$= 2c_iy_i+(1-c_i)[(2-p_1-p_2)+2(p_1+p_2-1)y_i]-2[c_i^2y_i+c_i(1-c_i)(1-p_1+2p_1-1+1-p_2+2p_2-1)y_i$$

$$+(1-c_i)^2\{(1-p_1)(1-p_2)+(1-p_2)(2p_1-1)y_i+(1-p_1)(2p_2-1)y_i+(2p_1-1)(2p_2-1)y_i\}]$$

$$= 2c_i(1-c_i)y_i+(1-c_i)[(2-p_1-p_2)+2(p_1+p_2-1)y_i]-2c_i(1-c_i)(p_1+p_2)y_i$$

$$-2(1-c_i)^2\{(1-p_1)(1-p_2)+(2p_1-4p_1p_2-1+p_2+2p_2-1+p_1+4p_1p_2+1-2p_1-2p_2)y_i\}$$

$$= 2c_i(1-c_i)(1-p_1-p_2)y_i+(1-c_i)[(2-p_1-p_2)+2(p_1+p_2-1)y_i]$$

$$-2(1-c_i)^2[(1-p_1)(1-p_2)+(p_1+p_2-1)y_i]$$

$$= (1-c_i)(2-p_1-p_2)+2(1-c_i)^2(p_1+p_2-1)y_i-2(1-c_i)^2(1-p_1)(1-p_2)-2(1-c_i)^2(p_1+p_2-1)y_i$$

$$= (1-c_i)(2-p_1-p_2)-2(1-c_i)^2(1-p_1)(1-p_2)$$

Thus,

$$\frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}E(Z_i-Z_i')^2 = \frac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(1-c_i)[(2-p_1-p_2)-2(1-c_i)(1-p_1)(1-p_2)] = V_R(r_i)$$

i.e. $v_i = \dfrac{(1-p_1)(1-p_2)}{(p_1-p_2)^2}(Z_i-Z_i')^2$ is an unbiased estimator of $V_R(r_i)$.

Appendix 2. ORR using Greenberg et al.'s (1969) unrelated question RR model (under Section 3.2)

Estimator $r_i = \dfrac{(1-p_2)Z_i - (1-p_1)Z_i'}{p_1 - p_2}$ (from Section 3.2)

$V_R(r_i) = \dfrac{(1-p_2)^2 V_R(Z_i) + (1-p_1)^2 V_R(Z_i')}{(p_1 - p_2)^2}$ and

$V_R(Z_i) = E_R(Z_i)(1 - E_R(Z_i)) = (y_i - x_i)^2 (1-p_1)(1-c_i)(p_1 + (1-p_1)c_i)$

Similarly, $V_R(Z_i') = (y_i - x_i)^2 (1-p_2)(1-c_i)(p_2 + (1-p_2)c_i)$

So, $V_R(r_i)$ can be written as,

$$V_R(r_i) = \frac{(y_i - x_i)^2 (1-c_i)(1-p_1)(1-p_2)}{(p_1 - p_2)^2}[2c_i(1-p_1)(1-p_2) + p_1 + p_2 - 2p_1p_2].$$

Now,

$E_R(Z_i - Z_i')^2 = E_R(Z_i) + E_R(Z_i') - 2E_R(Z_i)E_R(Z_i')$

$= (y_i - x_i)^2 (1-c_i)[(p_1 + p_2) - 2p_1p_2 + 2c_i(1-p_1)(1-p_2)]$

Thus, $v_i = \dfrac{(1-p_1)(1-p_2)}{(p_1 - p_2)^2}(Z_i - Z_i')^2$ is an unbiased estimator of $V_R(r_i)$.

ORR using Forced response model (under Section 3.3.)

Estimator $r_i = \dfrac{p_3 Z_i - p_1 Z_i'}{p_3 - p_1}$ (from Section 3.3) ,

$V_R(r_i) = \dfrac{p_3^2}{(p_3 - p_1)^2} V_R(Z_i) + \dfrac{p_1^2}{(p_3 - p_1)^2} V_R(Z_i')$ and

$V_R(Z_i) = (2y_i - 1)(1-c_i)\{(p_1 + p_2)y_i - p_1\} - (1-c_i)^2\{(p_1 + p_2)y_i - p_1\}^2$

$= p_1^2(2y_i - 1)(1-c_i)(\dfrac{p_1 + p_2}{p_1} y_i - 1)\{\dfrac{1}{p_1} - (1-c_i)(\dfrac{p_1 + p_2}{p_1} y_i - 1)(2y_i - 1)\}$ as $(2y_i - 1)^2 = 1$

Similarly,

$V_R(Z_i') = (2y_i - 1)(1-c_i)\{(p_3 + p_4)y_i - p_3\} - (1-c_i)^2\{(p_3 + p_4)y_i - p_3\}^2$

$= p_3^2(2y_i - 1)(1-c_i)(\dfrac{p_3 + p_4}{p_3} y_i - 1)\{\dfrac{1}{p_3} - (1-c_i)(\dfrac{p_3 + p_4}{p_3} y_i - 1)(2y_i - 1)\}$

Using the condition $p_1 p_4 = p_2 p_3$, (see Section 3.3)

$$V_R(r_i) = \frac{p_1^2 p_3^2}{(p_3 - p_1)^2}(2y_i - 1)(1 - c_i)(\frac{p_1 + p_2}{p_1} y_i - 1)(\frac{1}{p_1} + \frac{1}{p_3} - 2(1 - c_i)(\frac{p_1 + p_2}{p_1} y_i - 1)(2y_i - 1)) \cdot$$

Now,

$$E(Z_i - Z_i')^2 = p_1 p_3 (2y_i - 1)(1 - c_i)(\frac{p_1 + p_2}{p_1} y_i - 1)(\frac{1}{p_1} + \frac{1}{p_3} - 2(1 - c_i)(\frac{p_1 + p_2}{p_1} y_i - 1)(2y_i - 1)) \cdot$$

So, $v_i = \frac{p_1 p_3}{(p_3 - p_1)^2}(Z_i - Z_i')^2$ is an unbiased estimator for $V_R(r_i)$.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 89–117, DOI 10.21307/stattrans-2020-015
Submitted – 28.08.2019; paper accepted for published – 12.03.2020

A comparison study on a new five-parameter generalized Lindley distribution with its sub-models

Ramajeyam Tharshan [1] [2] **Pushpakanthie Wijekoon**[3]

ABSTRACT

In recent years, modifications of the classical Lindley distribution have been considered by many authors. In this paper, we introduce a new generalization of the Lindley distribution based on a mixture of exponential and gamma distributions with different mixing proportions and compare its performance with its sub-models. The new distribution accommodates the classical Lindley, Quasi Lindley, Two-parameter Lindley, Shanker, Lindley distribution with location parameter, and Three-parameter Lindley distributions as special cases. Various structural properties of the new distribution are discussed and the size-biased and the length-biased are derived. A simulation study is conducted to examine the mean square error for the parameters by means of the method of maximum likelihood. Finally, simulation studies and some real-world data sets are used to illustrate its flexibility in terms of its location, scale and shape parameters.

Key words: Lindley distribution, mixture distributions, size-biased distributions, maximum likelihood estimation.

1. Introduction

In the modeling of the lifetime data, especially biomedical science, engineering, actuarial science, several continuous distributions bounded to 0 and ∞ have been developed, which may have one or more parameter(s). Examples of such distributions are exponential, gamma, Lindley, log-normal, Weibull and their modifications. These distributions may have various abilities to cover the tail-heaviness of a data set. The tail-heaviness of a data set may be measured by the excess kurtosis (EK) and EK is defined as $\tau - 3$, where τ is the kurtosis of the data set. The $EK > 0$ is called a fatter tail (Leptokurtic) and $EK < 0$ is called a thinner tail (Platykurtic) distributions. Among the distributions mentioned-above, Lindley distribution (LD) which was developed by Lindley (1958), and its modifications are more flexible than the above-mentioned distributions, especially when considering less complexity of their mathematical forms, shapes, and failure rate criteria.

[1] Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka.
E-mail: tharshan10684@gmail.com. ORCID: https://orcid.org/0000-0002-6112-2517.
[2] Department of Mathematics and Statistics, University of Jaffna, Sri Lanka
[3] Department of Statistics and Computer Science, University of Peradeniya, Peradeniya, Sri Lanka.
E-mail: pushpaw@pdn.ac.lk. ORCID: https://orcid.org/0000-0003-4242-1017.

The LD is a one-parameter exponential family lifetime distribution defined over the interval $[0,\infty)$ having the density function:

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

where θ is the shape parameter, and y is the respective random variable. The density function of this distribution can be verified that a two-component mixture of two different continuous distributions namely exponential (θ) and gamma $(2,\theta)$ distributions with the mixing proportion, $p = \frac{\theta}{1+\theta}$. Ghitany (2008) has done a comprehensive study on the mathematical and statistical properties of the LD and showed that the LD is more flexible and provides a better fit than the exponential distribution for lifetime data.

Even though the LD is used for modeling of the lifetime data, researchers are more keen on its modified forms in terms of increasing the flexibility of LD's shapes and failure rate criteria in recent years. Therefore, many researchers have proposed several modified forms of the LD as an alternative to LD in the past few years. Proposed new distributions are developed in terms of introducing new parameter(s) to the existing distributions. The new parameter(s) might be introduced from the latent variable distribution or mixing components that may be exponential and gamma or gamma and gamma. In this line of new proposed distributions, we may make references to a considerable number of existing distributions that are actual mixing components of LD with an exponential(θ) and a gamma($2,\theta$) distributions mixture but different mixing proportions. The existing distributions are listed below.

Shanker et. al. (2013a) obtained Quasi Lindley distribution (QLD), and discussed its various statistical properties. The distribution is a two-parameter family distribution with density function:

$$f_Y(y) = \frac{\theta(\alpha+y\theta)}{1+\alpha}e^{-\theta y}; y > 0, \theta > 0, \alpha > -1, \tag{2}$$

where α and θ are shape, and scale parameters, respectively. The mixing proportion, $p = \frac{\alpha}{\alpha+1}$. Note that the LD is a special case of the QLD when $\alpha = \theta$.

Shanker et. al. (2013b) introduced the two-parameter Lindley distribution (TwPLD) and discussed its statistical properties. Its density function is given by:

$$f_Y(y) = \frac{\theta^2(1+\alpha y)}{\theta+\alpha}e^{-\theta y}; y > 0, \theta > 0, \alpha > -\theta, \tag{3}$$

where θ and α are shape parameters. The mixing proportion, $p = \frac{\theta}{\theta+\alpha}$. Note that the LD is a special case of TwPLD when $\alpha = 1$.

Shanker (2015) introduced a one-parameter family distribution, namely Shanker distribu-

tion (SD) with the probability density function:

$$f_Y(y) = \frac{\theta^2}{\theta^2+1}(\theta+y)e^{-\theta y}; y > 0, \theta > 0, \tag{4}$$

where θ is the shape parameter. The mixing proportion, $p = \frac{\theta^2}{\theta^2+1}$.

To increase more flexibility in this line of development, Abdol-Monsef (2016) introduced a new three-parameter family generalized Lindley distribution (TPLwLD) by adding the location parameter for the exponential and gamma components. In this paper, a clear clarification is given that the location parameter is an important parameter in a statistical model to estimate the starting point of the distribution. The density function of TPLwLD is given by:

$$f_Y(y) = \frac{\theta^2}{\theta+\alpha}(1+\alpha(y-\beta))e^{-\theta(y-\beta)}; y > \beta > 0, 1+\alpha y > 0, \theta > 0, \alpha+\theta > 0, \tag{5}$$

where θ and α are shape parameters and β is a location parameter. Equation (5) presents two-component mixture of an exponential (θ,β) and gamma (2,θ,β) distributions with the mixing proportion, $p = \frac{\theta}{\theta+\alpha}$. Here the location parameter is added from the mixing components when comparing with TwPLD. Note that LD is a special case of the TPLwLD when $\alpha = 1, \beta = 0$.

Shanker et. al. (2017) obtained the Three-parameter Lindley distribution (ThPLD) with the following density function:

$$f_Y(y) = \frac{\theta^2}{\theta\alpha+\beta}(\alpha+\beta y)e^{-\theta y}; y > 0, \theta > 0, \beta > 0, \theta\alpha+\beta > 0, \tag{6}$$

where θ, α and β are shape parameters. The mixing proportion, $p = \frac{\theta\alpha}{\theta\alpha+\beta}$. Note that the LD is a special case of ThPLD when $\alpha = 1, \beta = 1$.

It is clear that when introducing a new such types of LDs, the researchers incorporate with three types of parameters, namely shape parameters from the latent variable distribution, scale and location parameters from the mixing components. Table 1 summarizes the application of the three types of parameters of the above-mentioned distributions.

The aim of this paper is to introduce a new generalized LD that accommodates all the distributions given in Table 1, and study the importance of the location parameter in the model and different mixing proportions in the development process of the new Lindly family distributions. Further, the new distribution is based on the two-component mixture of exponential and gamma distributions with different mixing proportions and it will be called as the five-parameter generalized Lindley distribution (FPGLD). A simulation study will

be done to study the performance of the maximum likelihood estimators of FPGLD. Further, a comparison study will be done with its sub-models by using simulated data sets, and real-world applications. The characteristics of the data sets will be differentiated by their skewness, Excess kurtosis (EK), and Fano factor values.

Organization of this paper is as follows: in section 2 we introduce the FPGLD and its sub-models. Its statistical properties and reliability properties are presented in section 3 and section 4, respectively. Further, section 5 covers the size-biased form of the FPGLD. The parameter estimation is discussed in section 6. Finally, a simulation study is conducted to examine the performance of the maximum likelihood estimators for FPGLD, and simulated data sets and real-world data sets are used for the comparison study with its sub-models.

Table 1. Application of three types of parameters

Distribution	Authors	Parameters		
		shape	scale	location
LD(θ)	Lindley (1958)	θ	-	-
TwPLD(θ,α)	Shanker et.al.(2013a)	θ,α	-	-
QLD(θ,δ)	Shanker et.al.(2013b)	α	θ	-
SD(θ)	Shanker (2015)	θ	-	-
TPLwLD(θ,α,β)	Monsef (2016)	θ,α		β
ThPLD(θ,α,β)	Shanker et.al.(2017)	θ,α,β	-	-

2. Five parameter generalized Lindley distribution

In this section, we introduce the five-parameter generalized Lindley distribution (FPGLD) with its sub-models.

The probability density function (pdf) of the FPGLD with parameters $\theta,\beta,\alpha,\delta$ and η is defined by;

$$f_Y(y) = \frac{\theta}{\delta\alpha+\eta}\left(\delta\alpha+\eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}, \tag{7}$$

where $y>\beta\geq 0, \theta>0, \delta\alpha>-\eta, \delta\alpha>-\eta\theta(y-\beta)$, and the range of the parameters are based on the log-likelihood function. The proposed distribution is a two-component mixture of exponential distribution with parameters θ and β, and gamma distribution with parameters $2,\theta$ and β with mixing proportion, $p=\dfrac{\delta\alpha}{\delta\alpha+\eta}$, where δ,α,η are shape parameters, and θ and β are scale and location parameters, respectively. Note that the FPGLD has the same mixing components of TPLwLD but different mixing proportion.

The probability density function of the FPGLD has some desirable properties:

(i) $f(\beta)=\dfrac{\theta\delta\alpha}{\delta\alpha+\eta}$ (ii) $\lim_{y\to\infty} f(y)=0$

The first derivative of equation (7) is derived as:

$$f^{'}(y) = \frac{\theta^2 e^{-\theta(y-\beta)}}{\delta\alpha+\eta}\left(-(\delta\alpha+\eta\theta(y-\beta))+\eta\right).$$

Then, $f^{'}(y) = 0$ gives $y_0 = \dfrac{\eta(1+\theta\beta)-\delta\alpha}{\eta\theta}$, when $\eta > \dfrac{\delta\alpha}{1+\theta\beta}$.

Therefore, the mode of the FPGLD is given by:

$$mode(y) = \begin{cases} \dfrac{\eta(1+\theta\beta)-\delta\alpha}{\eta\theta} & \text{if } \eta > \dfrac{\delta\alpha}{1+\theta\beta} \text{ and } \eta > 0. \\ \beta & \text{otherwise} \end{cases}$$

Graphs in Figure 1 have drawn by fixing four parameters and changing the fifth parameter. Figure 1 presents the possible shapes of the pdf of the FPGLD at different parameter values.

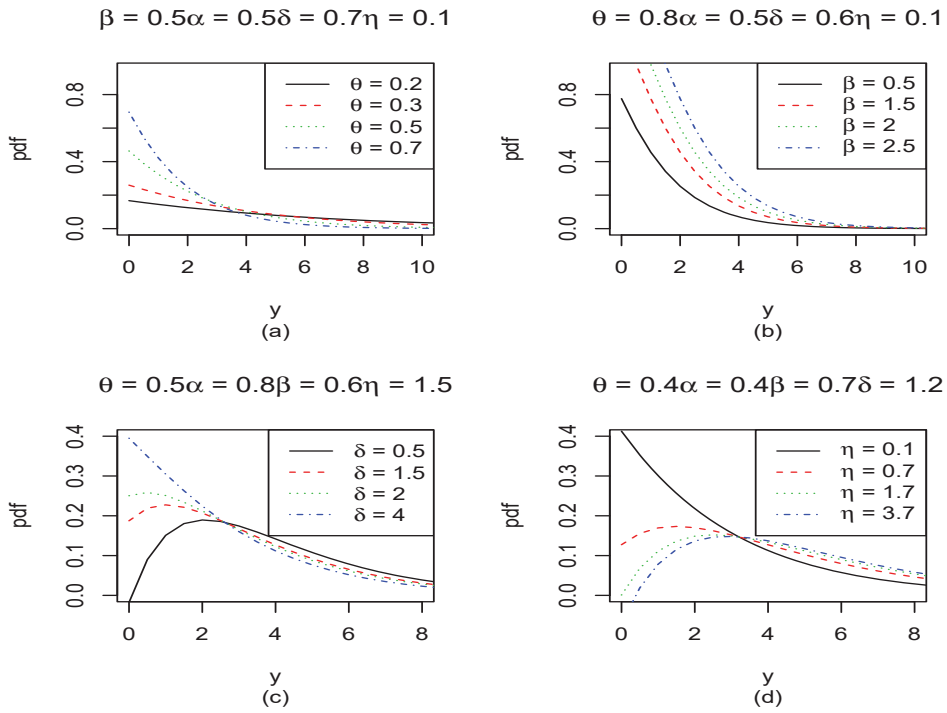


Figure 1: The probability density of FPGLD at different parameter values

(a) β, α, δ and η are fixed, and θ values are changed, (b) θ, α, δ and η are fixed, and β values are changed, (c) θ, β, α and η are fixed, and δ values are changed, (d) θ, β, α and δ are fixed, and η values are changed.

The corresponding cumulative distribution function of the FPGLD is given by:

$$F(y) = 1 - \left(1 + \frac{\eta\theta(y-\beta)}{\delta\alpha+\eta}\right)e^{-\theta(y-\beta)}, \tag{8}$$

where, $y > \beta \geq 0, \theta > 0, \delta\alpha > -\eta, \delta\alpha > -\eta\theta(y-\beta)$.

Sub-models of the FPGLD

The Five-parameter generalized Lindley distribution is nested with six existing Lindley family distributions when setting different particular numerical values of subsets of parameters, namely Lindley distribution (Lindley, 1958), Two-parameter Lindley distribution (Shanker et.al.,2013b), Quasi Lindley distribution (Shanker et.al.,2013a), Shanker distribution (Shanker, 2015), Lindley distribution with location parameter (Monsef, 2016), and Three-parameter Lindley distribution (Shanker et.al.,2017). Table 2 summarizes these modified Lindley distributions as sub-models of the FPGLD. From the knowledge of parameters in the sub-models of the FPGLD, the performance of the newly introduced shape parameters, δ and α in FPGLD in a data set could be studied comparing with TPLwLD, and the performance of the location parameter in a data set could be studied comparing TPLwLD and TwPLD.

Table 2. Sub-models of the FPGLD

Distribution	Parameters					References
	Shape			Scale	Location	
FPGLD(θ,β,α,δ,η)	δ	α	η	θ	β	in this paper
LD(θ)	θ	1	1	θ	0	Lindley (1958)
TwPLD(θ,η)	θ	1	η	θ	0	Shanker et. al.(2013)
QLD (θ,δ)	δ	1	1	θ	0	Shanker et. al.(2013)
SD(θ)	θ	θ	1	θ	0	Shanker (2015)
TPLwLD(θ,η,β)	θ	1	η	θ	β	Monsef (2016)
ThPLD(θ,α,η)	θ	α	η	θ	0	Shanker et.al.(2017)

3. Statistical properties

In this section, we provide basic statistical properties of the FPGLD such as r^{th} moment about the origin, central moments, moment generating function, and characteristic function.

3.1. Moments and related measures

The statistical properties of the central tendency, dispersion, skewness, and kurtosis can be studied through the moments. The following theorem gives the r^{th} moment about the origin.

Theorem 1. The r^{th} moment about the origin of the FPGLD is given by:

$$\mu_r^{'} = \frac{e^{\theta\beta}}{(\delta\alpha+\eta)\theta^r}\Bigg(r\Gamma(r,\theta\beta)(\delta\alpha-\eta\beta\theta+\eta(r+1))+\delta\alpha(\theta\beta)^r e^{-\theta\beta} + \eta(r+1)(\theta\beta)^r e^{-\theta\beta}\Bigg). \tag{9}$$

Proof.

$$\begin{aligned}
\mu_r^{'} &= \int_\beta^\infty y^r \frac{\theta}{\delta\alpha+\eta}\Big(\delta\alpha+\eta\theta(y-\beta)\Big)e^{-\theta(y-\beta)}dy \\
&= \frac{\theta e^{\theta\beta}}{\delta\alpha+\eta}\left(\delta\alpha\int_\beta^\infty y^r e^{-\theta y}dy+\eta\theta\int_\beta^\infty y^{r+1}e^{-\theta y}dy-\eta\theta\beta\int_\beta^\infty y^r e^{-\theta y}dy\right) \\
&= \frac{\theta e^{\theta\beta}}{\delta\alpha+\eta}\left(\frac{\delta\alpha}{\theta^{r+1}}\Gamma(r+1,\theta\beta)+\frac{\eta}{\theta^{r+1}}\Gamma(r+2,\theta\beta)-\frac{\eta\beta}{\theta^r}\Gamma(r+1,\theta\beta)\right) \\
&= \frac{e^{\theta\beta}}{(\delta\alpha+\eta)\theta^r}\Bigg(r\Gamma(r,\theta\beta)(\delta\alpha-\eta\beta\theta+\eta(r+1))+\delta\alpha(\theta\beta)^r e^{-\theta\beta}+ \\
&\qquad \eta(r+1)(\theta\beta)^r e^{-\theta\beta}\Bigg).
\end{aligned}$$

Subtituting $r = 1,2,3$ and 4 in equation (9), the first four moments about the origin are derived as:

$$\begin{aligned}
\mu_1^{'} &= \frac{1}{(\delta\alpha+\eta)\theta}\Big(\delta\alpha(1+\theta\beta)+\eta(2+\theta\beta)\Big)=\mu, \\
\mu_2^{'} &= \frac{1}{(\delta\alpha+\eta)\theta^2}\Big(\delta\alpha(2+\theta\beta(2+\theta\beta))+\eta(6+\theta\beta(4+\theta\beta))\Big), \\
\mu_3^{'} &= \frac{1}{(\delta\alpha+\eta)\theta^3}\Big(\delta\alpha\Big(6+\theta\beta(6+\theta\beta(3+\theta\beta))\Big)+\eta\Big(24+\theta\beta(18+\theta\beta(6+\theta\beta)\Big)\Big),
\end{aligned}$$

and

$$\mu_4^{'} = \frac{1}{(\delta\alpha+\eta)\theta^4}\Bigg(\delta\alpha\Big(24+\theta\beta(24+\theta\beta(12+\theta\beta(4+\theta\beta)))\Big)+\eta\Bigg(120+\beta\theta(96+ \theta\beta(36+\theta\beta(8+\theta\beta)))\Bigg)\Bigg).$$

Then, the r^{th}-order moments about the mean can be obtained by using the relationship between moments about the mean and moments about the origin, i.e)

$$\mu_r = E\left[(Y-\mu)^r\right] = \sum_{r=0}^{r}\binom{r}{i}(-1)^{r-i}\mu_i^{'}\mu^{r-i}.$$

Therefore, some r^{th}-order moments about the mean are:

$$\mu_2 = -\mu^2 + \mu_2^{'} = \frac{\delta\alpha(\delta\alpha+4\eta)+2\eta^2}{(\delta\alpha+\eta)^2\theta^2} = \sigma^2,$$

$$\mu_3 = 2\mu^3 - 3\mu_2^{'}\mu + \mu_3^{'} = \frac{2\left(\delta\alpha\left((\delta\alpha)^2+6\eta^2+6\eta(\delta\alpha)\right)+2\eta^3\right)}{(\delta\alpha+\eta)^3\theta^3}, \text{ and}$$

$$\mu_4 = -3\mu^4 + 6\mu_2^{'}\mu^2 - 4\mu_3^{'}\mu + \mu_4^{'}$$

$$= \frac{3\left(\delta\alpha\left(3(\delta\alpha)^3+24\eta(\delta\alpha)^2+44\eta^2(\delta\alpha)+32\eta^3\right)+8\eta^4\right)}{(\delta\alpha+\eta)^4\theta^4}.$$

Now, the coefficient of variation ($c.v$), measures of skewness (γ_1), measures of kurtosis (γ_2), and the Index of dispersion/Fano factor (γ_3) of the FPGLD can be derived as:

$$c.v = \frac{(\mu_2)^{1/2}}{\mu_1^{'}} = \frac{\sqrt{\delta\alpha(\delta\alpha+4\eta)+2\eta^2}}{\delta\alpha(1+\theta\beta)+\eta(2+\theta\beta)},$$

$$\gamma_1 = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{2\left(\delta\alpha\left((\delta\alpha)^2+6\eta^2+6\eta(\delta\alpha)\right)+2\eta^3\right)}{(\delta\alpha(\delta\alpha+4\eta)+2\eta^2)^{3/2}},$$

$$\gamma_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{3\left(\delta\alpha\left(3(\delta\alpha)^3+24\eta(\delta\alpha)^2+44\eta^2(\delta\alpha)+32\eta^3\right)+8\eta^4\right)}{(\delta\alpha(\delta\alpha+4\eta)+2\eta^2)^2}, \text{ and}$$

$$\gamma_3 = \frac{\mu_2}{\mu_1^{'}} = \frac{\delta\alpha(\delta\alpha+4\eta)+2\eta^2}{(\delta\alpha+\eta)\theta(\delta\alpha(1+\theta\beta)+\eta(2+\theta\beta))}.$$

The horizontal symmetry, and dispersion can be measured by γ_1, and γ_3, respectively. Figures 4 and 5 (Appendix) show various patterns of the kurtosis and the skewness functions of FPGLD at different parameter values, respectively. From these figures, it is clear that the kurtosis value is increasing when δ is increasing and decreasing when η is increasing for $\delta \leq 1$. Among the different formats of α; $\alpha = 1, \alpha = \delta, \alpha = \delta^2$, and $\alpha = \delta^3$, the maximum flexibility is obtained when $\alpha = 1$, i-e) $\delta\alpha = \delta$, in terms of having higher kurtosis value for $\delta \leq 1$. Further, the skewness value is increasing with δ and decreasing with η for $\delta \leq 1$. Figures 6 and 7 (Appendix) represent different shapes of the Fano factor function of FPGLD at different parameter values. Figure 6 (a) ,(b), (c), and (d) have drawn by fixing θ, β, and η and changing δ and α. Note that all shapes are anti-U shaped and the higher Fano factor values are obtained mostly when $\alpha = 1$. Figure 7 (a), and (b) have drawn by fixing α, δ and

η and changing θ, and β, respectively. All graphs show a monotonic decreasing pattern, and the Fano factor value is increasing when β or θ value is decreasing. When comparing Figures 6 and 7, it is clear that the effect on the Fano factor function of changing δ is totally different than the effect of changing θ.

3.2. Moment generating and characteristic function

The moment generating function is useful to determine the distribution of a random variable. The following theorem provides the moment generating function of the FPGLD.

Theorem 2. The moment generating function say $M_Y(t)$ of the FPGLD is given as follows:

$$M_Y(t) = \frac{\theta e^{\beta t}}{(\delta\alpha+\eta)(t-\theta)^2}\left(-\delta\alpha(t-\theta)+\eta\theta\right). \tag{10}$$

Proof.

$$\begin{aligned} M_Y(t) &= E(e^{ty}) \\ &= \int_\beta^\infty e^{ty}\frac{\theta}{\delta\alpha+\eta}\left(\delta\alpha+\eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}dy \\ &= \frac{\theta}{\delta\alpha+\eta}\left(\delta\alpha\int_\beta^\infty e^{ty-\theta(y-\beta)}dy+\eta\theta\int_\beta^\infty ye^{ty-\theta(y-\beta)}dy-\eta\theta\beta\int_\beta^\infty ye^{ty-\theta(y-\beta)}dy\right) \end{aligned}$$

.

The integrals of the above equation will be taken separately as follows:

$$\delta\alpha\int_\beta^\infty e^{ty-\theta(y-\beta)}dy = \frac{e^{\theta\beta}\delta\alpha}{t-\theta}\left(-e^{\beta(t-\theta)}\right) = \frac{\delta\alpha e^{\beta t}}{(\theta-t)}$$

$$\begin{aligned} &\eta\theta\int_\beta^\infty ye^{ty-\theta(y-\beta)}dy \\ &= e^{\theta\beta}\eta\theta\int_{\beta(\theta-t)}^\infty \frac{z}{\theta-t}e^{-z}\frac{dz}{\theta-t} \quad ; \ z=y(\theta-t) \\ &= \frac{\eta\theta e^{\theta\beta}}{(\theta-t)^2}\Gamma(2,\beta(\theta-t)) = \frac{\eta\theta e^{\beta\theta}}{(\theta-t)^2}\left(1+\beta(\theta-t)\right) \end{aligned}$$

Therefore,

$$\begin{aligned} M_Y(t) &= \frac{\theta}{\delta\alpha+\eta}\left(\frac{-\delta\alpha e^{\beta t}}{(t-\theta)}+\frac{\eta\theta e^{\beta\ t}}{(\theta-t)^2}(1+\beta(\theta-t))+\eta\theta\beta e^{\beta t}(t-\theta)\right) \\ &= \frac{\theta e^{\beta t}}{(\delta\alpha+\eta)(t-\theta)^2}\left(\delta\alpha(\theta-t)+\eta\theta\right). \end{aligned}$$

Similarly, the characteristic function say, $\psi(t)$ of the FPGLD can be derived as follows:

$$\psi_Y(t) = E(e^{ity}) = \frac{\theta e^{\beta it}}{(\delta\alpha+\eta)(\theta-it)^2}\left(\delta\alpha(\theta-it)+\eta\theta\right). \tag{11}$$

3.3. Quantile function

The quantile function of FPGLD can be found by solving $F(y) = u, 0 < u < 1$. It is useful for the quantile estimations and for simulation studies. So, the u^{th} quantile function of FPGLD is derived as:

$$F(y) = 1 - \left(1 + \frac{\eta\theta(y-\beta)}{\delta\alpha+\eta}\right)e^{-\theta(y-\beta)} = u$$

$$\Rightarrow \left(\delta\alpha+\eta+\eta\theta(y-\beta)\right)e^{-\theta(y-\beta)} = (1-u)(\delta\alpha+\eta).$$

This equation can be rewritten as:

$$-\left(\frac{\delta\alpha}{\eta}+1+\theta(y-\beta)\right)e^{-\theta(y-\beta)-\frac{\delta\alpha}{\eta}-1} = \frac{(u-1)(\delta\alpha+\eta)}{\eta}e^{\frac{-\delta\alpha}{\eta}-1}.$$

Clearly $-\left(\frac{\delta\alpha}{\eta}+1+\theta(y-\beta)\right)$ is the negative branch of Lambert function, and one writes it symbolically as W_{-1} . Therefore, the quantile function of the FPGLD can be written in terms of the negative branch of the Lambert function as:

$$-\left(\frac{\delta\alpha}{\eta}+1+\theta(y-\beta)\right) = W_{-1}\left(\frac{(u-1)(\delta\alpha+\eta)}{\eta}e^{-\frac{\delta\alpha}{\eta}-1}\right).$$

Hence,

$$y = \beta - \frac{\delta\alpha+\eta}{\eta\theta} - \frac{1}{\theta}W_{-1}\left(\frac{(u-1)(\delta\alpha+\eta)}{\eta}e^{-\frac{\delta\alpha}{\eta}-1}\right); y > \beta, 0 < u < 1. \tag{12}$$

Then, the first three quartiles of the FPGLD can be derived by substituting $u = 0.25, 0.5$ and 0.75 in equation (12) and given by:

$$Q_1 = \beta - \frac{\delta\alpha+\eta}{\eta\theta} - \frac{1}{\theta}W_{-1}\left(\frac{(-0.75)(\delta\alpha+\eta)}{\eta}e^{-\frac{\delta\alpha}{\eta}-1}\right),$$

$$Q_2 = \beta - \frac{\delta\alpha+\eta}{\eta\theta} - \frac{1}{\theta}W_{-1}\left(\frac{(-0.5)(\delta\alpha+\eta)}{\eta}e^{-\frac{\delta\alpha}{\eta}-1}\right), \text{ and}$$

$$Q_3 = \beta - \frac{\delta\alpha+\eta}{\eta\theta} - \frac{1}{\theta}W_{-1}\left(\frac{(-0.25)(\delta\alpha+\eta)}{\eta}e^{-\frac{\delta\alpha}{\eta}-1}\right).$$

4. Reliability properties

In this section, we study some important reliability properties of FPGLD, namely the survival function/reliability function $S(y)$, hazard rate function/failure rate function $h(y)$, reversed hazard rate function $r(y)$, cumulative hazard rate function $H(y)$, mean residual life function $m(y)$, Lorenz curve $L(F(y))$, and Benferroni curve $B(F(y))$.

4.1. Hazard rate and mean residual life function

1. The survival function of equation (7) is defined as:

$$S(y) = 1 - F(y) = \left(1 + \frac{\eta\theta(y-\beta)}{\delta\alpha+\eta}\right)e^{-\theta(y-\beta)}; y > \beta. \tag{13}$$

   It is clear that, $S(\beta) = 1$ and $\lim_{y\to\infty} S(y) = 0$.

2. The hazard rate function(hrf) of the FPGLD is defined as:

$$h(y) = \lim_{\Delta y\to 0}\frac{P(y<Y<y+\Delta y|Y>y)}{\Delta y} = \frac{f(y)}{S(y)} = \frac{\theta\left(\delta\alpha+\eta\theta(y-\beta)\right)}{\delta\alpha+\eta+\eta\theta(y-\beta)}; y > \beta. \tag{14}$$

   Further, it can be seen that, $h(\beta) = \dfrac{\theta\delta\alpha}{\delta\alpha+\eta} = f(\beta)$ and $\lim_{y\to\infty} h(y) = \theta$.
   Figure 2 illustrates the hazard rate function of FPGLD at different parameter values. It is approximately same hazard rate shape of the TPLwLD.

3. The reversed hazard function of FPGLD is defined as:

$$r(y) = \lim_{\Delta y\to 0}\frac{P(y<Y<y+\Delta y|Y<y)}{\Delta y} = \frac{\theta\left(\delta\alpha+\eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}}{\delta\alpha+\eta-\left(1+\eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}}; y > \beta. \tag{15}$$

4. The cumulative hazard rate function of FPGLD is defined as:

$$H(y) = \int_\beta^y h(t)dt = -log[S(y)] = -log\left(1+\frac{\eta\theta(y-\beta}{\delta\alpha+\eta}\right)e^{-\theta(y-\beta)}. \tag{16}$$

5. The following theorem gives the mean residual life function of FPGLD.

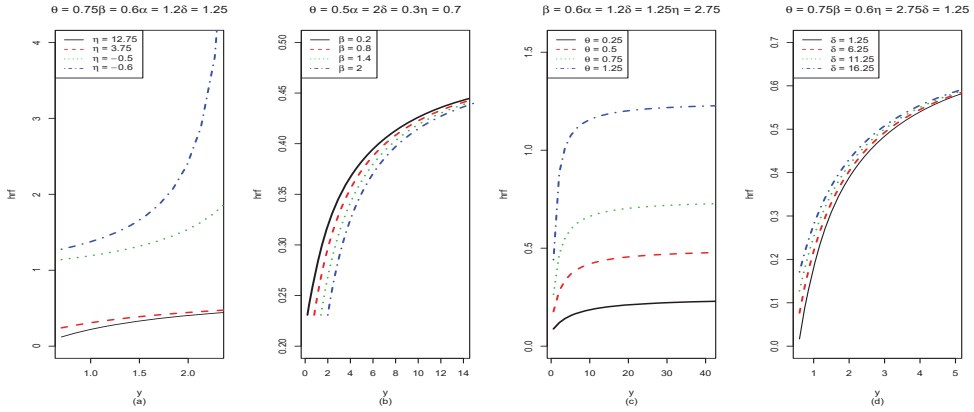


Figure 2: The hazard rate function of FPGLD at different parameter values

(a) θ, β, α, and δ are fixed, and η values are changed, (b) θ, α, δ and η are fixed, and β values are changed, (c) β, α, δ and η are fixed, and θ values are changed, (d) θ, β, η, and δ are fixed, and δ values are changed.

Theorem 3. The mean residual life function of FPGLD is given by:

$$m(y) = \frac{\delta\alpha + \eta(2 + \theta(y-\beta))}{\theta\left(\delta\alpha + \eta + \eta\theta(y-\beta)\right)}. \tag{17}$$

Proof.

$m(y) = \dfrac{1}{1-F(y)} \displaystyle\int_y^\infty t f(t)dt - y$, consider the integrals separately as follows:

$$\begin{aligned}
\int_y^\infty t f(t)dt &= \int_y^\infty t \frac{\theta}{\theta\alpha+\eta}\left(\delta\alpha + \eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}dt \\
&= \frac{\theta e^{\theta\beta}}{\delta\alpha+\eta}\left(\delta\alpha\int_y^\infty te^{-\theta t}dt + \eta\theta\int_y^\infty t^2e^{-\theta t}dt - \beta\eta\theta\int_y^\infty te^{-\theta t}dt\right) \\
&= \frac{\theta e^{\theta\beta}}{\delta\alpha+\eta}\left(\delta\alpha\left(\frac{\delta\alpha}{\theta^2}\Gamma(2,\theta y) + \frac{\eta\theta}{\theta^3}\Gamma(3,\theta y) - \frac{\beta\eta\theta}{\theta^2}\Gamma(2,\theta y)\right)\right. \\
&= \frac{(1+\theta y)(\delta\alpha+2\eta-\beta\eta\theta)+\eta(\theta y)^2}{\theta(\delta\alpha+\eta)}e^{-\theta(y-\beta)}.
\end{aligned}$$

Therefore,

$$m(y) = \frac{(1+\theta y)(\delta\alpha+2\eta-\beta\eta\theta)+\eta(\theta y)^2}{\theta\left(\delta\alpha+\eta+\eta\theta(y-\beta)\right)} - y = \frac{\delta\alpha+\eta(2+\theta(y-\beta))}{\theta\left(\delta\alpha+\eta+\eta\theta(y-\beta)\right)}.$$

Then, equation (17) satisfies the following properties:

$$m(y) \geq 0,\ m(\beta) = \frac{\delta\alpha + 2\eta}{\theta(\delta\alpha + \eta)},\ \text{and}\ \lim_{y\to\infty} m(y) = \frac{1}{\theta}.$$

4.2. Lorenz and Bonferroni curves

The concept of the Lorenz and Bonferroni curves were formulated by Bonferroni to measure the income inequalities. They are widely used in economics, reliability, demography, medicine, and insurance. The following theorem gives the function of the Lorenz curve of FPGLD.

Theorem 4. The Lorenz curve is defined for FPGLD as:

$$L(F(y)) = 1 - \frac{\int_y^\infty xf(x)dx}{\mu} = 1 - \frac{e^{-\theta(y-\beta)}\left[(1+y\theta)\left(\delta\alpha + \eta(2+\theta(y-\beta))\right)\right]}{\alpha\delta(1+\theta\beta) + \eta(2+\theta\beta)}. \quad (18)$$

Proof.

$$L(F(y)) = 1 - \frac{\int_y^\infty xf(x)dx}{\mu}.$$

Note that

$$\begin{aligned}
&\int_y^\infty xf(x)dx \\
&= \int_y^\infty x\frac{\theta}{\delta\alpha+\eta}\left(\delta\alpha + \eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}dx \\
&= \frac{\theta}{\delta\alpha+\eta}\left[e^{\theta\beta}\delta\alpha\int_y^\infty xe^{-\theta y}dx + e^{\theta\beta}\eta\theta\int_y^\infty x^2e^{-\theta y}dy - \eta\theta\beta e^{\theta\beta}\int_y^\infty xe^{-\theta y}dy\right] \\
&= \frac{\theta}{\delta\alpha+\eta}\left[\frac{e^{\theta\beta}\delta\alpha}{\theta^2}\Gamma(2,y\theta) + \frac{e^{\theta\beta}\eta\theta}{\theta^3}\Gamma(3,y\theta) - \frac{e^{\theta\beta}\eta\theta\beta}{\theta^2}\Gamma(2,y\theta)\right] \\
&= \frac{e^{-\theta(y-\beta)}}{(\delta\alpha+\eta)\theta}\left[(1+y\theta)\left(\delta\alpha + \eta(2+\theta(y-\beta))\right)\right].
\end{aligned}$$

Therefore, $$L(F(y)) = 1 - \frac{e^{-\theta(y-\beta)}\left[(1+y\theta)\left(\delta\alpha + \eta(2+\theta(y-\beta))\right)\right]}{\alpha\delta(1+\theta\beta) + \eta(2+\theta\beta)}.$$

Then, the function of Bonferroni curve for the FPGLD is defined as:

$$B(F(y)) = \frac{L(F(y))}{F(y)}$$

$$= \frac{(\delta\alpha+\eta)\left[\delta\alpha(1+\theta\beta)+\eta(2+\theta\beta)-e^{-\theta(y-\beta)}\left((1+y\theta)\left(\delta\alpha+\eta(2+\theta(y-\beta))\right)\right)\right]}{\left[(\delta\alpha+\eta)-\left((\delta\alpha+\eta)\eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}\right]\left[\alpha\delta(1+\theta\beta)+\eta(2+\theta\beta)\right]}. \tag{19}$$

4.3. Renyi entropy

The Renyi entropy (Renyi, 1961) is a basic uncertainty measure of a distribution say $H_R(\gamma)$ and an extension of Shannon entropy (Shannon et.al.,1949). This entropy is widely used in ecology and quantum information. The following theorem gives the Renyi entropy of FPGLD.

Theorem 5. The Renyi entropy of the FPGLD is given by:

$$H_R(\gamma) = \frac{1}{1-\gamma} log \int_\beta^\infty (f(y))^\gamma dy$$

$$= \frac{1}{1-\gamma} log\left(\frac{\theta^{\gamma-1}(\delta\alpha)^\gamma}{(\delta\alpha+\eta)^\gamma\gamma}\sum_{k=0}^{\gamma}\binom{\gamma}{k}\left(\frac{\eta}{\delta\alpha\gamma}\right)^k k\Gamma(k)\right); \gamma \geq 0, \gamma \neq 1. \tag{20}$$

Proof.

$$H_R(\gamma) = \frac{1}{1-\gamma} log \int_\beta^\infty (f(y))^\gamma dy$$

$$= \frac{1}{1-\gamma} log \int_\beta^\infty \left(\frac{\theta}{\delta\alpha+\eta}\left(\delta\alpha+\eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}\right)^\gamma dy$$

$$= \frac{1}{1-\gamma} log \int_\beta^\infty \frac{\theta^\gamma(\delta\alpha)^\gamma}{(\delta\alpha+\eta))^\gamma}\left(1+\frac{\eta\theta(y-\beta)}{\delta\alpha}\right)^\gamma e^{-\gamma\theta(y-\beta)} dy$$

$$= \frac{1}{1-\gamma} log \int_\beta^\infty \frac{\theta^\gamma(\delta\alpha)^\gamma}{(\delta\alpha+\eta))^\gamma}\sum_{k=0}^{\gamma}\binom{\gamma}{k}\left(\frac{\eta\theta(y-\beta)}{\delta\alpha}\right)^k e^{-\gamma\theta(y-\beta)} dy$$

$$= \frac{1}{1-\gamma} log\left(\frac{\theta^\gamma(\delta\alpha)^\gamma}{(\delta\alpha+\eta))^\gamma}\sum_{k=0}^{\gamma}\binom{\gamma}{k}\left(\frac{\eta\theta}{\delta\alpha}\right)^k \int_\beta^\infty (y-\beta)^k e^{-\gamma\theta(y-\beta)} dy\right)$$

$$= \frac{1}{1-\gamma} log\left(\frac{\theta^{\gamma-1}(\delta\alpha)^\gamma}{(\delta\alpha+\eta)^\gamma\gamma}\sum_{k=0}^{\gamma}\binom{\gamma}{k}\left(\frac{\eta}{\delta\alpha\gamma}\right)^k k\Gamma(k)\right).$$

5. The size-biased of FPGLD

The application of the size-biased distributions known as weighted distributions has been significantly used in forestry and wood product studies (Gove, 2003a) incorporating sampling probabilities that are proportional to weighted function $w(y)$. The size-biased distributions is defined as:

$$f_w(y) = \frac{w(y)f(y)}{E(w(y))}, \tag{21}$$

where, $w(y) = y^\gamma$ is a non-negative weighted function of order γ. Then, equation (21) can be rewritten as $f_Y^\gamma(y) = \dfrac{y^\gamma f(y)}{E(y^\gamma)}$,where $Y_r \sim f_Y^\gamma(y)$ is the size-biased random variable. The following theorem gives the density function for the size-biased FPGLD.

Theorem 6. The density function for sized-biased FPGLD is given by:

$$f_Y^\gamma(y) = y^\gamma \theta^{\gamma+1}\left(\frac{\delta\alpha+\eta\theta(y-\beta)}{A}\right)e^{-\theta y}; y>\beta, \gamma>0, \tag{22}$$

where, $A = \gamma\Gamma(\gamma,\theta\beta)\Big(\delta\alpha+\eta(\gamma+1-\theta\beta)\Big)+e^{-\theta\beta}(\theta\beta)^\gamma\Big(\delta\alpha+\eta(\gamma+1)\Big)$.

Proof.

$f_Y^\gamma(y) = \dfrac{y^\gamma f(y)}{E(y^\gamma)}$.

Note that

$$\begin{aligned}
E(y^\gamma) &= \int_\beta^\infty y^\gamma f(y)dy \\
&= \int_\beta^\infty y^\gamma \frac{\theta}{\delta\alpha+\eta}\Big(\delta\alpha+\eta\theta(y-\beta)\Big)e^{-\theta(y-\beta)}dy \\
&= \frac{\theta e^{\theta\beta}}{\delta\alpha+\eta}\left(\frac{\delta\alpha}{\theta^{\gamma+1}}\Gamma(\gamma+1,\theta\beta)+\frac{\eta\theta}{\theta^{\gamma+2}}\Gamma(\gamma+2,\theta\beta)-\frac{\eta\theta\beta}{\theta^{\gamma+1}}\Gamma(\gamma+1,\theta\beta)\right) \\
&= \frac{\theta e^{\theta\beta}}{(\delta\alpha+\eta)\theta^{\gamma+1}}\Bigg(\Gamma(\gamma,\theta\beta)\Big(\delta\alpha\gamma+\eta\gamma(\gamma+1)-\eta\theta\beta\gamma\Big)+\delta\alpha(\theta\beta)^\gamma e^{-\theta\beta}+ \\
&\qquad\qquad \eta(\gamma+1)(\theta\beta)^\gamma e^{-\theta\beta}\Bigg) \\
&= \frac{\theta e^{\theta\beta}}{(\delta\alpha+\eta)\theta^\gamma}\Bigg(\gamma\Gamma(\gamma,\theta\beta)\Big(\delta\alpha+\eta(\gamma+1-\theta\beta)\Big)+ \\
&\qquad\qquad e^{-\theta\beta}(\theta\beta)^\gamma\Big(\delta\alpha+\eta(\gamma+1)\Big)\Bigg).
\end{aligned}$$

Therefore,

$$f_Y^{\gamma}(y) = \frac{\dfrac{y^{\gamma}\theta}{\delta\alpha+\eta}\left(\delta\alpha+\eta\theta(y-\beta)\right)e^{-\theta(y-\beta)}}{\dfrac{\theta e^{\theta\beta}}{(\delta\alpha+\eta)\theta^{\gamma}}\left(\gamma\Gamma(\gamma,\theta\beta)\left(\delta\alpha+\eta(\gamma+1-\theta\beta)\right)+e^{-\theta\beta}(\theta\beta)^{\gamma}\left(\delta\alpha+\eta(\gamma+1)\right)\right)}$$

$$= y^{\gamma}\theta^{\gamma+1}\left(\frac{\delta\alpha+\eta\theta(y-\beta)}{\gamma\Gamma(\gamma,\theta\beta)(\delta\alpha+\eta(\gamma+1-\theta\beta))+e^{-\theta\beta}(\theta\beta)^{\gamma}(\delta\alpha+\eta(\gamma+1))}\right)e^{-\theta y}.$$

The length biased probability density function can be derived from size-biased pdf of FPGLD by substituting $\gamma = 1$. The length-biased probability density function is given by:

$$f_Y^{1}(y) = y\theta^{2}\left(\frac{\delta\alpha+\eta\theta(y-\beta)}{\delta\alpha+\eta(2-\theta\beta)+\theta\beta(\delta\alpha+2\eta)}\right)e^{-\theta(y-\beta)}; y>\beta,\gamma>0. \qquad (23)$$

6. Parameter estimation and inference

In this section, the parameter estimation and inference are given. In the parameter estimation of FPGLD, the method of moment estimators (MME) and maximum likelihood estimators (MLE) methods are introduced.

6.1. Method of moment estimation

The method of moment estimators can be derived by equating the raw-moments, say $\mu_r^{'}$, to the sample moments, say $\dfrac{\sum_{i=1}^{n} y_i^r}{n}, r = 1,2,3,4,5$

Then, we need to solve the following system of non-linear equations.

$$n\left(\delta\alpha(1+\theta\beta)+\eta(2+\theta\beta)\right)-\theta(\delta\alpha+\eta)\sum_{i=1}^{n}y_i=0$$

$$n\left(\delta\alpha(2+\theta\beta(2+\theta\beta))+\eta(6+\theta\beta(4+\theta\beta))\right)-\theta^2(\delta\alpha+\eta)\sum_{i=1}^{n}y_i^2=0$$

$$n\left(\delta\alpha\left(6+\theta\beta(6+\theta\beta(3+\theta\beta))\right)+\eta\left(24+\theta\beta(18+\theta\beta(6+\theta\beta)\right)\right)-$$
$$\theta^3(\delta\alpha+\eta)\sum_{i=1}^{n}y_i^3=0$$

$$n\left(\delta\alpha\left(24+\theta\beta(24+\theta\beta(12+\theta\beta(4+\theta\beta)))\right)+\eta\left(120+\beta\theta(96+\right.\right.$$

$$\theta\beta(36+\theta\beta(8+\theta\beta)))\Bigg)\Bigg)-\theta^4(\delta\alpha+\eta)\sum_{i=1}^{n}y_i^4=0$$

$$n\Bigg(\delta\alpha\Bigg(120+\theta\beta(120+\theta\beta(60+\theta\beta(20+\theta\beta(5+\theta\beta))))\Bigg)+\eta\Bigg(720+$$

$$\theta\beta(600+\theta\beta(240+\theta\beta(60+\theta\beta(10+\theta\beta))))\Bigg)\Bigg)-\theta^5(\delta\alpha+\eta)\sum_{i=1}^{n}y_i^5=0$$

6.2. Maximum likelihood estimation

Let $y_1, y_2, ...y_n$ be identically and independently distributed random variables from FPGLD with the likelihood function of the i^{th} sample value y_i as:

$$L(\theta,\beta,\alpha,\delta,\eta|y_i)=\frac{\theta}{\delta\alpha+\eta}\Bigg(\delta\alpha+\eta\theta(y_i-\beta)\Bigg)e^{-\theta(y_i-\beta)}.$$

Then, the log-likelihood function is given by:

$$log\Bigg(L(\theta,\beta,\alpha,\delta,\eta|y_i)\Bigg)$$

$$=l=nlog\theta+\sum_{i=1}^{n}log(\delta\alpha+\eta\theta(y_i-\beta))-\sum_{i=1}^{n}\theta(y_i-\beta)-nlog(\delta\alpha+\eta).$$

The maximum likelihood estimators (MLE), say $\hat{\theta},\hat{\beta},\hat{\alpha},\hat{\delta},\hat{\eta}$ can be derived by equating the partial derivatives of the l with respect to each parameter to zero. Then, we have:

$$\frac{\partial l}{\partial\theta}=\frac{n}{\theta}+\sum_{i=1}^{n}\frac{\eta(y_i-\beta)}{\delta\alpha+\eta\theta(y_i-\beta)}-\sum_{i=1}^{n}(y_i-\beta)=0,$$

$$\frac{\partial l}{\partial\beta}=\sum_{i=1}^{n}\frac{-\eta\theta}{\delta\alpha+\eta\theta(y_i-\beta)}+n\theta=0,$$

$$\frac{\partial l}{\partial\alpha}=\sum_{i=1}^{n}\frac{\delta}{\delta\alpha+\eta\theta(y_i-\beta)}-\frac{n\delta}{\delta\alpha+\eta}=0,$$

$$\frac{\partial l}{\partial\delta}=\sum_{i=1}^{n}\frac{\alpha}{\delta\alpha+\eta\theta(y_i-\beta)}-\frac{n\alpha}{\delta\alpha+\eta}=0, \text{ and}$$

$$\frac{\partial l}{\partial\eta}=\sum_{i=1}^{n}\frac{\theta(y_i-\beta)}{\delta\alpha+\eta\theta(y_i-\beta)}-\frac{n}{\delta\alpha+\eta}=0.$$

The asymptotic confidence intervals for the parameters of FPGLD, say $\theta,\beta,\alpha,\delta,\eta$ are derived under the regularity conditions of the maximum likelihood estimations. The second partial derivatives of the log-likelihood function are:

$$\frac{\partial^2 l}{\partial\theta^2}=\frac{-n}{\theta^2}+\sum_{i=1}^{n}\frac{-\eta^2(y_i-\beta)^2}{(\delta\alpha+\eta\theta(y_i-\beta))^2},$$

$$\frac{\partial^2 l}{\partial \beta^2} = \sum_{i=1}^{n} \frac{(\eta\theta)^2}{(\delta\alpha+\eta\theta(y_i-\beta))^2},$$
$$\frac{\partial^2 l}{\partial \alpha^2} = \sum_{i=1}^{n} \frac{-\delta^2}{(\delta\alpha+\eta\theta(y_i-\beta))^2} + \frac{n\delta^2}{(\delta\alpha+\eta)^2},$$
$$\frac{\partial^2 l}{\partial \delta^2} = \sum_{i=1}^{n} \frac{-\alpha^2}{(\delta\alpha+\eta\theta(y_i-\beta))^2} + \frac{n\alpha^2}{(\delta\alpha+\eta)^2},$$
$$\frac{\partial^2 l}{\partial \eta^2} = \sum_{i=1}^{n} \frac{-\theta^2(y_i-\beta)^2}{(\delta\alpha+\eta\theta(y_i-\beta))^2} + \frac{n}{(\delta\alpha+\eta)^2},$$
$$\frac{\partial^2 l}{\partial \theta\partial \beta} = \sum_{i=1}^{n} \frac{-\eta\delta\alpha}{(\delta\alpha+\eta\theta(y_i-\beta))^2} + n,$$
$$\frac{\partial^2 l}{\partial \theta\partial \alpha} = \sum_{i=1}^{n} \frac{-\eta\delta(y_i-\beta)}{(\delta\alpha+\eta\theta(y_i-\beta))^2},$$
$$\frac{\partial^2 l}{\partial \theta\partial \delta} = \sum_{i=1}^{n} \frac{-\eta\alpha(y_i-\beta)}{(\delta\alpha+\eta\theta(y_i-\beta))^2},$$
$$\frac{\partial^2 l}{\partial \theta\partial \eta} = \sum_{i=1}^{n} \frac{\delta\alpha(y_i-\beta)}{(\delta\alpha+\eta\theta(y_i-\beta))^2},$$
$$\frac{\partial^2 l}{\partial \beta\partial \alpha} = \sum_{i=1}^{n} \frac{\eta\theta\delta}{(\delta\alpha+\eta\theta(y_i-\beta))^2},$$
$$\frac{\partial^2 l}{\partial \beta\partial \delta} = \sum_{i=1}^{n} \frac{\eta\theta\alpha}{(\delta\alpha+\eta\theta(y_i-\beta))^2},$$
$$\frac{\partial^2 l}{\partial \beta\partial \eta} = \sum_{i=1}^{n} \frac{-\eta\delta\theta}{(\delta\alpha+\eta\theta(y_i-\beta))^2},$$
$$\frac{\partial^2 l}{\partial \alpha\partial \delta} = \sum_{i=1}^{n} \frac{\eta\theta(y_i-\beta)}{(\delta\alpha+\eta\theta(y_i-\beta))^2} - \frac{n\eta}{(\delta\alpha+\eta)^2},$$
$$\frac{\partial^2 l}{\partial \alpha\partial \eta} = \sum_{i=1}^{n} \frac{-\delta\theta(y_i-\beta)}{(\delta\alpha+\eta\theta(y_i-\beta))^2} + \frac{n\delta}{(\delta\alpha+\eta)^2}, \text{ and}$$
$$\frac{\partial^2 l}{\partial \delta\partial \eta} = \sum_{i=1}^{n} \frac{-\alpha\theta(y_i-\beta)}{(\delta\alpha+\eta\theta(y_i-\beta))^2} + \frac{n\alpha}{(\delta\alpha+\eta)^2}.$$

Let $\hat{p} = (\hat{\theta}, \hat{\beta}, \hat{\alpha}, \hat{\delta}, \hat{\eta})$ be MLE of p. By the asymptotic theory the estimators are asymptotically normal 5-variate with mean $(\theta, \beta, \alpha, \delta, \eta)$, and observed information matrix is given

by:

$$I(y)=\begin{pmatrix} -\frac{\partial^2 l}{\partial\theta^2} & -\frac{\partial^2 l}{\partial\theta\partial\beta} & -\frac{\partial^2 l}{\partial\theta\partial\alpha} & -\frac{\partial^2 l}{\partial\theta\partial\delta} & -\frac{\partial^2 l}{\partial\theta\partial\eta} \\ -\frac{\partial^2 l}{\partial\beta\partial\theta} & -\frac{\partial^2 l}{\partial\beta^2} & -\frac{\partial^2 l}{\partial\beta\partial\alpha} & -\frac{\partial^2 l}{\partial\beta\partial\delta} & -\frac{\partial^2 l}{\partial\beta\partial\eta} \\ -\frac{\partial^2 l}{\partial\alpha\partial\theta} & -\frac{\partial^2 l}{\partial\alpha\partial\beta} & -\frac{\partial^2 l}{\partial\alpha^2} & -\frac{\partial^2 l}{\partial\alpha\partial\delta} & -\frac{\partial^2 l}{\partial\alpha\partial\eta} \\ -\frac{\partial^2 l}{\partial\delta\partial\theta} & -\frac{\partial^2 l}{\partial\delta\partial\beta} & -\frac{\partial^2 l}{\partial\delta\partial\alpha} & -\frac{\partial^2 l}{\partial\delta^2} & -\frac{\partial^2 l}{\partial\delta\partial\eta} \\ -\frac{\partial^2 l}{\partial\eta\partial\theta} & -\frac{\partial^2 l}{\partial\eta\partial\beta} & -\frac{\partial^2 l}{\partial\eta\partial\alpha} & -\frac{\partial^2 l}{\partial\eta\partial\delta} & -\frac{\partial^2 l}{\partial\eta^2} \end{pmatrix}$$

at $\theta=\hat{\theta},\beta=\hat{\beta},\alpha=\hat{\alpha},\delta=\hat{\delta},\eta=\hat{\eta}$. By the asymptotic theory, the estimates are approximately multivariate normal. Therefore, the $(1-\alpha)100\%$ confidence interval for the parameters $\theta,\beta,\alpha,\delta,\eta$ are given by:

$$\hat{\theta}\pm z_{a/2}\sqrt{var(\hat{\theta})}, \qquad \hat{\beta}\pm z_{a/2}\sqrt{var(\hat{\beta})}, \qquad \hat{\alpha}\pm z_{a/2}\sqrt{var(\hat{\alpha})},$$

$$\hat{\delta}\pm z_{a/2}\sqrt{var(\hat{\delta})}, \qquad \hat{\eta}\pm z_{a/2}\sqrt{var(\hat{\eta})}$$

wherein, the $var(\hat{\theta}), var(\hat{\beta}), var(\hat{\alpha}), var(\hat{\delta})$, and $var(\hat{\eta})$ are the variance of $\hat{\theta},\hat{\beta},\hat{\alpha},\hat{\delta}$, and $\hat{\eta}$, respectively, and can be derived by diagonal elements of $I^{-1}(y)$ and $z_{a/2}$ is the critical value at a level of significance.

7. Applications

In this section, we perform a simulation study to examine the behavior of FPGLD's parameter estimates by MLE method, performance of location parameter β, and performance of scale parameter θ when it is incoperated in the mixing proportion. Further,the real-world applications are used to study the performance of the FPGLD with TPLwLD, LD, TwPLD, QLD, SD, and ThPLD. The estimates of the parameters for each distribution has been derived by the MLE method.

7.1. Simulation study

7.1.1 Performance of maximum likelihood method

Here, we discuss the simulation study for the unknown parameter estimations of FPGLD by maximum likelihood method for different sample sizes. The combination of parameter values are set to $\theta=0.5,\delta=0.1,\alpha=0.2,\eta=0.4,\beta=1.5$. Then, the steps of the simulation study are given below:

1. Generate 1000 samples for each of the sample size, $n=20,n=50,n=80$ and $n=100$ using equation (12).

2. Calculate the average MSE for the parameters of FPGLD using the equation

$$\text{MSE}(p)=\frac{\sum_{i=1}^{1000}(\hat{p}_i-p)^2}{1000}, \text{ where } p=(\theta,\delta,\alpha,\eta,\beta), \text{ represents the parameter set.}$$

Table 3 summarizes the average mean square error(MSE) values of FPGLD at different sample sizes. According to Table 3, the average MSE values for parameters $\theta,\beta,\ \alpha,\delta$ and η decreases when sample size increases. Further, it is notable that decreasing rates of average MSE for the parameters θ and β are higher than decreasing rates of average MSE for the parameters δ,α, and η. This indicates that the parameters of the mixing components, θ and β are highly sensitive than the parameters δ,α, and η that are introduced from the latent variable distribution in the unknown parameter estimations for this model.

Table 3. Simulation results for the average MSE values

Parameter	MSE			
	$n=20$	$n=50$	$n=80$	$n=100$
$\theta=0.5$	0.014061	0.004528	0.004479	0.002029
$\delta=0.1$	0.009918	0.009916	0.009871	0.009870
$\alpha=0.2$	0.039809	0.039541	0.039524	0.039456
$\eta=0.4$	0.158916	0.158795	0.156714	0.156594
$\beta=1.5$	0.135841	0.040434	0.025789	0.019406

7.1.2 Performance of the FPGLD when the location parameter $\beta=0$

In this subsection, the performance of the FPGLD is examined by a simulation study when the location parameter $\beta=0$. It was done by comparing FPGLD $(\theta,\beta,\alpha,\delta,\eta)$ and FPGLD$(\theta,\beta=0,\alpha,\delta,\eta)$ for selected values of skewness, EK, and Fano factor. The study is designed as follows:

1. Generate random samples of size, $n=150$ from FPGLD $(\theta,\beta,\alpha,\delta,\eta)$ with various skewness (SK), Excees kurtosis (EK), and Fano factor (FF) values by setting the parameter values.

2. Fit the FPGLD $(\theta,\beta,\alpha,\delta,\eta)$ and FPGLD $(\theta,\beta=0,\alpha,\delta,\eta)$ to the generated data sets.

3. Calculate the differences of negative log-likelihood $(-2logL)$ values for every generated data sets as:
$$\Big(-2logL(\text{FPGLD }(\theta,\beta=0,\alpha,\delta,\eta))\Big)-\Big(-2logL(\text{FPGLD }(\theta,\beta,\alpha,\delta,\eta))\Big)$$

The table 6 (Appendix) summarizes the differences of $-2logL$ values between FPGLD $(\theta,\beta=0,\alpha,\delta,\eta)$ and FPGLD $(\theta,\beta,\alpha,\delta,\eta)$. We may notice that $-2logL$ difference is decreasing when skewness, EK, and Fano factor values are increasing. Hence, this simulation study reveals that the inclusion of the location parameter in this distribution resists the flexibility to cover the higher skewness, EK, and Fano factor values.

7.1.3 Performance of scale parameter θ when that is incorporated in the mixing proportion

Here, we compare the LD and QLD using a simulation study since they just differ in their defined mixing proportion. i-e) while the LD′s mixing proportion is defined incorporating the scale parameter of the mixing component θ, the QLD′s mixing proportion is not incorporated with θ. The similar steps that have designed in section 7.1.2 are followed and $-2logL$ differences are calculated as: $(-2logL(\text{QLD}(\theta,\alpha)))$-$(-2logL(\text{LD}(\theta)))$. Table 7 summarizes the differences of $-2logL$ values between QLD and LD. We may notice that $-2logL$ difference is decreasing when skewness, EK, and Fano factor values are increasing. The results indicates that the incorporation of the scale parameter in the mixing proportion in LD resists the flexibility to cover the higher skewness, EK, and Fano factor values.

7.2. Real-world applications

The performance of the FPGLD with respect to the sub-models is now considered by using real-world applications. The negative log-likelihood $(-2logL)$, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Kolmogorov-Smirnov Statistics (K-S Statistics) are utilized to compare the performance of distributions. The estimates of the parameters for each distribution has been derived by the MLE method. The following four real-world data sets have been fitted to the distributions for the goodness of fit of distributions.

Data set 1: This data set is the relief times (in minutes) of the 20 patients receiving an analgesic and reported by Gross and Clark (1975).
1.1, 1.4, 1.3, 1.7, 1.9, 1.8, 1.6, 2.2, 1.7, 2.7, 4.1, 1.8, 1.5 1.2, 1.4, 3.0 , 1.7, 2.3, 1.6, 2.0.

Data set 2: The data set reported by Bjerkedal (1960) that represents the survival times(in days) of 72 guinea pigs infected with virulent tubercle bacilli is given below:
12, 15, 22, 24, 24, 32, 32, 33, 34, 38, 38, 43, 44, 48, 52, 53, 54, 54, 55, 56, 57, 58, 58, 59, 60, 60, 60, 60, 61, 62, 63, 65, 65, 67, 68, 70, 70, 72, 73, 75, 76, 76, 81, 83, 84, 85, 87, 91, 95, 96, 98, 99, 109, 110, 121, 127, 129, 131, 143, 146, 146, 175, 175, 211, 233, 258, 258, 263, 297, 341, 341, 376.
Data set 3: The data set was given by Fuller et. al. (1994) that represents the strength data of glass of the aircraft window is given below:
18.83, 20.80, 21.657, 23.03, 23.23, 24.05, 24.321, 25.50, 25.52, 25.80, 26.69, 26.77, 26.78, 27.05, 27.67, 29.90, 31.11, 33.20, 33.73, 33.76, 33.89, 34.76, 35.75, 35.91, 36.98, 37.08, 37.09, 39.58, 44.045, 45.29, 45.381.
Data set 4: The data set was used by Lawless (1982) and the data were recorded in tests on the endurance of deep groove ball bearings. The corresponding random variable is the number of million revolutions before failure for each of the 23 ball bearings in the life tests.
17.88, 28.92, 33, 41.52, 42.12, 45.6, 48.8, 51.84, 51.96, 54.12, 55.56, 67.8, 68.44, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.4.

Some of the important statistical measures for the data set 1 to 4 are summarized in Table 4:

Table 4. Statistical measures for data set 1 to 4

Data	Sample size	Minimum value	Mean	Median	Skewness	EK	Fano factor
Data 1	20	1.100	1.900	1.700	1.720	2.924	0.261
Data 2	72	12.000	99.819	70.000	1.796	2.614	65.920
Data 3	31	18.830	30.811	29.900	0.405	-0.713	1.708
Data 4	23	17.880	72.230	67.800	0.941	0.488	19.448

Figure 3 (Appendix) shows the density plots that compare the fitted densities of each model with the empirical histogram of the real-world data sets. We can observe that the fitted densities for the FPGLD and TPLwLD show a closer fit with the empirical distributions for real-data sets 1, 3 and 4, and both fitted densities are approximately the same. Further, QLD shows a closer fit with the empirical distribution for the data set 2. Table 8 (Appendix) shows the values of $-2logL$, AIC, BIC and K-S statistics and critical values of the K-S statistics. According to Table 8, we may note that AIC and BIC values increase when the number of parameters of the distributions increases. Therefore, we use $-2logL$ values and K-S statistics for the comparison of all models.

Based on the minimum $-2logL$, and the significant results by K-S statistics, FPGLD and TPLwLD provide a better fit than all other sub-models for the data sets 1, 3, and 4. Data set 1, 3, and 4 have considerably smaller skewness and EK values or smaller Fano factor value. There is no difference between the log-likelihood values of FPGLD and TPLwLD for these data sets. This indicates that $\delta\alpha = \theta$ and the performance of both distributions are the same. Further, when we compare TPLwLD and TwPLD, the likelihood ratio (LR) test statistics for the hypothesis testing $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$ for data 1, 3 and 4 are 22.686, 53.413, and 16.663, respectively, and all are greater than $\chi^2_{1,0.05} = 3.841$. These results indicate the importance of the location parameter in such type of lifetime data analysis than introducing new shape parameters from the latent variable distribution to give different weights.

On the other hand, it is notable that in most of the real-data applications, the performance of TwPLD, QLD, and ThPLD are the same except for data set 2, where the QLD shows the minimum $-2logL$ significant result by K-S statistic. The data set 2 has considerably higher skewness, EK and Fano factor values. To show the effect of the higher skewness, EK, and Fano factor values, data set 5 (Appendix) was also used to fit the distributions, and Table 5 summarizes the results of the goodness of fittest. These results indicate that QLD performs well than other distributions for the data sets with considerably higher skewness, EK and Fano factor values. A possible reason may be that it has the flexibility with the format $\delta\alpha = \delta \neq \theta$ and exclusion of the location parameter. Therefore, when developing a best-fitted distribution for the data sets that have higher skewness, EK, and Fano factor values, it is recommended to use proper mixing weights and mixing components without a location parameter.

We hope these findings could be helpful for the researchers when they develop a new Lindley family distribution.

Table 5. $-2logL$, and K-S statistics of the NGAD and LwLD for different data sets with various EK values

Data	Distribution	Sample size	Skewness	EK	Fano factor	$-2logL$	AIC
	FPGLD					252.416	262.416
	TPLwLD					252.416	258.416
	TwPLD					266.401	270.401
Data 5	QLD	60	2.437	7.018	2.547	**250.920**	**254.920**
	ThPLD					266.401	272.401
	LD					259.171	261.171
	SD					257.096	259.096

8. Conclusions

In this paper, we have introduced a new five-parameter generalized Lindley distribution(FPGLD) based on exponential and gamma mixtures with different mixing proportions and done a comparison study with its sub-models. The FPGLD generalizes the Lindley distribution with location parameter(TPLwLD), Quasi Linley distribution (QLD), Two-parameter Lindley distribution (TwPLD), Three-parameter Lindley distribution (ThPLD), Shanker distribution (SD), and classical Lindley distribution (LD). Hence, using FPGLD a researcher can compare the other existing lifetime distributions without considering its sub-models separately. The statistical properties and estimates of parameters are obtained for the FPGLD and compared it with its sub-models.

Acknowledgements

We thank the Postgraduate Institute of Science, University of Peradeniya, Sri Lanka for providing all facilities to do this research and editor-in-chief and the reviewers for their comments, which significantly improved the paper.

REFERENCES

BJERKEDAL, T., (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli, *American Journal of Hygiene*, 72(1), pp. 130–148.

FULLER, E. R., FREIMAN, S. W., QUINN, J. B., QUINN, G., CARTER, W., (1994). Fracture mechanics approach to the design of glass aircraft windows - A case study, *Proceedings of SPIE - The International Society for Optical Engineering*, pp. 419–430.

GHITANY, M. E., ATIEH, B., NADARAJAH, S., (2008). Lindley distribution and its Applications, *Mathematics and Computers in Simulation*, 78(4), pp. 493–506.

GOVE, J. H., (2003a). Estimation and applications of size-biased distributions in forestry, *Modeling forest systems*. Edited by A. Amaro, D. Reed, and P. Soares. CABI Publishing, Walligford, UK., pp. 201–212.

GROSS, A. J., CLARK,V.A., (1975). Survival Distributions, *Reliability Applications in the Biometrical Sciences*, John Wiley, New York, USA.

HIBATULLAH, R., WIDYANINGSIH, Y., ABDULLAH, S., (2018). Marshall - Olkin extended power Lindley distribution with application, *J. Ris. and Ap. Mat.*, 2(2), pp. 84 – 92.

LAWLESS, J. F., (1982). Statistical models and methods for lifetime data, John Wiley and Sons, New York, USA.

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

LINDLEY, D. V., (1965). Introduction to Probability and Statistics from Bayesian viewpoint, part II, Inference, *Cambridge university press, New York*.

MONSEF, M. M. E. A., (2016). A new Lindley distribution with location parameter, *Communications in Statistics-Theory and Methods*, 45(17), pp. 5204–5219.

NICHOLS, M. D., PADGETT, W. J., (2006). A bootstrap control chart for Weibull percentiles, *Quality and Reliability Engineering International*, 22(2), pp. 141–151.

RENYI, A. (1961). On measures of entropy and information, In Fourth Berkeley Symposium on Mathematical Statistics and Probability, pp. 547–561.

SHANKER, R., MISHR, A., (2013a). A Quasi Lindley Distribution, *African Journal of Mathematics and Computer Science Research*, 6(4), pp. 64 –71.

SHANKER, R., MISHR, A., (2013b). A two-parameter Lindley distribution, *Statistics in Transition new Series*, 14(1), pp. 45–56.

SHANKER, R., SHARMA, S., SHANKER, R., (2013). A Two-Parameter Lindley Distribution for Modeling Waiting and Survival Times Data, *Applied Mathematics*, 4, pp. 363–368.

SHANKER, R., (2015). Shanker Distribution and Its Applications, *International Journal of Statistics and Applications*, 5(6), pp. 338–348.

SHANKER, R., SHUKLA, K. K., SHANKER, R., TEKIE, A. L., (2017). A Three-parameter Lindley Distribution, *American Journal of Mathematics and Statistics*, 7(1), pp. 15–26.

SHANNON, C., WEAVER, W., (1949). *The mathematical theory of communication*, Chicago: University of Illinois Press.

APPENDIX

Table 6. Differences of $-2logL$ values between FPGLD $(\theta, \beta = 0, \alpha, \delta, \eta)$ and FPGLD $(\theta, \beta, \alpha, \delta, \eta)$

$\downarrow SK(EK)$	FF $\rightarrow$						
	6.70	7.70	12.70	15.80	20.50	28.30	34.60
0.918 (0.496)	37.442	32.105	19.087	15.009	11.112	7.656	6.147
0.920 (0.498)	34.758	29.892	16.788	12.822	9.151	5.896	4.435
0.928 (0.509)	31.223	26.807	13.746	9.840	6.392	3.392	2.239
0.969 (0.571)	28.037	23.353	10.201	6.563	3.370	1.160	0.421
1.044 (0.709)	27.116	22.081	8.693	5.069	2.164	0.368	0.014
1.102 (0.837)	27.108	22.031	8.321	4.656	1.822	0.189	0.002
1.208 (1.107)	27.082	22.003	8.242	4.540	1.634	0.089	0.001

Table 7. Differences of $-2logL$ values between QLD (θ, α) and LD (θ)

$\downarrow SK(EK)$	FF $\rightarrow$						
	6.70	7.70	12.70	15.80	20.50	28.30	34.60
0.918 (0.496)	71.804	70.212	64.911	62.578	59.815	56.668	54.969
0.920 (0.498)	70.585	69.019	63.197	60.659	57.701	54.250	52.264
0.928 (0.509)	68.506	66.906	60.250	57.248	53.771	49.493	47.189
0.969 (0.571)	65.186	63.199	54.971	51.261	46.605	41.265	38.026
1.044 (0.709)	62.391	60.046	50.563	46.135	40.695	34.049	29.988
1.102 (0.837)	60.962	58.641	48.365	43.576	37.731	30.405	27.083
1.208 (1.107)	57.306	52.694	41.795	36.615	30.199	22.009	20.084

Data set 5 (Hibatullah.et.al.,(2018): average wind speed per month.
1.04525, 2.78426, 2.54918, 6.90446, 2.46577, 2.83905, 2.09819, 0.47927, 1.41378, 4.77888, 2.28740 ,4.79976, 1.32359 ,1.71967, 3.52471, 0.38095, 10.9028 ,1.38314 ,1.89628, 1.03046, 2.44529, 13.1893 ,2.16495 ,3.78884, 2.20266, 0.71543 ,16.4941, 3.14792, 7.72747, 2.84926 ,2.68460, 5.45061, 1.32353, 1.48582, 5.10102, 3.00342, 1.77735 ,4.88295, 0.80280, 5.02584, 1.50003, 2.01266, 1.74341, 3.11761,0.80668 ,2.65187, 4.64156, 1.65586, 6.95507, 5.83996 ,3.33749, 1.27453, 2.29751 ,3.26983,2.65993, 4.53323, 5.73434, 2.09596, 1.52554, 2.71060.

Table 8. MLE, AIC, BIC, K-S statistics and its critical value of the fitted distributions

Data	Model	MLE	$-2logL$	AIC	BIC	K-S Statistic	Critical value
Data 1	FPGLD	$\hat{\theta}=2.284,\hat{\beta}=1.024,\hat{\alpha}=2.655*10^{-10}$, $\hat{\delta}=7.554*10^{-6},\hat{\eta}=2.319*10^{-5}$	30.988	40.988	45.966	0.121	0.304
	TPLwLD	$\hat{\theta}=2.284,\hat{\beta}=1.024,\hat{\alpha}=4.982*10^{3}$	**30.988**	**36.988**	**39.975**	**0.121**	
	TwPLD	$\hat{\theta}=0.526,\hat{\alpha}=-5.583*10^{-6}$	65.674	69.674	71.666	0.390	
	QLD	$\hat{\theta}=0.527,\hat{\alpha}=1.322*10^{3}$	65.674	69.674	71.666	0.390	
	ThPLD	$\hat{\theta}=0.053,\hat{\beta}=1.025,\hat{\alpha}=8.175*10^{3}$	65.674	71.674	74.661	0.390	
	LD	$\hat{\theta}=0.816$	60.499	62.499	63.495	0.341	
	SD	$\hat{\theta}=0.795$	59.788	61.788	62.784	0.309	
Data 2	FPGLD	$\hat{\theta}=0.021,\hat{\beta}=6.886,\hat{\alpha}=7.174*10^{-12}$ $\hat{\delta}=1.635*10^{-7},\hat{\eta}=4.475*10^{-8}$,	786.511	796.511	807.894	0.145	0.160
	TPLwLD	$\hat{\theta}=0.022,\hat{\beta}=6.891,\hat{\alpha}=225.796$	786.511	792.511	799.341	0.145	
	TwPLD	$\hat{\theta}=0.010,\hat{\alpha}=2.407*10^{-15}$	806.844	810.884	815.438	0.198	
	QLD	$\hat{\theta}=0.023,\hat{\alpha}=-0.241$	**783.145**	**787.145**	**791.698**	**0.152**	
	ThPLD	$\hat{\theta}=0.010,\hat{\beta}=2.356*10^{-14},\hat{\alpha}=145.12$	806.844	812.884	819.714	0.198	
	LD	$\hat{\theta}=0.020$	789.039	791.039	793.316	0.133	
	SD	$\hat{\theta}=0.020$	788.570	790.570	792.847	0.133	
Data 3	FPGLD	$\hat{\theta}=0.159,\hat{\beta}=1.821,\hat{\alpha}=6.031*10^{-4}$ $\hat{\delta}=1.519*10^{-12},\hat{\eta}=7.327*10^{-5}$	209.115	219.115	226.285	0.106	0.244
	TPLwLD	$\hat{\theta}=0.159,\hat{\beta}=18.211,\hat{\alpha}=566.720$	**209.115**	**215.115**	**219.418**	**0.106**	
	TwPLD	$\hat{\theta}=0.032,\hat{\alpha}=1.552*10^{-18}$	274.528	278.528	281.396	0.426	
	QLD	$\hat{\theta}=0.032,\hat{\alpha}=2.274*10^{3}$	274.528	278.528	281.396	0.426	
	ThPLD	$\hat{\theta}=0.003,\hat{\beta}=1.352*10^{-13}$, $\hat{\beta}=1.352*10^{-13},\hat{\alpha}=13.603$	274.528	280.529	284.831	0.426	
	LD	$\hat{\theta}=0.063$	253.988	255.988	257.422	0.333	
	SD	$\hat{\theta}=0.065$	252.343	254.353	255.787	0.326	
Data 4	FPGLD	$\hat{\theta}=0.035,\hat{\beta}=14.277,\hat{\alpha}=2.969*10^{-12}$ $\hat{\delta}=1.320*10^{-6},\hat{\eta}=6.005*10^{-7}$	226.210	236.210	241.887	0.090	0.284
	TPLwLD	$\hat{\theta}=0.035,\hat{\beta}=14.281,\hat{\alpha}=239.790$	**226.210**	**232.210**	**235.616**	**0.090**	
	TwPLD	$\hat{\theta}=0.014,\hat{\alpha}=1.635*10^{-14}$	242.873	246.874	249.141	0.263	
	QLD	$\hat{\theta}=0.014,\hat{\alpha}=3.815*10^{4}$	242.873	246.874	249.141	0.263	
	ThPLD	$\hat{\theta}=0.014,\hat{\beta}=15.372,\hat{\alpha}=-2.652*10^{-15}$	242.873	248.873	252.300	0.263	
	LD	$\hat{\theta}=0.027$	231.471	233.471	234.607	0.149	
	SD	$\hat{\theta}=0.028$	231.061	233.061	234.196	0.145	

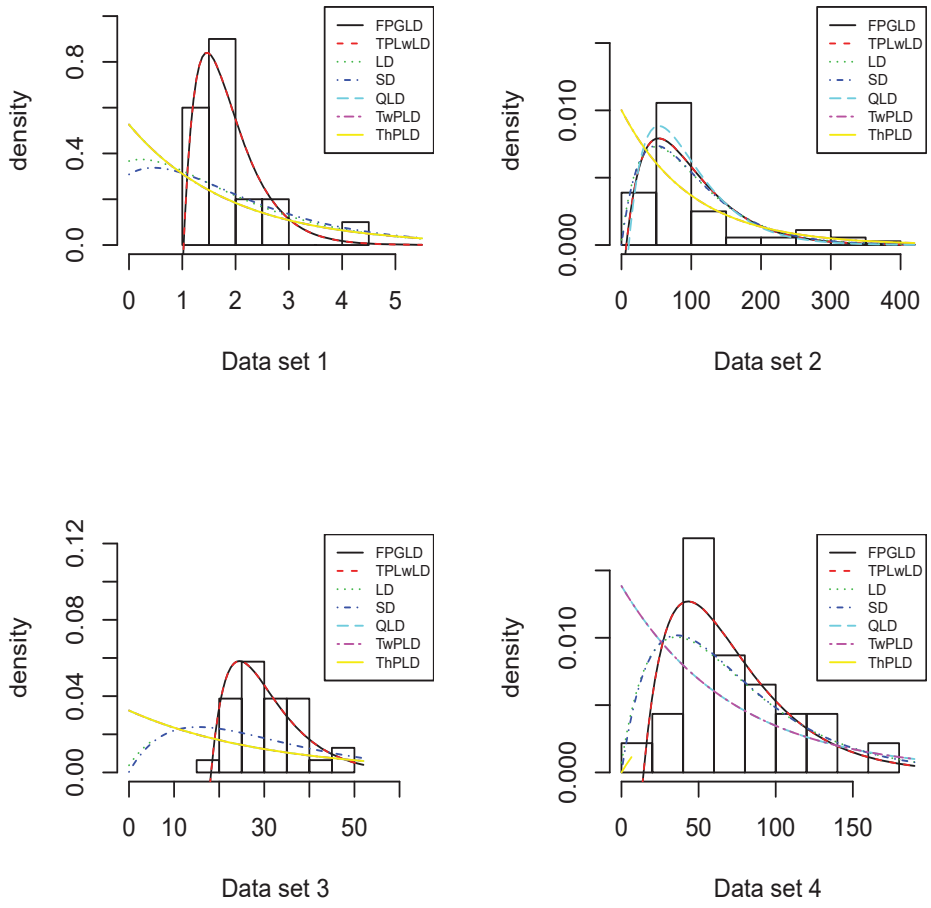


Figure 3: Empirical histograms with fitted densities of distributions

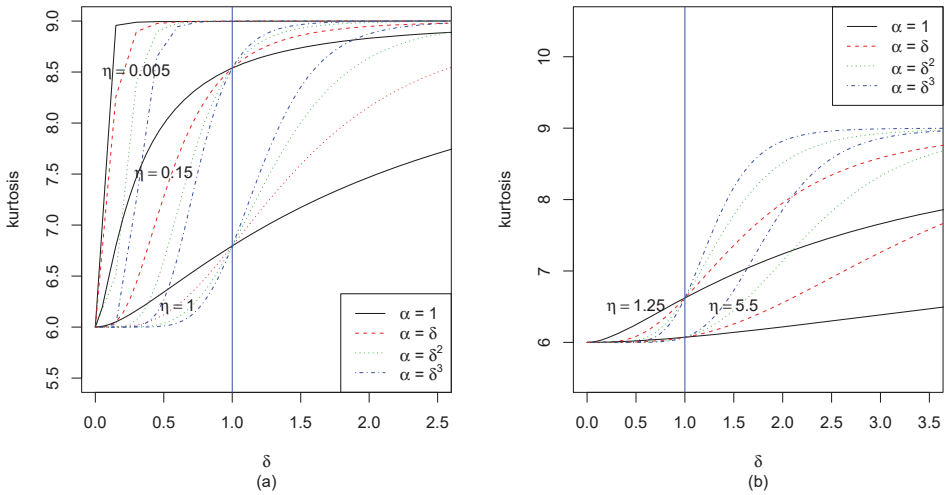


Figure 4: The kurtosis values of FPGLD at different parameter values of δ, α and η

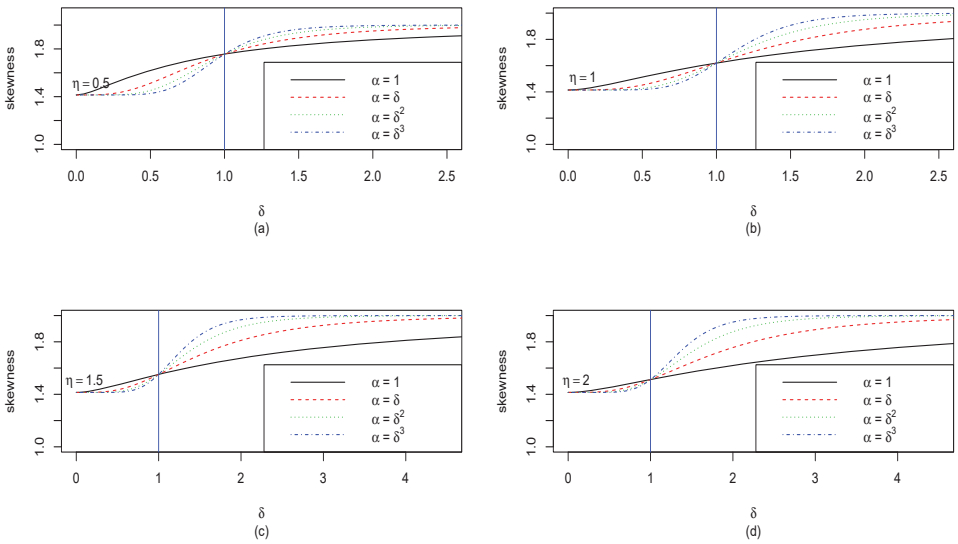


Figure 5: The skewness values of FPGLD at different parameter values of δ, α and η

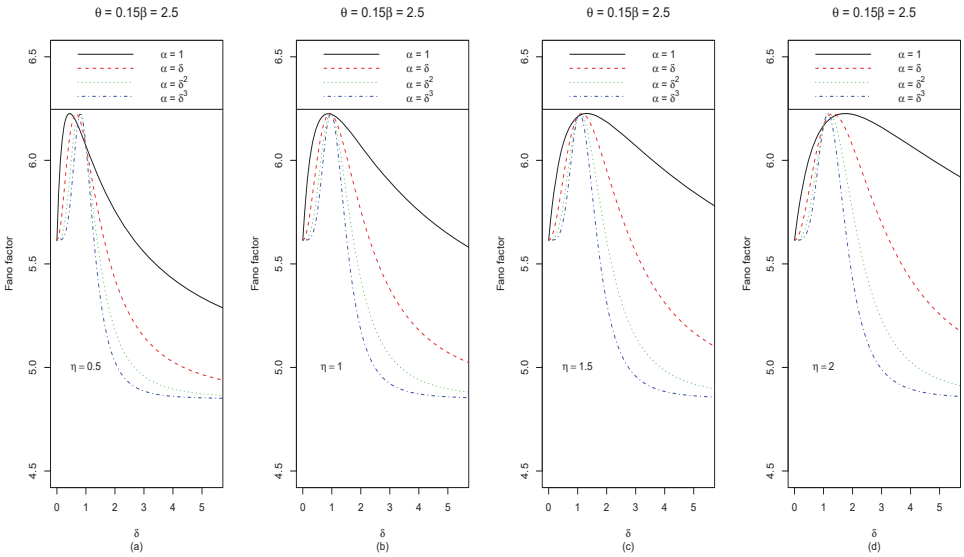


Figure 6: The Fano factor values of FPGLD at different parameter values of δ, α and η

(a) to (d): θ, β and η are fixed, and δ and α values are changed

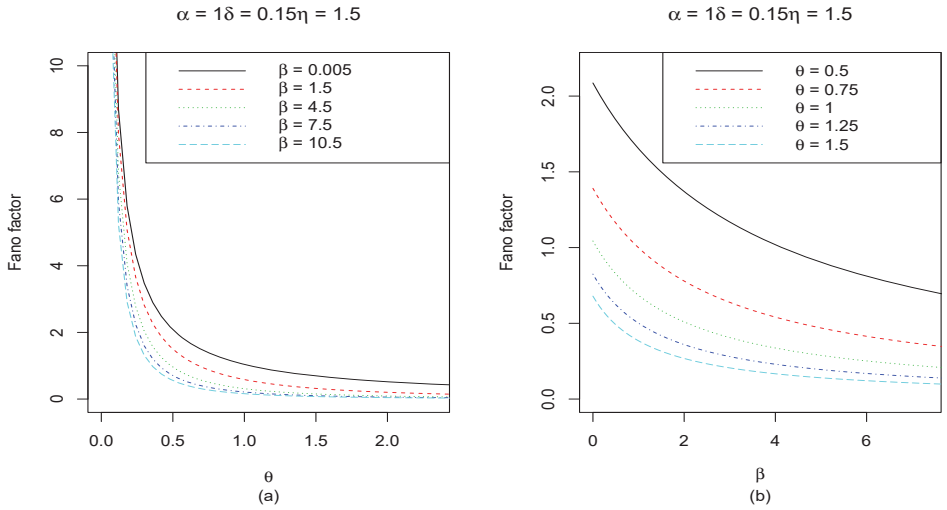


Figure 7: The Fano factor values of FPGLD at different parameter values of β and θ

(a) and (b): δ, α and η are fixed, and θ and β values are changed

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 119–141, DOI 10.21307/stattrans-2020-016
Submitted – 03.05.2017; paper accepted for published – 13.02.2020

Statistical properties and different methods of estimation for extended weighted inverted Rayleigh distribution

Abhimanyu Singh Yadav[1], **S. K. Singh**[2], **Umesh Singh**[3]

ABSTRACT

The aim of this paper is to introduce a new weighted probability distribution to model the non-monotone failure rate pattern for survival data. The proposed distribution is generalized by considering inverted Rayleigh distribution as a baseline distribution called an extended weighted inverted Rayleigh distribution. Different statistical properties such as moment, quantile function, moment generating function, entropy measurement, Bonferroni and Lorenz curve, stochastic ordering and order statistics have been derived. Different estimation procedures have also been discussed to estimate the unknown parameters of the proposed probability distribution. The Monte Carlo simulation study has been conducted to compare the performances of the proposed estimators obtained through various methods of estimation. Finally, two real data sets have been used to show the applicability of the proposed model in a real-life scenario.

Key words: moments and inverse moments, entropy measurements, order statistics, classical methods of estimation.

1. Introduction

In reliability analysis, numerous methods are available to generalize new probability distribution by adding an extra parameter with specific baseline distributions. For example, the well-known lifetime distributions, namely Weibull and gamma, are generalized by using power and Laplace transform of exponential random variates. Also, Gupta and Kundu (1999) introduced exponentiated exponential distribution by adding a shape parameter as a power of cumulative distribution function (CDF) of an exponential distribution. Nadarajah et al. (2011) proposed an extension of exponential distribution by a simple modification in the survival function of the exponential model. In reliability analysis, Rayleigh distribution (2005) is one of the most popular lifetime distribution and several generalizations based on Rayleigh distribution are advocated from time to time using the similar approach. The inverted versions of these models are also frequently used and well justified for the real life situations. For example, Voda (1972) introduced the inverted version of the Rayleigh model and discussed its different statistical properties. Ahmad et al. (2014) derived the weighted version of Rayleigh distribution and debated about the descriptive measure of statistics and

[1]Department of Statistics, Banaras Hindu University, Varanasi-221005, India. E-mail: asybhu10@gmail.com. ORCID: https://orcid.org/0000-0002-2411-5190.
[2]Department of Statistics, Banaras Hindu University, Varanasi-221005, India.
[3]Department of Statistics, Banaras Hindu University, Varanasi-221005, India.

estimation procedure for the unknown parameters. Exponentiated inverse Rayleigh distribution (EIRD) is a generalized form of inverse Rayleigh distribution as suggested by Nadarajah and Kotz (2006), etc., although the weighted version of the inverted Rayleigh distribution (IRD) has already been developed by Fatima and Ahmad (2017) using weighting function (length-area biased) approach. In this paper, a new weighted version of IRD has been proposed and studied with IRD as a base line distribution. IRD is the most popular lifetime distribution and frequently used to model the data with non-monotone failure rate. Let us assume that the variable Y is distributed as IRD with parameter θ. The probability density and distribution functions of IRD are given by;

$$f_Y(x) = \frac{2\theta}{x^3} e^{-\frac{\theta}{x^2}}, \quad \theta > 0, x > 0 \tag{1}$$

and

$$F_Y(x) = e^{-\frac{\theta}{x^2}} \tag{2}$$

where, θ is scale parameter.
The proposed extended weighted version of IRD has been derived by using the approach discussed by Azzalini (1985). The method mentioned by Azzalini is not new. It was given in 1985 and introduced various skew-symmetric distributions namely skew-normal, skew-chai, skew-Cauchy, skew-t, etc. Recently, Gupta and Kundu (2009) derived a new class of weighted distribution by introducing shape parameters to exponential distributions using the same approach. The lifetime distribution generated by this method possesses several good properties and can be used as a good alternative to other popular distributions such as gamma, Weibull, Rayleigh or generalized exponential distribution, etc.

The organization of the paper is as follows.The introduction of the considered problem is given in Section 1. Section 2 discusses the model genesis and its reliability characteristics. Statistical properties have been discussed in Sections 3, 4 & 5 respectively. Estimation of the unknown parameters is proposed in Section 6. Monte Carlo simulation study has been performed in Section 7. Section 8 has described the applicability of the model and study using two real data sets. Finally, Section 9 concludes the paper.

2. The model

Let X_1 and X_2 be the two i.i.d. random variables, with probability density function (PDF) $f_Y(\cdot)$ and CDF $F_Y(\cdot)$, then for any $\alpha > 0$, consider a new random variable $X = X_1$ given that $\alpha X_1 > X_2$. Then, the PDF of the new random variable X is;

$$f_X(x, \alpha) = \frac{1}{P[\alpha X_1 > X_2]} f_Y(x) F_Y(\alpha x) \quad ; \alpha > 0 \ \& \ x > 0 \tag{3}$$

Now, by using the Equations (1) and (2) in (3), the resulting probability distribution is called as extended weighted inverted Rayleigh distribution (EWIRD). Hence, the PDF and CDF

of the new model are given by;

$$f_w(x,\alpha,\theta)=\left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{x^3}\right)e^{-\frac{\theta}{x^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} \quad ;x>0\ \&\ \alpha,\theta>0 \tag{4}$$

and

$$F_w(x,\alpha,\theta)=e^{-\frac{\theta}{x^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} \tag{5}$$

where, α is the shape parameter of the model.

2.1. Reliability characteristics

- The reliability function $R(t)$ for specified value of t is given by

$$R_w(t)=1-F_w(t)=1-e^{-\frac{\theta}{t^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} \quad ;t>0 \tag{6}$$

- The hazard rate, i.e. instantaneous failure rate $h(t)$ is the conditional probability of failure in time interval $(t,t+\delta t)$ given that units has survived at least time t. Mathematically, it is given by the following equation:

$$h(t)=\frac{f_w(t)}{R_w(t)}=\frac{\left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{t^3}\right)e^{-\frac{\theta}{t^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}}{1-e^{-\frac{\theta}{t^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}} \tag{7}$$

- The reverse hazard function $H(t)$ can be interpreted as the ratio of the probability density function to the distribution function and is defined as;

$$H(t)=\frac{f_w(t)}{F_w(t)}=\left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{t^3}\right) \tag{8}$$

The different shape of the distribution, i.e. curve of density function and reliability function are presented in Figure 1. The shape of the hazard is presented in Figure 2. From Figure 2, it is clear that the proposed model is unimodal and exhibits the pattern of non-monotone failure rate. Usually, the problem of non-monotone failure rate is arising in medical and engineering sciences. In survival analysis, several times it has been realized that the failure rate of survival data is reached to a pick in the beginning stage and then declined abruptly until it stabilized. Such behaviour of the hazard rate is called a non-monotone failure rate, and the same behaviour of the failure rate is accommodated by the proposed model which, would be more flexible and used as a alternative survival model.

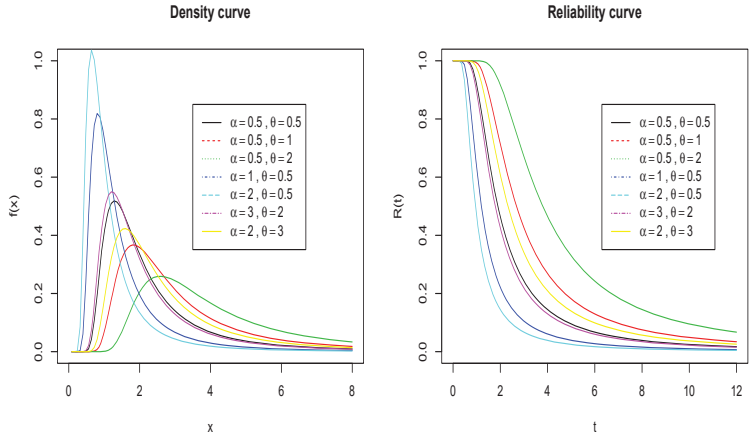


Figure 1: Density and reliability curve.

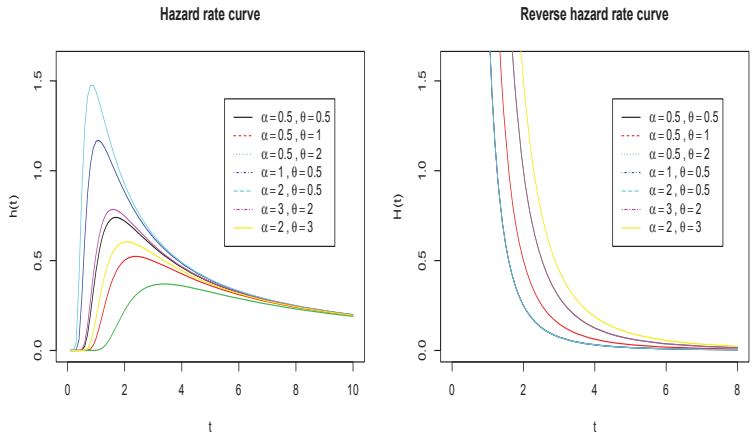


Figure 2: Hazard rate and reverse hazard rate.

3. Statistical properties

The different statistical properties of EWIRD are discussed in the following subsections.

3.1. Moments

The r^{th} moment about the origin is defined as;

$$\begin{aligned}\mu_r^{'} = E(X^r) &= \int_{x=0}^{\infty} x^r f_w(x,\alpha,\theta)dx \\ &= \int_{x=0}^{\infty} \left(\frac{1+\alpha^2}{\alpha^2}\right) 2\theta x^{r-3} e^{-\frac{\theta}{x^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} dx \\ &= \Gamma\left(1-\frac{r}{2}\right)\left[\frac{\theta(1+\alpha^2)}{\alpha^2}\right]^{\frac{r}{2}}\end{aligned} \tag{9}$$

The above expression is valid only for $r \leq 1$. Therefore, only mean of the distribution will exist in closed form and obtained by setting $r = 1$ in the Equation (9).

$$Mean = E(X) = \sqrt{\frac{\pi\theta(1+\alpha^2)}{\alpha^2}} \tag{10}$$

3.2. Inverse moments

The r^{th} inverse moment about origin $(M_{r^{-1}}^{'})$ is evaluated by the following expression:

$$\begin{aligned}M_{r^{-1}}^{'} = E\left(\frac{1}{X}\right)^r &= \int_{x=0}^{\infty} x^{-r} f_w(x,\alpha,\theta)dx \\ &= \int_{x=0}^{\infty} \left(\frac{1+\alpha^2}{\alpha^2}\right) 2\theta x^{-r-3} e^{-\frac{\theta}{x^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} dx \\ &= \Gamma\left(1+\frac{r}{2}\right)\left[\frac{\theta(1+\alpha^2)}{\alpha^2}\right]^{\frac{-r}{2}}\end{aligned} \tag{11}$$

- The values of different inverse moments are obtained by putting $r = 1,\cdots,4$ in the above equation and we get,

$$E\left(\frac{1}{X}\right) = \frac{1}{2}\sqrt{\frac{\pi\alpha^2}{\theta(1+\alpha^2)}}$$

$$E\left(\frac{1}{X^2}\right) = \frac{\alpha^2}{\theta(1+\alpha^2)}$$

$$E\left(\frac{1}{X^3}\right) = \frac{3}{4}\sqrt{\frac{\pi\alpha^6}{\theta^3(1+\alpha^2)^3}}$$

and

$$E\left(\frac{1}{X^4}\right) = \frac{2\alpha^4}{\theta^2(1+\alpha^2)^2}$$

respectively.

- Variance of the inverse variable is

$$Var\left(\frac{1}{X}\right) = \frac{\alpha^2(4-\pi)}{4\theta(1+\alpha^2)}$$

- Coefficient of variation (CV) is evaluated as

$$CV = \frac{\sqrt{Var\left(\frac{1}{X}\right)}}{E\left(\frac{1}{X}\right)} = \sqrt{\frac{4-\pi}{\pi}} \approx 0.5223$$

3.3. Quantile function

The quantile function for the proposed model is computed by the following expression:

$$F(Q_i) = \frac{i}{\zeta} \tag{12}$$

where, i and ζ indicates the position and total partition respectively. After simplification, the quartile function is

$$Q_i = \sqrt{\frac{\theta}{[\ln\zeta - \ln i]}\left(\frac{1+\alpha^2}{\alpha^2}\right)} \tag{13}$$

- **Quartile:** The first, second and third quartiles are evaluated by taking $\zeta = 4$ & $i = 1,2,3$ respectively.

- **Decile:** The deciles are calculated by assuming $\zeta = 10$, then the consecutive deciles are obtained by taking $i = 1,2,\cdots,9$.

- **Percentile:** The percentiles are evaluated by assuming $\zeta = 100$, then the consecutive deciles are obtained by taking $i = 1,2,\cdots,99$.

3.4. Median and Mode

The median for the PDF (4) is evaluated by using the following expression:

$$P[X \leq M] = P[X \geq M] = \frac{1}{2} \tag{14}$$

where, M is the median value. Also, it can be directly extracted from quantile expressions, e.g. 2^{nd} (in quartile), 5^{th} (in deciles) and 50^{th} (in percentile) quantiles are median. Thus, the

median is

$$M = \sqrt{\frac{\theta(1+\alpha^2)}{(\ln 2)\alpha^2}} \tag{15}$$

If the proposed probability distribution is moderately skewed then it has been verified that the difference between mean and mode is almost equal to three times the difference between the mean and median. Hence, we have

$$\text{Mode=3 Median-2 Mean}$$

which yield

$$Mode = 0.06\sqrt{\frac{\theta(1+\alpha^2)}{\alpha^2}} \tag{16}$$

3.5. Sample generation

The random sample for EWIRD can be generated using an inverse CDF transformation method as follows.

- Generate random deviates (U) from uniform distribution.
- Equate $F_w(X) = U \Rightarrow X = F^{-1}(U)$
- After simplification, we get

$$X = \sqrt{\frac{\theta(1+\alpha^2)}{\ln(U^{-1})\alpha^2}} \tag{17}$$

3.6. Moment generating function

The moment generating function (mgf) of a continuous r.v. x is defined by the following equation:

$$M_X(t) = E(e^{tX}) = \int_{x\in R^+} e^{tx} f_w(x,\alpha,\theta)dx \tag{18}$$

Thus, using PDF (4) we get;

$$\begin{aligned} M_X(t) &= \int_{x=0}^{\infty} e^{tx} \left[\left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{x^3}\right) e^{-\frac{\theta}{x^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} \right] dx \\ &= \sum_{r=0}^{\infty} \frac{t^r}{r!}\Gamma\left(1-\frac{r}{2}\right)\sqrt[r]{\left(\frac{\theta+\theta\alpha^2}{\alpha^2}\right)} \end{aligned} \tag{19}$$

Also, the moment can be obtained by using the above expression of m.g.f. Also, the characteristics function is simply obtained by replacing t by it in the above equation.

4. Entropy measurements

In information theory, entropy measurement plays a vital role in studying the uncertainty associated with the probability distribution. In this section, we discuss a different measure of change. For more detail about entropy, measurement see, Reniyi (1961).

4.1. Renyi entropy

Renyi entropy (RE) of a r.v. X is defined as

$$\begin{aligned} RE &= \frac{1}{(1-\kappa)} \ln \left[\int_{x=0}^{\infty} f_w^{\kappa}(x,\alpha,\theta) dx \right] \\ &= \frac{1}{(1-\kappa)} \ln \left[\int_{x=0}^{\infty} \left\{ \left(\frac{1+\alpha^2}{\alpha^2} \right) \left(\frac{2\theta}{x^3} \right) e^{-\frac{\theta}{x^2}\left(\frac{1+\alpha^2}{\alpha^2} \right)} \right\}^{\kappa} dx \right] \end{aligned} \tag{20}$$

Hence, after solving the internal, we get the following

$$RE = \frac{1}{2} \ln(\theta + \theta\alpha^2) - \ln\alpha - \ln 2 - \frac{(3\kappa-1)}{2(\kappa-1)} \ln\kappa + \frac{1}{(1-\kappa)} \ln\Gamma\left(\frac{(3\kappa-1)}{2} \right) \tag{21}$$

4.2. β-Entropy

The β-entropy (BE) is obtained as follows:

$$BE = \frac{1}{\beta-1} \left[1 - \int_{x=0}^{\infty} f_w^{\beta}(x,\alpha,\theta) dx \right] \tag{22}$$

Using PDF (4) and after simplification the expression for β-entropy is given by

$$BE = \frac{1}{\beta-1} \left[1 - \phi(\alpha,\theta,\beta) \right] \tag{23}$$

where, $\phi(\alpha,\theta,\beta) = \left(\frac{\theta+\theta\alpha^2}{\alpha^2} \right)^{\frac{1-\beta}{2}} \frac{2^{\beta-1}}{\beta^{(3\beta-1)/2}} \Gamma\left(\frac{(3\beta-1)}{2} \right)$

4.3. Generalized entropy

The generalized entropy (GE) is obtained by

$$GE = \frac{\nu_{\lambda}\mu^{-\lambda} - 1}{\lambda(\lambda-1)} \quad ; \lambda \neq 0, 1 \tag{24}$$

where, $\nu_\lambda = \int_{x=0}^{\infty} x^\lambda f_w(x,\alpha,\theta)dx$. The value of ν_λ is calculated as;

$$\begin{aligned}\nu_\lambda &= \int_{x=0}^{\infty} x^\lambda \left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{x^3}\right) e^{-\frac{\theta}{x^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} dx \\ &= \left(\frac{\theta(1+\alpha^2)}{\alpha^2}\right)^{\lambda/2} \Gamma\left(1-\frac{\lambda}{2}\right)\end{aligned} \tag{25}$$

Using the above expression

$$GE = \frac{\left(\frac{\theta(1+\alpha^2)}{\alpha^2}\right)^{\lambda/2} \Gamma\left(1-\frac{\lambda}{2}\right)\mu^{-\lambda}-1}{\lambda(\lambda-1)} \quad ;\lambda \neq 0,1 \tag{26}$$

4.4. Bonferroni and Lorenz curves

Bonferroni and Lorenz curves have good link-up to each other, and their extensive applications can be found in economics to study the income and poverty level. In the present time, these are frequently used in reliability theory also. These were initially proposed and analysed by Bonferroni (1930) and are defined by

$$B(c) = \frac{1}{cm}\int_0^a x f_w(x,\alpha,\theta)dx \tag{27}$$

$$L(c) = \frac{1}{m}\int_0^a x f_w(x,\alpha,\theta)dx \tag{28}$$

respectively. where $a = F_w^{-1}(c)$ and $m = E(x)$. Hence, using the Equation (4), the above two equations are reduced as

$$B(c) = \frac{1}{c\sqrt{\pi}} IG\left(\frac{1}{2}, \frac{\theta+\theta\alpha^2}{a^2\alpha^2}\right) \tag{29}$$

$$L(c) = \frac{1}{\sqrt{\pi}} IG\left(\frac{1}{2}, \frac{\theta+\theta\alpha^2}{a^2\alpha^2}\right) \tag{30}$$

where, $IG(a_1,b_1)$ stands for incomplete gamma function.

4.5. Stochastic ordering

The concept of stochastic ordering is used to show the ordering mechanism in lifetime distributions. For more detail about stochastic ordering see, Shaked and Shanthikumar (1988). The random variable X and Y is said to possess the following ordering behaviour:

- stochastic order $(X \leq_{st} Y)$ if $F_X(x) \geq F_Y(x)$ for all x.
- hazard rate order $(X \leq_{hr} Y)$ if $h_X(x) \geq h_Y(x)$ for all x.
- mean residual life order $(X \leq_{mrl} Y)$ if $m_X(x) \geq m_Y(x)$ for all x.

- likelihood ratio order $(X \leq_{lr} Y)$ if $\left(\frac{f_w^X(x)}{f_w^Y(x)}\right)$ decreases in x.

From the above relations, we analyzed that;

$$(X \leq_{lr} Y) \Rightarrow (X \leq_{hr} Y) \Downarrow (X \leq_{st} Y) \Rightarrow (X \leq_{mrl} Y)$$

The proposed distribution is also ordered with respect to the strongest likelihood ratio ordering as shown in the following lemma.

Lemma: Let $X \sim f_w(\alpha_1, \theta_1)$ and $Y \sim f_w(\alpha_2, \theta_2)$. If $\alpha_1 > \alpha_2$, then $(X \leq_{lr} Y)$ and hence $(X \leq_{hr} Y)$, $(X \leq_{mrl} Y)$ and $(X \leq_{st} Y)$.

Proof: According to the definition of likelihood ratio order, first we obtain the ratio $\left[\frac{f_w^X(x)}{f_w^Y(x)}\right]$ i.e.

$$\begin{aligned}\psi = \frac{f_w^X(x)}{f_w^Y(x)} &= \frac{\left(\frac{1+\alpha_1^2}{\alpha_1^2}\right)\left(\frac{2\theta_1}{x^3}\right)e^{-\frac{\theta_1}{x^2}\left(\frac{1+\alpha_1^2}{\alpha_1^2}\right)}}{\left(\frac{1+\alpha_2^2}{\alpha_2^2}\right)\left(\frac{2\theta_2}{x^3}\right)e^{-\frac{\theta_2}{x^2}\left(\frac{1+\alpha_2^2}{\alpha_2^2}\right)}} \\ &= \frac{\theta_1\alpha_2^2(1+\alpha_1^2)}{\theta_2\alpha_1^2(1+\alpha_2^2)}e^{-\frac{1}{x^2}\left[\frac{\theta_1(1+\alpha_1^2)}{\alpha_1^2}+\frac{\theta_2(1+\alpha_2^2)}{\alpha_2^2}\right]}\end{aligned}$$

Therefore,

$$\psi' = \frac{2\theta_1\alpha_2^2(1+\alpha_1^2)}{x^3\theta_2\alpha_1^2(1+\alpha_2^2)}\left[\frac{\theta_1(1+\alpha_1^2)}{\alpha_1^2}+\frac{\theta_2(1+\alpha_2^2)}{\alpha_2^2}\right]e^{-\frac{1}{x^2}\left[\frac{\theta_1(1+\alpha_1^2)}{\alpha_1^2}+\frac{\theta_2(1+\alpha_2^2)}{\alpha_2^2}\right]} \quad (31)$$

from above equation, we observed that if $\psi' > 0 \; \forall \; \alpha_1, \alpha_2$, hence $(X \leq_{lr} Y)$. The remaining statements can be established in the same way.

5. Order statistics

Let us consider $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ are the n ordered random sample from (4) then $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denote the corresponding order statistics. $X_{(1)}$, $X_{(n)}$ and $X_{(r)}$ denote the minimum, maximum and r^{th} order statistics respectively. Then the PDF of r^{th} order statistics, as suggested by (1970), is given by

$$f_r(X_{(r)}, \alpha, \theta) = \frac{n!}{(r)!(n-r)!}\left[F_w(x_{(r)})\right]^{r-1}\left[1-F_w(x_{(r)})\right]^{n-r} f_w(x_{(r)}, \alpha, \theta) \quad (32)$$

Using the expressions (4) and (5) for $X_{(r)}$ in the above equation we get the expression for r^{th} order statistics, i.e.

$$f_r(X_{(r)},\alpha,\theta) = \frac{n!}{(r)!(n-r)!}\left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{x_{(r)}^3}\right)\left[e^{-\frac{\theta}{x_{(r)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^r \times \left[1-e^{-\frac{\theta}{x_{(r)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^{n-r} \tag{33}$$

5.1. Distribution of minimum, maximum and median

Let $X_{(1)},X_{(2)},\cdots,X_{(n)}$ be the n independent ordered random sample observed, then the distribution of minimum $X_{(1)}$, maximum $X_{(n)}$ order statistics are obtained by putting $r=1$ & $r=n$ in the Equation (35). Hence, after simplification we get

$$\begin{aligned} f_{mini}(X_{(1)},\alpha,\theta) &= n\left[1-F_w(x_{(1)})\right]^{n-1} f_w(x_{(1)},\alpha,\theta) \\ &= n\left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{x_{(1)}^3}\right)\left[e^{-\frac{\theta}{x_{(1)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]\left[1-e^{-\frac{\theta}{x_{(1)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^{n-1} \end{aligned} \tag{34}$$

$$\begin{aligned} f_{max}(X_{(n)},\alpha,\theta) &= n\left[F_w(x_{(n)})\right]^{n-1} f_w(x_{(n)},\alpha,\theta) \\ &= n\left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{x_{(n)}^3}\right)\left[e^{-\frac{\theta}{x_{(n)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^n \end{aligned} \tag{35}$$

Now, the density function for sample median for order statistics is given by $X_{(m+1:n)}$. It is computed by

$$\begin{aligned} f_{m+1}(\tilde{X}_{(m+1:n)},\alpha,\theta) &= \frac{(2m+1)!}{(m)!(m)!}\left[F_w(\tilde{x}_{m+1})\right]^m\left[1-F_w(\tilde{x}_{m+1})\right]^m f_w(\tilde{x}_{m+1},\alpha,\theta) \\ &= \frac{(2m+1)!}{(m)!(m)!}\left(\frac{1+\alpha^2}{\alpha^2}\right)\left(\frac{2\theta}{\tilde{x}_{m+1}^3}\right)\left[e^{-\frac{\theta}{\tilde{x}_{m+1}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^{m+1} \\ &\times\left[1-e^{-\frac{\theta}{\tilde{x}_{m+1}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^m \end{aligned} \tag{36}$$

5.2. Joint distribution of r^{th} and s^{th} order statistics

The joint density function of r^{th} and s^{th} $(r<s)$ order statistics is obtained by considering the following expression:

$$\begin{aligned} &f_{r:s:n}(X_r,X_s,\alpha,\theta) = \\ &= xi\left[F_w(x_{(r)})\right]^{r-1}\left[1-F_w(x_{(s)})\right]^{n-s}\left[F_w(x_{(s)})-F_w(x_{(r)})\right]^{s-r-1} f_w(x_{(r)})f_w(x_{(s)}) \end{aligned} \tag{37}$$

Using PDF and CDF of EWIRD, the density function for r^{th},s^{th} is given by

$$\begin{aligned} f_{r:s:n}(X_r,X_s,\alpha,\theta) &= \xi\left[e^{-\frac{\theta}{x_{(r)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^r\left[1-e^{-\frac{\theta}{x_{(s)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^{n-s} \\ &\times\left[e^{-\frac{\theta}{x_{(s)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}-e^{-\frac{\theta}{x_{(r)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^{s-r-1} \\ &\times\left(\frac{1+\alpha^2}{\alpha^2}\right)^2\left(\frac{4\theta^2}{x_{(r)}^3x_{(s)}^3}\right)e^{-\frac{\theta}{x_{(s)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} \end{aligned}$$

In particular, when $r=1$ and $s=n$, we have the joint distribution of minimum and maximum order statistics and it is written as

$$f_{1:n:n}(X_1,X_n,\alpha,\theta)=\frac{n!}{(n-2)!}\left(\frac{1+\alpha^2}{\alpha^2}\right)^2\left(\frac{4\theta^2}{x_{(1)}^3x_{(n)}^3}\right)e^{-\frac{\theta}{x_{(n)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\left[e^{-\frac{\theta}{x_{(1)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]$$
$$\times\left[e^{-\frac{\theta}{x_{(n)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}-e^{-\frac{\theta}{x_{(1)}^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\right]^{n-2} \tag{38}$$

where, $\xi=\dfrac{n!}{(r-1)!(s-r-1)!(n-s)!}$

6. Estimation of the parameters

In this section, we discuss different estimation procedures for estimating the unknown model parameters of the proposed model. These methods are presented below;

6.1. Maximum Likelihood Estimation

Let $X_1,X_2,\cdots,X_n$ be the random sample of size n from density function (4). The likelihood function is written as;

$$\begin{aligned}L(\alpha,\theta)&=\prod_{i=1}^{n}f_w(x_i,\alpha,\theta)\\&=\left(\frac{1+\alpha^2}{\alpha^2}\right)^n\frac{(2\theta)^n}{\prod_{i=1}^n x_i^3}e^{-\sum_{i=1}^n\frac{\theta}{x_i^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)}\end{aligned} \tag{39}$$

Hence, log-likelihood by ignoring the constant is written as;

$$L_1=\ln L(\alpha,\theta)=n\ln(1+\alpha^2)-2n\ln\alpha+n\ln\theta-3\sum_{i=1}^{n}\ln x_i-\left(\frac{\theta+\theta\alpha^2}{\alpha^2}\right)\sum_{i=1}^{n}\frac{1}{x_i^2} \tag{40}$$

Thus, the MLEs are obtained by maximizing the above equation w.r.t. to the parameters and as a result we have two likelihood equations which yield the MLEs of the unknown parameters.

$$\frac{n}{\theta}-\frac{1+\alpha^2}{\alpha^2}\sum_{i=1}^{n}\frac{1}{x_i^2}=0 \tag{41}$$

and

$$\frac{n\alpha}{1+\alpha^2}-\frac{n}{\alpha}+\frac{\theta}{\alpha^3}\sum_{i=1}^{n}\frac{1}{x_i^2}=0 \tag{42}$$

6.1.1 Interval estimate based on MLEs

From above, it is clear that the exact distribution for MLEs is not easy to find. Thus, we considered the asymptotic distribution of MLE to construct $100(1-\alpha)\%$ approximate confidence interval. Thus, for this purpose we evaluate the Fisher information matrix and it is obtained as

$$I(\hat{\alpha},\hat{\theta})=\begin{pmatrix}-l_{\alpha\alpha} & -l_{\alpha\theta}\\ -l_{\theta\alpha} & -l_{\theta\theta}\end{pmatrix}_{\hat{\alpha},\hat{\theta}} \tag{43}$$

where, $l_{\alpha\alpha}=\dfrac{\partial^2 \ln L(\alpha,\theta)}{\partial\alpha^2}, l_{\alpha\theta}=\dfrac{\partial^2 \ln L(\alpha,\theta)}{\partial\alpha\partial\theta}, l_{\theta\alpha}=\dfrac{\partial^2 \ln L(\alpha,\theta)}{\partial\theta\partial\alpha}, l_{\theta\theta}=\dfrac{\partial^2 \ln L(\alpha,\theta)}{\partial\theta^2}$

The above matrix is inverted and its diagonal elements provide the asymptotic variance of the estimates for the parameter. Hence, approximate CI is given by

$$[\hat{\alpha}_L,\hat{\alpha}_U]\in[\hat{\alpha}\mp Z_{\gamma/2}\sqrt{\hat{\sigma}^2{}_{\alpha\alpha}}]$$

and

$$[\hat{\theta}_L,\hat{\theta}_U]\in[\hat{\theta}\mp Z_{\gamma/2}\sqrt{\hat{\sigma}^2{}_{\theta\theta}}]$$

6.2. Maximum Product Spacing method of estimation

In this subsection, we described a very effective and alternative method to MLEs named maximum product spacing method. It was initially introduced and extensively studied by Chen and Amin (1979). Coolen and Newby (1990) studied its invariance properties and concluded that it possesses the similar features as MLEs. Recently, the utility of this method has been nicely explained by Singh et al. (2014). Under this method of estimation techniques the likelihood function is defined on the basis of spacing of two consecutive CDFs and is given by

$$L^{'}(\alpha,\theta)=\sqrt[n+1]{\prod_{i=1}^{n+1}D_i} \tag{44}$$

such that $\sum_{i=1}^{n}D_i=1$. Taking log both side, we get;

$$\begin{aligned}\ln L^{'}(\alpha,\theta)&=\frac{1}{n+1}\sum_{i=1}^{n+1}\ln D_i\\ &=\frac{1}{n+1}\left[\ln D_1+\sum_{i=2}^{n}\ln D_i+\ln D_{n+1}\right]\end{aligned} \tag{45}$$

where, $D_1=F(x_1)$, $D_i=F(x_i)-F(x_{i-1})$ and $D_{n+1}=1-F(x_n)$

The MPS estimates of the parameter α, θ are obtained by maximizing the above equation w. r. t. the parameters.

6.2.1 Interval estimate based on MPS

Here, we consider the asymptotic confidence intervals based on maximum product spacing estimates. It was mentioned by Cheng and Amin (1979), Ghosh and Jammalamadaka (2001) that the MPS method also shows asymptotic properties like the Maximum likelihood estimator and is asymptotically equivalent to MLE. Keeping this in mind, we have to consider the Fisher information matrix and it is obtained as;

$$I^{'}(\hat{\alpha},\hat{\theta}) = \begin{pmatrix} -W_{\alpha\alpha} & -W_{\alpha\theta} \\ -W_{\theta\alpha} & -W_{\theta\theta} \end{pmatrix}_{\hat{\alpha}_{MP},\hat{\theta}_{MP}} \tag{46}$$

where, $W_{\alpha\alpha} = \dfrac{\partial^2 \ln L^{'}(\alpha,\theta)}{\partial\alpha^2}, W_{\alpha\theta} = \dfrac{\partial^2 \ln L^{'}(\alpha,\theta)}{\partial\alpha\partial\theta}, W_{\theta\alpha} = \dfrac{\partial^2 \ln L^{'}(\alpha,\theta)}{\partial\theta\partial\alpha}, W_{\theta\theta} = \dfrac{\partial^2 \ln L^{'}(\alpha,\theta)}{\partial\theta^2}$

Using similar approach as MLE the $100\ (1-\gamma)$ asymptotic confidence interval is given by

$$\hat{\alpha}_{MP} \mp Z_{\gamma/2}\sqrt{(\sigma^2_{\hat{\alpha}\alpha})_{MP}}$$

and

$$\hat{\theta}_{MP} \mp Z_{\gamma/2}\sqrt{(\sigma^2_{\hat{\theta}\theta})_{MP}}$$

6.3. Estimators based on percentile

Estimation of the parameters based on percentile is not the new one and frequently used when the distribution function is in closed form. It was proposed and extensively studied by Kao (1958, 1959). Recently, this method has gained some popularity in the statistical literature and used was by Gupta and Kundu (2001), Kundu, and Raqab (2005) based on different lifetime models. Most of the time estimators obtained by this method have a nice closed form. The percentile-based estimators are mainly obtained by minimizing the Euclidean distance between the sample percentile and population percentile points. Hence, using the expression of CDF, we get

$$x = \sqrt{\frac{\theta(1+\alpha^2)}{\ln[F^{-1}(x,\alpha,\theta)]\alpha^2}} \tag{47}$$

The percentile estimators of α, θ are obtained by minimizing

$$PE = \sum_{i=1}^{n}\left[x_i - \sqrt{\frac{\theta(1+\alpha^2)}{\ln[F^{-1}(x,\hat{\alpha},\hat{\theta})]\alpha^2}}\right]^2 \tag{48}$$

where, $F(x,\hat{\alpha},\hat{\theta})$ denotes the estimated value of CDF. It can be assumed as

$$E[F(x_i,\alpha,\theta)] = \frac{i}{n+1} = p_i$$

6.4. Ordinary and Weighted Least Squares Estimation

The theory of least squares estimation was proposed by Swain et al. (1988) to estimate the parameters of Beta distribution using the principal of least squares. The LSEs of the unknown parameters of EWIRD are evaluated by minimizing

$$LSE = \sum_{i=1}^{n}\left[F(x_i,\alpha,\theta) - \frac{i}{n+1}\right]^2 \tag{49}$$

using the Equation (5) in the above equation, we get

$$LSE = \sum_{i=1}^{n}\left[e^{-\frac{\theta}{x^2}\left(\frac{1+\alpha^2}{\alpha^2}\right)} - \left(\frac{i}{n+1}\right)\right]^2 \tag{50}$$

Hence, the least square estimators of the parameter α and θ are obtained by minimizing the above equation w.r.t α and θ respectively.

7. Simulation study

In this section, Monte Carlo simulation study has been performed to compare the performance of the obtained estimators in previous subsections. The comparisons of these estimators are made in terms of average mean square error (mse) based on 5000 replications. Since, all estimators are not in closed form, thus non-linear optimization iterative procedure has been used to obtain the estimates of the parameters. In result Tables, $(\hat{\theta}_{ml},\hat{\alpha}_{ml})$, $(\hat{\theta}_{mp},\hat{\alpha}_{mp})$, $(\hat{\theta}_{lse},\hat{\alpha}_{lse})$, $(\hat{\theta}_{pse},\hat{\alpha}_{pse})$ denotes the estimators obtained by the method of MLE, MPSE, LSE and PSE for scale and shape parameters respectively. The simulation study has been carried out for $n = 10,20,30,50,80,$ & 120 when $\theta = 2,\alpha = 3$. Average estimates of the parameters and corresponding mse are reported in Table 1. From Table 1, it has been noticed that the mse of all estimator is decreasing when the sample size is increasing, which guarantees the consistency of the estimators. Also, among the estimators obtained by different method of estimation the following patterns have been noticed in terms of their average mse.

$$mse(\hat{\theta}_{lse}) < mse(\hat{\theta}_{mp}) < mse(\hat{\theta}_{mle}) < mse(\hat{\theta}_{pse})$$

and

$$mse(\hat{\alpha}_{ml}) < mse(\hat{\alpha}_{mp}) < mse(\hat{\alpha}_{pse}) < mse(\hat{\alpha}_{lse})$$

respectively. Thus, for the scale parameter θ, the LSE method performs better as compared to the other methods of estimation, however in the case of the shape parameter α, MLE

Table 1: Average estimate and corresponding mse (in each second row) of the estimators when $\theta = 2,\ \alpha = 3$.

n	$\hat{\theta}_{ml}$	$\hat{\theta}_{mp}$	$\hat{\theta}_{lse}$	$\hat{\theta}_{pse}$	$\hat{\alpha}_{ml}$	$\hat{\alpha}_{mp}$	$\hat{\alpha}_{lse}$	$\hat{\alpha}_{pse}$
10	2.2359	1.8034	1.8182	2.0608	3.2627	2.2831	1.9998	2.4456
	0.6672	0.3576	0.3439	0.8425	0.3391	0.5349	1.0549	0.6935
20	2.1047	1.8103	1.8181	2.2620	3.0717	2.3244	2.0647	2.4320
	0.2843	0.1999	0.1963	0.8403	0.2866	0.4674	0.9078	0.5880
30	2.0877	1.8238	1.8185	2.0987	3.1902	2.3333	2.0858	2.2867
	0.1622	0.1327	0.1274	0.7143	0.2479	0.4513	0.8958	0.5624
50	2.0733	1.8299	1.8348	2.0993	3.3729	2.3333	2.0624	2.3249
	0.0923	0.0922	0.0907	0.6439	0.1439	0.4485	0.8860	0.5584
80	2.0856	1.8499	1.7936	2.1270	3.1526	2.3263	2.0719	2.5151
	0.0561	0.0596	0.0546	0.5961	0.0389	0.4456	0.8646	0.4825
120	2.0720	1.8495	1.8430	2.2823	3.1546	2.3245	2.0567	2.5827
	0.0442	0.0483	0.0421	0.5926	0.0379	0.4358	0.8593	0.4576

method provides a better result.

8. Application to lifetime data

In this section, survival/reliability data applications of the proposed model are provided. For this purpose, we have considered two data sets and checked the suitability of the proposed model.

Data Set 1: Cancer data

These data represent the survival times (in days) of 45 head and neck cancer patients treated with combined radiotherapy and chemotherapy. Firstly, the data set is reported by Efron (1988). To check the suitability of the considered data set for the proposed model different model selection tools are used such as AIC, BIC, and log-likelihood criterion. These statistical tools are defined as follows:

$$AIC = -2 * \ln L(x, \hat{\alpha}, \hat{\theta}) + 2 * k$$

$$BIC = -2 * \ln L(x, \hat{\alpha}, \hat{\theta}) + k * \ln(n)$$

where, k is the number of parameters involved in the probability distribution, and n is the sample size. Smaller values of AIC, BIC and the LogL test statistic are indicators of better fit of distributions. The proposed model is compared with the most commonly used non-monotone failure rate models namely inverted exponential distribution (IED), generalized inverted exponential distribution (GIED) and Inverse Weibull distribution (IWD). Among these models, it has been observed that the proposed model has the least AIC, BIC and negative LogL, see Table 4. Hence, the proposed model can be taken as an alternative to these models when data have the non-monotone failure rate.

Data Set 2: Ball bearing failure data

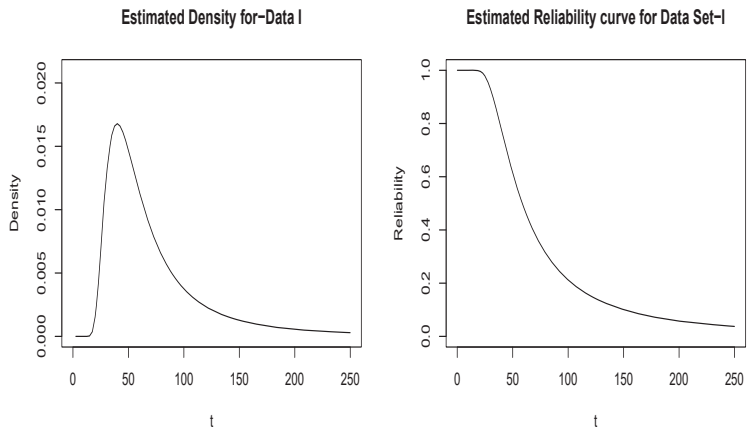


Figure 3: Estimated plot for the real data-I

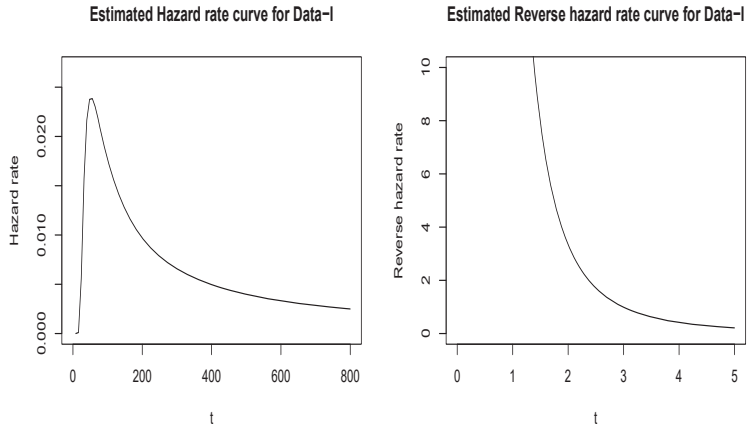


Figure 4: Estimated plot for the real data-I

Table 2: Estimates and values of various measures for data set-I

Models	Estimate	AIC	BIC	-LogL
IED	59.125	773.37	775.44	385.69
GIED	(0.777,49.241)	773.18	777.30	384.59
IWD	(28.505,0.786)	767.16	771.28	381.58
EWIRD	(6.679,0.053)	691.04	694.65	343.52

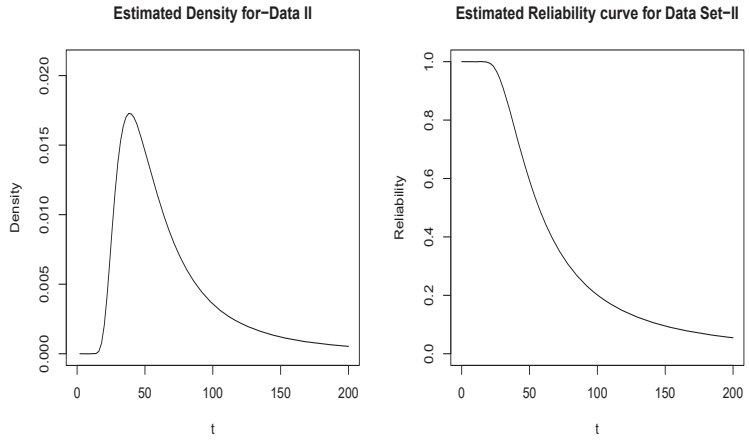


Figure 5: Estimated plot for the real data-II

The considered data set represents the 23 ball bearing failure times (millions of cycles) for units tested at one level of stress and it was firstly reported and analysed by Lawless (1982). The summary of the ball bearing data set is

Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
17.88	47.20	67.80	72.22	95.88	173.40

To check the validity of the proposed model, we used Kolmogrov-Smirnov test. Thus, the hypothesis is

H_0 : Samples are observed from proposed model

H_1 : Samples are not observed from proposed model

Hence, test statistic for testing the null hypothesis is

$KS_{Cal} = Sup_x|\hat{F}_n(x) - \hat{F}(x)| = 0.098$ and $KS_{tab} = 0.276$

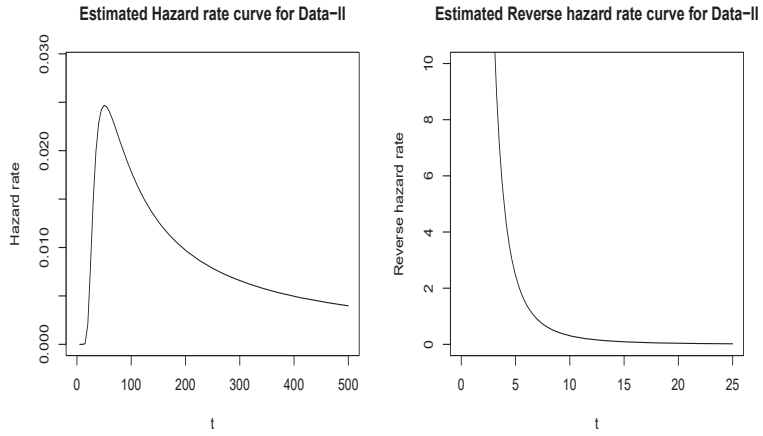


Figure 6: Estimated plot for the real data-II

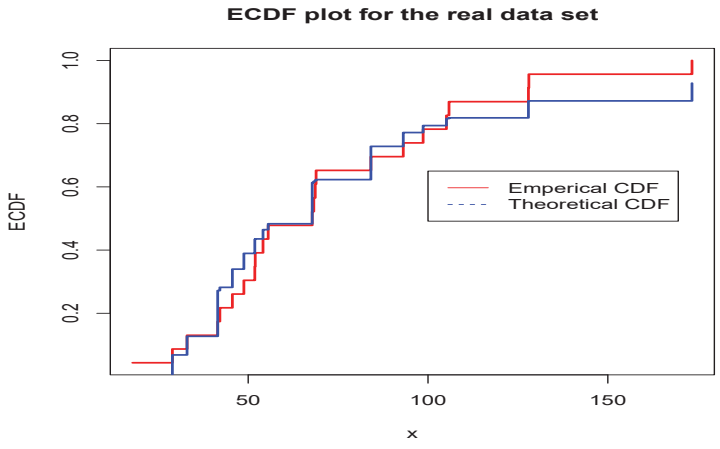


Figure 7: Empirical CDF plot for the real data

We see that the calculated value of is less than the tabulated value. Hence the null hypothesis may be accepted at $\alpha = 5\%$ level of significance. Also, from the empirical cumulative distribution function plot (see Figure 7) it is clear that the data set-II provides excellent fit to the proposed model and hence, one may use EWIRD as an alternative lifetime model. The estimated plots for the density function, reliability function, hazard function and reverse hazard rate function are given in Figure 3, 4, 5 and 6 respectively. These plots indicate that cancer data and ball bearing data both adequately accommodate the new model.

9. Conclusion

In this article, a new version of weighted probability distribution, named EWIRD has been introduced. Different statistical properties such as moments, inverse moments, moment generating function, entropy, stochastic ordering and order statistics have been discussed. Different estimation procedures are also described to estimate the unknown parameters, and their performances are compared through Monte Carlo simulations. The applications of the proposed model are provided based on two real data sets and it has been found that it can efficiently be used to model the data with a non-monotone failure rate pattern.

Acknowledgments

The authors are thankful to the anonymous reviewer for their valuable comments and suggestions regarding improvements of the current manuscript.

REFERENCES

AZZALINI, A., (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, 12, pp. 171–178.

AHMAD, A., AHMAD, S. P., AHMED, A., (2014). Characterization and Estimation of Double Weighted Rayleigh Distribution, *Journal of Agriculture and Life Sciences*, 1(2), pp. 2375–4222.

BONFERRONI, C. E., (1930). Elementi di Statistica General, *Seeber, Firenze.*

CHENG, R. C. H., AMIN, N. A. K., (1979). Maximum product of spacings estimation with applications to the lognormal distribution, *University of Wales IST, Math Report*, pp. 79–1.

COOLEN, F. P. A., NEWBY, M. J., (1990). A note on the use of the product of spacings in Bayesian inference, *Memorandum COSOR*; Vol. 9035), Eindhoven: Technische Universiteit Eindhoven.

DAVID, H. A., (1970). Order Statistics, *New York, Wiley.*

EFRON, B., (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve, *Journal of American Statistical Association*, 83, pp. 414–425.

FATIMA, K., AHMAD, S. P., (2017). Weighted Inverse Rayleigh Distribution, *International Journal of Statistics and Systems*, 12(1), pp. 119–137.

GHOSH, K., JAMMALAMADAKA, S., R., (2001). A general estimation method using spacings, *Journal of Statistical Planning and Inference*, 93.

GUPTA, R. D., KUNDU, D., (1999). Generalized exponential distributions, *Australia and New Zealand Journal of Statistics*, 41, pp. 173–188.

GUPTA, R. D., KUNDU, D., (2001). Generalized exponential distributions: different methods of estimation, *Journal of Statistical Computation and Simulation*, 69, pp. 315–338.

GUPTA, R. D., KUNDU, D. (2009,). A new class of weighted exponential distributions, *Statistics*, 43, pp. 621–634.

KAO, J. H. K., (1958). Computer methods for estimating Weibull parameters in reliability studies, *IRE Transactions on Reliability and Quality Control*, 13, pp. 15–22.

KAO, J. H. K., (1959). A graphical estimation of mixed Weibull parameters in life testing electron tube, *Technometrics*, 1, pp. 389–407.

KUNDU, D., RAQAB, M. Z., (2005). Generalized Rayleigh distribution: different methods of estimation, *Computational Statistics and Data Analysis*, 49, pp. 187–200.

LAWLESS, J. F., (1982). Statistical Models and Methods for Lifetime Data, *New York: John Wiley & Sons*.

MAKKAR, P., SRIVASTAVA, P. K. , SINGH R. S., UPADHYAY, S. K., (2014). Bayesian survival analysis of head and neck cancer data using log normal model, *Communications in Statistics - Theory and Methods*, 43, pp. 392–407.

NADARAJAH, S., HAGHIGHI, F., (2011). An extension of the exponential distribution, *Statistics: Journal of Theoretical and Applied Statistics*, 45(6), pp. 543–558.

NADARAJAH, S., KOTZ, S., (2006). The exponentiated type distributions, *Acta Applicandae Mathematicae*, 92(2), pp. 97–111.

RAYLEIGH, J. W. S., (1880). On the resultant of a large number of vibrations of the some pitch and of arbitrary phase. *Philosophical Magazine*, 5-th Series, 10, pp. 73–78.

RANNEBY, B., (1984). The Maximum Spacings Method. An Estimation Method Related to the Maximum Likelihood Method, *Scandinavian Journal of Statistics*, 11, pp. 93–112.

RENYI, A., (1961). On measures of entropy and information, in: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, *University of California Press, Berkeley*.

SINGH, U., SINGH, S. K., SINGH, R. K., (2014). Comparative study of traditional estimation method and maximum product spacing method in Generalized inverted exponential Distribution, *Journal of Statistics Application and Probability*, 3(2), pp. 1–17.

SWAIN, J. J., VENKATARAMAN, S., WILSON, J. R., (1988). Least squares estimation of distribution functions in Johnson's translation system, *Journal of Statistical Computation and Simulation*, 29, pp.271–297.

SHAKED, M., SHANTHIKUMAR, J. G., (1994). Stochastic Orders and Their Applications, *New York, Academic Press.*

VODA, V. G., (1972). On the inverse Rayleigh random variable, *Rep. Statistical Application Res. Jues*, 19 (4), pp. 13–21.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 143–156, DOI 10.21307/stattrans-2020-017
Submitted – 02.01.2020; accepted for publishing – 29.02.2020

Asymmetry of foreign trade turnover between Ukraine and Poland

Ruslan Motoryn[1], **Kateryna Prykhodko**[2], **Bogusław Ślusarczyk**[3]

ABSTRACT

The article identifies the determinants of the asymmetry of foreign trade turnover between Ukraine and Poland based on an analysis of competitiveness indicators of the studied countries in the period 2003–2017. The emphasis is on calculation of the comparative advantages of particular commodity headings in Polish exports in the domestic market of Ukraine. Potential directions of the intensification of bilateral trade were evaluated.

Key words: asymmetry, competitiveness, foreign trade, international cooperation.

1. Introduction

Asymmetry of trade integration is caused by differences in the levels of economic development of countries, the size of the market, the degree of integration of countries into the global economy and other factors. The current structure of exports and imports demonstrates a critical technological imbalance for Ukraine: raw materials are exported from Ukraine and high technology is imported from EU countries; trade deficit maintains for most product groups; domestic businesses show limited EU market entry due to high level of non-tariff protection, primarily on agricultural products. In this regard, the experience of Poland as an EU member demonstrates the possibility of overcoming technological imbalance.

The main purpose of the empirical analysis undertaken in the article is to study the development trends of foreign trade of Poland with Ukraine in 2003–2017. This partnership is explained by these facts:

– Ukraine borders Poland. Upon its accession to the European Union, Ukrainian eastern borders will also become the borders of the European Union, which will

[1] Kyiv National University of Trade and Economics, Ukraine. E-mail: ruslan.motoryn@gmail.com.
ORCID: https://orcid.org/0000-0001-9344-2315.
[2] Taras Shevchenko National University of Kyiv, Ukraine. E-mail: kateryna.prykhodko@gmail.com.
ORCID: https://orcid.org/0000-0001-5415-1662.
[3] University of Rzeszow, Poland. E-mail: boguslaw.slusarczyk@gmail.com.
ORCID: https://orcid.org/0000-0003-0567-8470.

undoubtedly influence the building of relations and forms of cooperation between these partners in the future;
- these are countries with unequal levels of participation in international division of labour and levels of economic development, which, in our view, will allow us to broadly verify theories of international trade and, moreover, to answer the question whether they form the basis of trade policy and to what extent;
- the hypothesis is accepted in the analysis that the development of each country's international trade, especially with its key partners, is a reflection of the static and dynamic dimensions of its economy. In other words, the development of foreign trade is a proof of the formation of international competitive position and at the same time international competitiveness of the national economy.

2. Methodical approaches

The concept of international competitiveness of the national economy has so far been ambiguously defined. Therefore, there are many suggestions and postulates measurement of international competitiveness of the national economy Broll, U., Gilroy, M., (1994), Dunning, J.H., (1992), Falvey, R.E., Kierzkowski, H., (1987), Grzywacz, W., (2001), The World Competitiveness Report 1994, (1994). The literature presents indicators of international competitiveness (competitiveness or competitive advantage) and indicators of international competitive position of the economy Misala, J., (1991, 2004).

Among the various methods for assessing a country's competitive position on the international market, ex ante competitiveness indicators (predictive modelling of economic phenomena and processes based on theoretical concepts) deserve attention. In our opinion, the greatest successes in this area of research have been achieved in Germany: Giersch, H., (1979), Horn, E.J., (1985), Kojima, K.A., (1974), Von Stackelberg, K., (1991). According to German scientists T. Griez, S. Enchel and B. Wigger (1992), the essence of international competitiveness of each national economy at any moment is to optimize the use of resources on an international scale. Scientists have suggested the Revealed Absolute Competitiveness index – RAC, based on domestic and foreign resources, consisting of two components: Revealed Absolute Internal Competitiveness (RAIC) and Revealed Absolute External Competitiveness (RAEC).

RAIC level setting depends on the internal economic performance of accessible production factors usage, due to more or less favourable exchange of the part of domestic resources for external (imported) resources. Indexes, calculated in this way, relate to the scale of the national economy. However, from a theoretical standpoint, it is possible to use them in particular directions, sectors and even particular products, under the condition of implementation of a specific database of statistics, which will

allow for comparative analysis. The absence of such base limits the range of analyses and forecasts and is a barrier to building indicators that are lower in aggregation.

The empirical analysis is based on data of Foreign Economic Activity Commodity Nomenclature on single, double, four- and nine-digit numeric expression. After all, the degree of disaggregation of data and how they are compared significantly affect the statistical picture of the phenomena and processes discussed (their trends). For example, only as deep as possible disaggregation of data can allow comparisons of the same products, and then the real level of "overlap" in the value of exports and imports can be determined, or the actual level of intra-industry trade intensity Petrose, E., (1959). Therefore, the foreign trade turnover of Ukraine is analysed on the basis of the EU data in four-stage disaggregation, which makes studied product groups meet the theoretical concept of sector in the productive industrial classification.

Based on EU statistics for the years 2003–2017, calculations have been made and the following are presented:

1) Ukraine's participation in Poland's trade turnover;

2) foreign trade balance of Poland with Ukraine;

3) changes in the share of Poland in the markets of Ukraine, calculated by the formula(1):

$$Ci = \frac{x1i}{m1i} : \frac{x0i}{m0i} * 100 \tag{1}$$

where Ci– index of change in shares;

x – export;

m- import;

i – specific product group or product;

1 – analysed period;

o – base period;

4) the commodity structure of Poland's trade turnovers with Ukraine;

5) typical structure (according to the apacity of production factors) of Poland's trade turnovers with Ukraine;

6) the balance of trade turnover of Poland with Ukraine and the balance of exports in the group of resource-intensive products;

7) Revealed Competitive Advantages indexes of Poland in the markets of Ukraine, calculated by the logarithmic formula (2) Grubel, H.G., Lloyd, P.J., (1975):

$$RCA_i = \ln\left[\frac{x_i}{m_i} : \frac{\sum x_i}{\sum m_i}\right] \tag{2}$$

where RCA_i – Revealed Competitive Advantages indexes;

x – export;

m – import;

i – specific product group or product.

The value of this formula lies in the simultaneous consistency and symmetry of the presented RCA indexes;

8) Intra-industry trade intensity indexes – IIT_i in Poland's trade with Ukraine, calculated by the formula (3) Mączyńska, E., (1999):

$$IIT_i = \frac{x_i+m_i-|x_i-m_i|}{x_i+m_i} \quad (3)$$

where IIT_i _ the Intra-industry trade intensity;

x – export;

m – import;

i - specific product group or product;

9) RCA_i and IIT_i indexes of Poland's turnovers in general, including with Ukraine by capacity codes. From a methodological standpoint, the analysis of intra-industry trade intensity is complemented by indexes of the comparative advantages of RCA_i. Such a supplement might help to determine to what extent intra-industry trade of high intensity can be a source of synergies for future export-import trade. Furthermore, it should be maintained that high levels of intra-industry trade intensity do not always keep up with a high share of the industry's exports in global exports. Therefore, when making forecasts, it is necessary to analyse the indicators of export dynamics and indicators of the level of export.

The analysis of Poland's foreign trade flows with Ukraine is supplemented by indicators of 50 most important positions (with the highest or the lowest indexes):

- in the value of export;
- in the value of import;
- in the share of trade turnover with selected countries in Poland's exports in general;
- in the value of the C_i index;
- in the competitiveness of exports calculated by RCA_i;
- Intra-industry trade intensity, expressed in IIT_iindicators.

3. Empirical analysis

The dynamics of absolute indexes of Poland's foreign trade turnover with Ukraine in 2003-2017 was generally characterized by an upward trend. In terms of the nature and trends of these indexes, their time series can be divided into 5 segments: 2003–2008, 2008–2009, 2009–2013, 2013–2015, 2015–2017. In the period from 2003 to 2008, according to Polish statistics, export from Poland to Ukraine increased at a higher rate than import from Ukraine to Poland, which led to the behaviour of the

balance of trade, which grew at almost the same pace as exports. As a result of the 2008 financial crisis, all foreign trade turnover indicators declined sharply. However, starting from 2009, the process of restoring the growth of Poland's trade relations with Ukraine began and continued until 2013. In the subsequent years, from 2013 to 2015, there was a tendency towards a decrease in Poland's foreign trade turnover with Ukraine. But in 2015, the situation changed to the opposite and lasted until 2017.

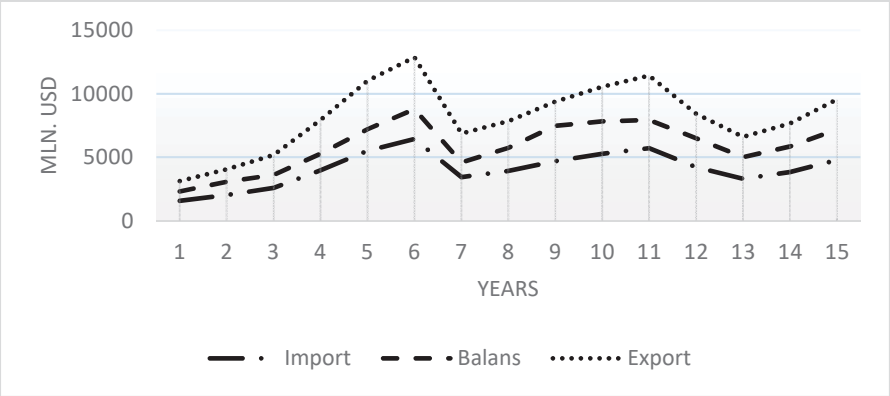


Figure 1. Dynamics of foreign trade turnover of Poland with Ukraine for 2003-2017, USD million
Source: Rocznik statystyczny handlu zagranicznego, GUS, wybrane wydania, www.stat.gov.pl

C_i indexes should be applied, taking them as an introductory analysis of a country's competitive position relative to another trading partner country.

If we analyse the dynamics of C_i indexes of Poland in general, that is Poland's trade turnovers with all their trading partners, in 2003–2017 Poland's position improved at different speeds, showing an upward trend compared to the base year (that is 2003). The export / import ratio during this period was advantageous for Poland because export growth rates were higher than imports (Table 1). Undoubtedly, this trend was influenced by favourable export / import ratios in trade with major partners.

Table 1. Dynamics of Ci indexes of Poland with Ukraine from 2003 to 2017, %

Countries	Years							
	2003	2005	2007	2009	2011	2013	2015	2017
Poland in general	100.0	111.7	107.3	116.0	113.7	125.3	128.6	127.2
Ukraine	100.0	120.9	155.2	142.0	80.1	122.8	92.7	94.6

Source: Authors' calculations according to Eurostat database.

At the same time, Poland's foreign trade turnover with Ukraine during the studied period is characterized by ambiguous dynamics. Compared to 2003, in 2004 the value of C_i of Poland with Ukraine was 7.1 pp. lower, but in the next years there was an increasing trend until 2010, when it was above 100%. However, in 2011, the situation changed dramatically to the opposite, in favour of Ukraine, and lasted until 2012. In 2013, the $\boldsymbol{C}_i$ index was in favour of Poland. However, from 2014 to 2017, it was again in favour of Ukraine.

A more detailed analysis of changes in Poland's shares in Ukrainian markets, the definition of structure and their evaluation require disaggregation of the statistical database. After all, the analysis by sections of product groups gives us an opportunity to evaluate changes – from the standpoint of the economy – whether they are positive or negative.

An analysis of the C_i indexes, which represent export-import ratios for Poland and Ukraine, makes it possible to formulate such generalized conclusions.

First, in the exchange between Poland and Ukraine, the quantitative asymmetry between export and import in certain product groups, which has increased significantly since 2003, is manifesting itself. Quantitative asymmetry is manifested in the following sections:

- basic metals and articles;
- fats and oils;
- mineral products;
- wood and wood products;
- plant-based products.

Considering the share of these products in the total exchange of Poland with Ukraine, it can be stated that they had a negative impact on the Polish balance of trade and, at the same time, a positive one on the Ukrainian balance of trade.

Secondly, at the same time, during the period there were positive structural changes in Poland's exchange between Poland and Ukraine, reflected by the increase in $C_i$indicators in the following sections:

- optical, photographic, measuring instruments and apparatus;
- machinery and equipment, electrical and electrical engineering appliances;
- sawdust, paper, cardboard and articles;
- various finished products - furniture, prefabricated buildings, toys;
- live animals and products of animal origin;
- artificial materials and articles;
- articles of stone and ceramics, glass;
- pearls, precious stones, precious metals and articles;

– chemical industry products;
– transport;
– shoes, hats.

Thus, it can be stated that, despite the positive changes, C_i indicators clearly inform that in Ukraine's exports, products with a low level of processing, labour-intensive and capital-intensive continue to prevail.

A much wider geographical diversification of Ukrainian exports could reduce the risk of fluctuations in the situation. However, the results of the analysis of the formation of competitiveness indexes for Poland and Ukraine do not confirm this. For many product groups, values of these indexes have declined due to the 2008 crisis. This is especially noticeable in the following sections: XV "Base metals and articles thereof", V "Mineral products"; IX "Wood and articles of wood"; XVII "Transport equipment"; IV "Prepared foodstuffs". Instead, for sections XI "Textiles and textile articles"; XX "Miscellaneous manufactured articles" and XVI "Machinery and mechanical appliances, electrical and electrotechnical equipment" - a characteristic wavy change in the ratio of exports and imports.

In the foreign trade turnover of Poland with Ukraine in only five sections, C_i values were less than 100%. The share in the total turnover of these sections was negligible. In the other product groups, C_i values were much higher than 100, and in some of them – even higher in a few dozen of times (for example, in the XV "Precious metals"). In many sections, C_i values were gaining wavy values and tending to decline, especially after 2008. However, in other sections, there was a clear upward trend: XVI "Machinery and mechanical appliances, electrical and electrotechnical equipment; XI "Textiles and textile articles"; XVII "Transport equipment". The persistence of trends throughout this period of course indicates the comparative advantage of Poland in these sectors, as well as the level of technological development and unequal status of the economy of Poland and Ukraine.

Summarizing the above calculations, it should not be overlooked that the informational value of C_i is limited as they relate to the exports and imports of the surveyed partners only (Poland - Ukraine). They are a measure of internal specific advantage in terms of mutual exchange. At the same time, they allow to determine the participation of these countries in the international division of labour, the level of economic and technological development. In addition, the analysis, which covers a long period (more than 5 years), allows to determine the direction and pace of structural changes in the economies of the partner countries. However, formulation of conclusions and proposals in this field requires in-depth study and analysis using other methods and criteria that will increase the plausibility, thoroughness and adequacy of the results of the study, on which the strategy and economic policy is based.

The Ukrainian industry was restructuring during the study period, moreover, at a slow pace; as well as the external trade flows were transforming, which was manifested in the gradual decline in the value of inter-branch exchange. This is reflected in a decreasing trend of absolute value of RCA, but negative values of RCA for many goods indicate a low level of competitiveness of the Ukrainian economy. The predominant reason for this situation is, as we wrote earlier, an anachronistic and clearly asymmetric assortment and species structure. Imports were dominated by transformed products with a relatively high share of value added, while exports were dominated by products with relatively low levels of processing (Table 2).

Table 2. Dynamics of RCA Poland with Ukraine by Capacity Codes 2003-2017

Product group	**2003**	**2005**	**2007**	**2009**	**2011**	**2013**	**2015**	**2017**
Raw materials	-0.87	-0.83	-0.58	-0.46	-0.39	-0.29	-0.31	-0.32
Labour-intensive goods	-	-	0.70	0.55	0.53	0.47	0.43	0.41
Capital-intensive goods	-0.24	-0.19	-0.10	-0.06	0.06	0.22	0.20	0.25
Technology-intensive products easy to imitate	-	0.04	-0.17	0.30	0.41	0.37	0.44	0.44
Technology-intensive products difficult to imitate	-	-	0.52	0.16	0.32	0.23	0.19	0.27
Non-classified goods	-	-	-	-0.17	-0.24	-0.42	-0.28	-0.26

Source: Authors' calculations according to: United Nations Database.

The intensity of Poland's inter-branch exchange with Ukraine was the highest among technology-intensive, easy to imitate and labour-intensive goods.

Poland has shown a relative advantage of many goods, for example, in the export of furniture, parts of houses, products of vine and straw, clothing and accessories, fruits and vegetables, machinery and electrical equipment and spare parts, sports equipment and toys.

At the same time, Poland has not demonstrated comparative advantages in material-intensive, capital-intensive, non-transformed land-use plants. The detailed analysis suggests that the exchange in Poland was predominantly complementary to the cross-industry type.

Indexes of the revealed comparative advantage should also be looked into from the perspective of geographical directions of foreign trade, that is Poland's trade relations with Ukraine.

Analysis of indexes of the revealed comparative advantage enables to make generalized conclusions, namely:

- the structure of RCA indicators in trade between Poland and Ukraine is inherent to countries with lower levels of economic development;
- the exchange of products was dominated by land-intensive, non-transformed and raw materials, and Ukraine has a significant comparative advantage in these products.

Therefore, it can be argued that trade relations between partner countries were determined by different product and species structure. Poland's foreign trade was clearly dominated by cross-industry exchanges, as measured by RCAs. The link between Poland's economic potential and the intensity and structure of external turnover with Ukraine is not only weak, but also heterogeneous.

Of course, the form and dynamics of the development of relations were influenced and continue to be influenced by the new geopolitical system. The intensity of Poland's trade turnovers with Ukraine was much lower than their economic potential. There could be many reasons for this, but the most important are the structural factors. It was they who had a decisive influence on the asymmetry of indicators of Ukraine's revealed comparative advantage.

However, the pragmatic value of the revealed comparative advantage indexes for creating a foreign trade development strategy is limited as they inform of the extent of the advantage or lack of it in the past and in the cross-sectoral dimension. From a methodological standpoint, this kind of analysis – as the basis of the concept of development – should, first of all, be supplemented by the indexes of the intensity of intra-industry trade.

In modern international trade, the values of specialization and intra-industry trade ("intra-industry trade" towards "two-way trade") are constantly increasing. Its essence lies in the simultaneous characterization of imports and exports by products and their components, which belong to the same industry, usually during the year Soete, L.L.G., (1990), by one country or group of countries. Intra-industry trade was studied at the 4-digit CN level of disaggregated data, aggregated to the 2-digit and 1-digit levels. An empirical analysis based on data from the Polish and Ukrainian foreign trade nomenclatures based on 4-digit CN disaggregation shows that the studied product groups in this classification correspond to the theoretical understanding of the industry in the industrial classification Petrose, E., (1959).

The intensity of intra-industry exchange increases with the positive values of IIT. The analysis of IIT indexes for Poland's trade with Ukraine in the selected years shows that the highest intensities of intra-industry trade were detected in the sections: XX "Miscellaneous manufactured articles", IV "Prepared foodstuffs", V "Mineral products", XV "Base metals and articles thereof", XIII "Articles of stone, ceramic

products, glass" and XXI "Works of art, collectors' pieces and antiques" (Table 3). The values of IIT in these sections ranged from 0.75 to 0.98. Considering the above RCA indexes, whose values fluctuated within 0.25–0.44 in 2017, we can conclude that the products of external differentiation were dominated by vertical differentiation in Poland, while intra-industry exchange of horizontally differentiated products was of subordinary importance. The turnovers were dominated by slightly transformed products, when their share in total exports was negligible.

Table 3. Dynamics of IIT Poland indicators with Ukraine by Product Groups in 2003-2017

No.		2003	2005	2007	2009	2011	2013	2015	2017
I	Live animals; animal products	0.50	0.24	0.28	0.12	0.19	0.24	0.43	0.57
II	Vegetable products	0.60	0.86	0.47	0.71	0.78	0.97	0.71	0.74
III	Fats and oils	0.46	0.31	0.04	0.05	0.04	0.10	0.13	0.05
IV	Prepared foodstuffs	0.29	0.32	0.58	0.69	0.80	0.79	0.97	0.91
V	Mineral products	0.20	0.12	0.64	0.68	0.47	1.00	0.75	0.84
VI	Products of the chemical industry	0.70	0.29	0.74	0.32	0.65	0.31	0.32	0.38
VII	Plastics and rubber and articles thereof	0.14	0.09	0.08	0.04	0.09	0.03	0.10	0.14
VIII	Raw hides and skins, articles thereof	0.72	0.29	0.38	0.43	0.34	0.40	0.19	0.37
IX	Wood and articles of wood	0.72	0.91	0.96	0.88	0.81	0.97	0.31	0.33
X	Pulp of wood, paper, paperboard and articles thereof	0.04	0.04	0.06	0.09	0.09	0.15	0.18	0.31
XI	Textiles and textile articles	0.17	0.09	0.09	0.11	0.07	0.06	0.21	0.16
XII	Footwear, headgear, etc.	0.04	0.02	0.01	0.03	0.04	0.02	0.04	0.10
XIII	Articles of stone, ceramic products, glass	0.03	0.05	0.07	0.14	0.21	0.19	0.66	0.75
XIV	Pearls, precious stones and metals, articles thereof	0.04	0.07	-	0.01	0.01	-	-	-
XV	Base metals and articles thereof	0.98	0.98	0.88	0.78	0.78	0.85	0.76	0.79

Table 3. Dynamics of IIT Poland indicators with Ukraine by Product Groups in 2003-2017 (cont.)

No.		2003	2005	2007	2009	2011	2013	2015	2017
XVI	Machinery and mechanical appliances, electrical and electrotechnical equipment	0.17	0.10	0.14	0.39	0.34	0.21	0.35	0.29
XVII	Transport equipment	0.02	0.02	0.04	0.05	0.06	0.03	0.06	0.05
XVIII	Optical, photographic, measuring, checking instruments, etc.	0.20	0.07	0.04	0.03	0.04	0.03	0.11	0.05
XIX	Arms and ammunition	0.37	-	-	-	-	-	-	-
XX	Miscellaneous manufactured articles	0.04	0.16	0.12	0.20	0.28	0.17	0.61	0.98
XXI	Works of art, collector's items and antiques	0.73	0.00	0.12	0.77	0.12	0.01	-	0.67

Source: Authors' calculations according to: Rocznik statystyczny handlu zagranicznego, GUS, wybrane wydania, www.stat.gov.pl

Thus, the structure of intra-industry trade indicators for these partners was shaped by the exchange of low-conversion, material-intensive and labour-intensive goods.

This is characteristic of the initial phase of development of intra-industry division of labor between partners, which differ in the level of technological progress, the development of system-restructuring transformation processes, as well as the level of gross domestic product per capita, which is a source of demand stimulation in intra-industry division of labor Michalet, Ch.A., (1984).

With the objective of determining the concentration of intra-industry trade, at the 4-digit CN level codes aggregated to 2-digit codes, it was showed that:

– Unfavourable trends in trade relations between Poland and Ukraine were revealed by a comparative analysis of their trade from the perspective of intra-industry division of labour. Namely: intra-industry exchange rates with values higher than 0.50 were only in 14 product groups in 2003, whereas, at the same time, IIT index was higher than 0.70 only in 8 groups. In 2017, 12 product groups' indexes were more than 0.50, and only 7 groups with an index of more than 0.70. In 2003, the number of product

groups with an index of more than 0.50 was 15, and more than 0.70 - 10; in 2017, the number of groups decreased to 13 and 7. Consequently, there has been a clear tendency for these indicators to decline for more than 14 years. Moreover, the structure of intra-industry IIT indexes encompassed, first of all, low-grade raw materials and labour-intensive products, as mentioned above. The IIT structure was conditioned by an anachronistic commodity and species structure with a clear quantitative and, above all, qualitative asymmetry.

In trade with Ukraine, only a few commodity groups have achieved intra-industry trade intensities of more than 0.85: Wood and articles of wood such as gluing and plywood, laminated timber; clothing, including women's and children's coats and kits; locomotives, in particular parts to them and to rolling-stock.

The share of the 50 product groups and products with highest IIT indexes reached about 55% of Poland's total exports. But if these groups had accounted for about 54% of exports to the EU, and especially to Germany, then their part in exchange with Ukraine would have been minimal.

The highest indicators of IIT in Poland's trade with Ukraine are typical for land-intensive animal products that are not transformed and labour-intensive products, which require hard work. However, the participation of these goods in Poland's general turnover was subordinate.

There was a noticeable convergence between the structures of imports and exports of Poland and Ukraine as an effect of the development of cooperation and partnership between these countries in the 2000s.

Adaptation of the labour model to the intra-industry division is significant not only considering Ukraine's future place in industrial and trade integration, but also because of the scale of the benefits of international division of labour within the EU (measured by GDP per capita). In addition, improving the competitiveness of the Ukrainian economy in both the export structure (transition from traditional, inter-sectoral, to modern, intra-sectoral division of labour) and in prices (reduction of interest rates and claims on speculative capital turnover) is closely linked to the improvement of balance of trade, and thus with the reduction of the foreign trade deficit.

4. Conclusions

One of the priority line of developments of Ukrainian foreign economic policy is to build relations with Poland, which is caused not only by the long tradition of Ukrainian-Polish relations, but, first of all, by the unity of political and strategic interests, active cooperation in all areas of public life of both countries. The deepening of cooperation with Poland, which has formed a qualitatively new economic system

within the EU, gives Ukraine the opportunity to use its experience to intensify its own transformation processes.

Internal and external factors of innovation remain, above all, the key source of growth in intra-industry trade intensities in Ukraine and Poland. Increasing their use will reduce not only the technological gap of Ukraine, but also the difference in gross domestic product per capita between Poland and Ukraine, thereby attracting the demand factor to the sources of intensification of intra-industry trade. Intensification of internal (sources based on accelerating market transformation processes, especially privatization and implementation of the legal and institutional market system, and increasing the share of Research & Development expenditures in GDP and accelerating the pace of technological and technological restructuring) and external innovation sources (technology import, know-how, the intense inflow of foreign direct investment and the intensification of the international division of labour with EU countries) – the only rational way of real adaptation of the Ukrainian economy and its partners to changes in the markets of the EU and the world at large.

REFERENCES

BROLL, U., GILROY, M., (1994). Aussenwirtschaftstheorie. Einführung und Neuere Ansätze, *München–Wien*., pp. 14–36.

DUNNING, J. H., (1992). The Competitive Advantage of Countries and the Activities of Transnational Corporations, *Transnational Corporations*, No. 2, pp. 11–16.

EUROSTAT, database, https://ec.europa.eu/eurostat/data/database

FALVEY, R. E., KIERZKOWSKI, H., (1987). Product Quality, Intra-Industry Trade and Perfect Competition, Protection and Competition in International Trade, Essays in Honor of W.M. Corden, /red./ H. Kierzkowski, *Oxford*, pp. 67–72.

GIERSCH, H., (1979). Aspects of Growth. Structural Change and Employment. A Schumpeterian Perspective, *Weltwirttschaftliches Archiv*, Vol. 115, pp. 1–22

GRIES, T., HENTSCHEL, C., WIGGER, B., (1992). Internationale Wettbewerbsfähigkeit – theoretisches Konzept und empirischer Befund, Universität Göttinge, pp. 27–32.

GRUBEL, H. G., LLOYD, P. J., (1975). Intra-Industry trade. The theory and measurement of international trade in differentiated products, *New York*, p. 35.

GRZYWACZ, W., (2001).Czynniki wzrostu polskiej konkurencyjności gospodarczej, *VII Kongres Ekonomistów Polskich*, Warszawa, Sesja II, Zeszyt 17, pp. 1–17.

HORN, E. J., (1985). Internationale Wettbewerbsfähigkeit von Ländern, *Tübingen*, pp. 188–196.

KOJIMA, K. A., (1974). International Trade and Foreign Investment: Substitutes or Complements, *Hitotsubashi Journal of Economics*, No. 169, pp. 237–262.

MĄCZYŃSKA, E., (1999). Bezpośrednie inwestycje zagraniczne. Światowe i lokalne czynniki dynamizujące, *Ekonomista*, No. 1–2, pp. 96.

MICHALET, CH. A., (1984). *L'integration*, pp. 7–17.

MISALA, J., (1991). Handel zagraniczny Polski a teoria handlu międzynarodowego, *Ekonomista*, nr 4–6, pp. 487–499.

MISALA, J., (2004). Współpraca gospodarcza Polski z krajami sąsiedzkimi w okresie transformacji, *Wydawnictwo Politechniki Radomskiej*, Radom, pp.16–28.

PETROSE, E., (1959). The Theory of Growth of the Firm, *Basil Blackwell, Oxford*, pp. 29–43.

ROCZNIK STATYSTYCZNY HANDLU ZAGRANICZNEGO, *GUS*, wybrane wydania, www.stat.gov.pl.

SIEBERT, H., RAUSCHER M., (1993). Neuere Entwicklung in der Aussenhandelstheorie, *Kiel Working Paper, Institute of World Econ*omics, Kiel, No 476, pp. 17–27.

SOETE, L. L. G., (1990). Technical Change Theory and International Trade Competition, *Science Technology and Free Trade*, /red./ J. de la Monthe, L. M. Ducharme, London, pp. 123–146.

THE WORLD COMPETITIVENESS REPORT 1994, (1994). *World Economic Forum*, Luosanne, pp. 18–37.

UNITED NATIONS DATABASE, www.unstats.un.org.

VON STACKELBERG, K., (1991). Internationale Wettbewerbsfähigkeit bei zur nehmenden intra–industriellen Handelsbezichungen mit Schwellenländern: Analyse des Handels der Bundesrepublik Deutschalnd, Niedersachens und Japans mit den Schwellenländern Ost–/Südost–Asien/, Berlin, pp. 61–92.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 157–172, DOI 10.21307/stattrans-2020-018
Submitted – 15.02.2019; accepted for publishing – 18.02.2020

The positional MEF-TOPSIS method for the assessment of complex economic phenomena in territorial units

Aleksandra Łuczak[1], **Małgorzata Just**[2]

ABSTRACT

In this paper, the authors propose a new methodological approach to the construction of a synthetic measure, where the objects are described by variables with strong asymmetry and extreme values (outliers). Even a single extreme value (very large or very small) of a variable for the object may significantly affect the attribution of an excessively high or low rank in the final ranking of objects. This dependence is particularly apparent when using the classical TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method. The aim of the study is to present the application potential of the positional MEF-TOPSIS method for the assessment of the level of development of complex economic phenomena for territorial units. In the positional TOPSIS method, the application of the spatial median of Oja, which limits the impact of strong asymmetry, is proposed. In order to weaken the influence of extreme values, the Mean Excess Function (MEF) is used, by means of which it is possible to identify the limits of extreme values and establish model objects. The proposed approach is used to assess the financial self-sufficiency of Polish municipalities in 2016. The study finally compares the results of applications of positional MEF-TOPSIS and the classic and positional TOPSIS methods.

Key words: synthetic measure, TOPSIS, spatial median of Oja, Mean Excess Function.

1. Introduction

The complex nature of the economics phenomena taking place in the real world causes many problems in the research for territorial units. Complex economics phenomena (e.g. financial self-sufficiency, socio-economic development, standard of living) include many various problems, which are often difficult to identify and quantify. Therefore, these phenomena cannot be measured directly, but they can only be evaluated based on different criteria and variables. As there are many aspects to the

[1] Faculty of Economics and Social Sciences, Poznań University of Life Sciences, Poland. E-mail: aleksandra.luczak@up.poznan.pl. ORCID: https://orcid.org/0000-0002-3149-7748.
[2] Faculty of Economics and Social Sciences, Poznań University of Life Sciences, Poland. E-mail: malgorzata.just@up.poznan.pl. ORCID: https://orcid.org/0000-0001-7655-6046.

process, various analyses are carried out. One type of analysis is the assessment of the phenomenon level by a synthetic measure. In this case, the use of the classical statistical methods imposes some restrictions, which often lead to the excessive simplification of the analysis. In such studies the non-classical multi-criteria quantitative methods are very useful. Therefore, the authors proposed a novel hybrid approach to the construction of a synthetic measure in the study.

The procedure for constructing the synthetic measure is a multi-stage process. One of the most important stages is selecting variables for studies. It is a very complicated issue, especially when unusual data values (e.g. extreme values) appear or strong asymmetry occurs within variables. It may result from the specifics of the complex phenomenon studied. These anomalous observations have a crucial impact on the results of the research. The occurrence of even only one problematic value of the variable for the object may significantly affect the attribution of the incorrect (excessively high or excessively low) rank in the final ranking of objects. This also leads to incorrect identification of types of the complex economics phenomena on its basis. Therefore, it is necessary to seek optimal methods to identify extreme values and develop new methodological approaches which are resistant to these phenomena.

In this study the authors propose a novel hybrid methodological approach to the construction of the synthetic measure, where the objects are described by the variables with extreme values and strong asymmetry. The aim of the study is to present the application potential of the positional MEF-TOPSIS method to assess the development level of the complex economics phenomena for territorial units. The proposed method is based on the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method (Hwang and Yoon, 1981). The TOPSIS method is very useful in constructing the ranking of objects described by many variables. The synthetic measure is constructed on the base of the distances from the model values (positive ideal solution and negative ideal solution). In the case of data set with unusual variables, the assumption that the maximum and minimum values of the variables are model values leads to excessive remoteness from typical values of the considered variables and consequently narrows the range of variability of the constructed synthetic measure. The problem may be solved by application of the Mean Excess Function (MEF) for identifying the limits of extreme values and establishing the model objects. As a result, the influence of extreme values (outliers) was reduced, whereas the spatial median of Oja was used in order to limit the impact of strong asymmetry. This novel hybrid approach was used in the assessment of financial self-sufficiency of local administrative units in Poland in 2016. The research hypothesis was that the construction of a synthetic measure for complex economic phenomena, described by variables with

extreme values, using positional MEF-TOPSIS allows to perform more accurate classifications of objects and to distinguish more homogeneous types than approaches using the classical and positional TOPSIS methods.

2. Methods

The classical TOPSIS method was first presented by Hwang and Yoon (1981) and is the most established technique for solving Multi-Criteria Decision Making (MCDM) problems. TOPSIS is based on the idea that the best object should have the shortest distance from the positive ideal solution and the longest distance from the negative ideal solution. The main assumption of the method is that the variables monotonically increase or decrease. TOPSIS was further developed by Yoon (1987) and also Hwang, Lai and Liu (1993). Nowadays, many different extensions of TOPSIS exist. They are based on triangular fuzzy numbers (Chen, 2000), interval data (Jahanshahloo, Lotfi and Izadikhah, 2006), interval-valued fuzzy sets (Chen and Tsao, 2008), interval type-2 fuzzy sets (Chen and Lee, 2010), interval-valued intuitionistic fuzzy sets (Li, 2010) and multi-granularity linguistic assessment information (Liu, Chan and Ran, 2013), positional notation (Wysocki 2010, Kozera, Łuczak and Wysocki, 2016).

Extended versions of TOPSIS have solved many methodological problems in the assessment of the development level of the complex economics phenomena. The interval TOPSIS method is employed when determining the variable values of the object precisely is difficult and the values can be presented by means of intervals, i.e. two extreme variable values, which are minimum and maximum (Łuczak, 2015). The fuzzy TOPSIS method allows for the construction of the synthetic measure and the linear ordering of the objects described by means of both metrical and non-metrical (ordering) variables, owing to the transformation of the ordering characteristics into fuzzy numbers, which is one of the ways to strengthen the measurement scale (Wysocki and Łuczak, 2009). Furthermore, many hybrid approaches and their application have been presented. The approaches combine TOPSIS with the following methods: AHP (Kusumawardani and Agintiara, 2015), Pareto and genetic algorithm method (Taleizadeh, Niaki and Aryanezhad, 2009), SAW and GRA (Wang, Zhu and Wang, 2016), POT (Łuczak, Just and Kozera, 2018). A broad review of different versions of the TOPSIS method and their application was carried out by Behzadian et al. (2012); Velasquez and Hester (2013); Mardani, Jusoh and Zavadskas (2015); Afsordegan et al. (2016); Nădăban, Dzitac and Idzitac (2016), Zavadskas et al. (2016).

The novel hybrid approach to the construction of the synthetic measure, proposed by the authors, combines Technique for Order of Preference by Similarity to Ideal

Solution (TOPSIS) and the Mean Excess Function (MEF). The procedure based on the modified positional MEF-TOPSIS method includes seven main stages:

Stage 1. Selection of variables on the complex phenomenon and identification of extreme values by application of the Mean Excess Function,

Stage 2. Determination of the impact direction of variables in relation to the complex phenomenon,

Stage 3. Normalization of the variable values with utilization of the spatial median of Oja,

Stage 4. Calculation of the positive ideal solution and negative ideal solution,

Stage 5. Calculation of the distance of each object from positive and negative ideal solutions,

Stage 6. Calculation of values of the synthetic measure,

Stage 7. Ranking classification of objects and identification of the types.

The first stage is the selection of variables of the complex phenomenon and the identification of extreme values. The selection of variables for objects (e.g. territorial units: countries, regions, states, districts, municipalities) is to be carried out based on substantive and statistical analysis. The set of variables describing the complex phenomenon (e.g. financial self-sufficiency, socio-economic development, standard of living) for territorial units is usually characterized by strong asymmetry or includes extreme values. In the real data studied, the choice of a suitable threshold of extreme values is frequently a very difficult task. In order to identify extreme values, an approach based on the Extreme Value Theory (EVT) was used. The EVT is a powerful and robust theory for studying the tail behaviour of a distribution of variable. Two approaches are used in the EVT to model extreme values. The first approach is based on the Block Maxima Model (BMM), estimating the distribution of extremes. The second is based on the Peaks over Threshold Model (POT), estimating the tail of the distribution of the variable. The Mean Excess Function plot is useful for determining the appropriate thresholds for extreme values of the variable in the POT. The MEF is also a convenient visual tool for examining whether a variable has a specific distribution (Chen et al., 2015). In the research, the MEF allows a threshold (limit) of extreme values to be established.

In the POT (see e.g. McNeil, 1999, Echaust, 2014), the tail of the distribution of the variable is modelled using the Generalized Pareto Distribution (GPD), while the beginning of this tail is determined by specifying a threshold value (ul_k). In this approach, the starting point for the considerations is the conditional distribution of excess over ul_k of random variable X_k (k^{th} variable), which is defined by the formula:

$$F_{ul_k}(y_k) = P(X_k - ul_k \leq y_k \,|\, X_k > ul_k) = \frac{F(y_k + ul_k) - F(ul_k)}{1 - F(ul_k)}, \qquad (1)$$

where: $y_k = x_k - ul_k > 0$, F – an unknown distribution function of random variable X_k. According to the Pickands–Balkema–de Haan theorem, for a sufficiently large ul_k, the distribution function F_{ul_k} is definite and well approximated by the GPD with the distribution function:

$$G_{\xi,\beta}(x_k - ul_k) = \begin{cases} 1-(1+\xi(x_k-ul_k)/\beta)^{-1/\xi}, & \xi \neq 0 \\ 1-\exp(-(x_k-ul_k)/\beta), & \xi = 0 \end{cases}, \tag{2}$$

where: $\beta > 0,\ x_k - ul_k \geq 0$ for $\xi \geq 0$ and $0 \leq x_k - ul_k \leq -\beta/\xi$ for $\xi < 0$. This distribution has two parameters: ξ – the shape parameter determining the thickness of the tail and β – the scale parameter. The positive values of the shape parameter mean that the distribution has fat tails. It is connected with an increased probability of extreme variable values. In turn, negative values of the shape parameter denote that the distribution has the finite right endpoint. The choice of the threshold value ul_k is very important, because it affects the obtained values of the GPD parameter estimators. If N is the number of observations, N_{ul_k} is the number exceeding ul_k, the estimator of the distribution function F is calculated from the following formula:

$$\hat{F}(x_k) = 1 - \frac{N_{ul_k}}{N}\left(1+\hat{\xi}\frac{(x_k-ul_k)}{\hat{\beta}}\right)^{-1/\hat{\xi}}. \tag{3}$$

Selecting the ul_k threshold should take into account the specifics of the variable and their number. The threshold selection methods have been described, for example, by Coles (2001). One of the methods is the analysis of the stability of the GPD parameters estimates. This method was used in POT-TOPSIS by Łuczak, Just and Kozera (2018). The next method is based on the analysis of the graph of the Mean Excess Function. In this method, the starting point is the conditional expected value:

$$E(X_k - ul_k \,|\, X_k > ul_k) = \frac{\beta(ul_k)}{1-\xi}, \quad \xi < 1. \tag{4}$$

Since $\beta(ul_k)$ depends linearly on ul_k, the empirical estimator of the conditional expected value also must depend linearly on ul_k. Therefore, the graph of the Mean Excess Function:

$$\left\{\left(ul_k, \frac{1}{N_{ul_k}}\sum_{i=1}^{N_{ul_k}}(x_{ik}-ul_k)\right) : ul_k < x_{k\max}\right\} \tag{5}$$

after exceeding ul_k should be linear. The lower limit (ll_k) of the variable is determined by performing calculations for the values of the variable with a negative coefficient.

Identification of even one variable with extreme values (or even one value) does not allow the use of a classical approach to the construction of a synthetic measure. In the

second stage, the impact direction of variables in relation to the complex phenomenon is determined. The selected variables have a positive (stimulating) or negative (de-stimulating) influence on the phenomenon. Variables that have a stimulating influence, contribute to increasing the phenomenon level. These variables are called stimulants. Variables that have a de-stimulating influence, decrease the phenomenon level. The destimulating variables are called destimulants. Destimulants should be converted into stimulants with the use of a negative coefficient transformation:

$$x_{ik} = a - b \cdot x_{ik}^{D}, \tag{6}$$

where: x_{ik}^{D} – value of the k^{th} variable, identified as a destimulant $(k \in I_D)$, in the i^{th} object (i = 1, …, N), x_{ik} – value of the k^{th} variable (k = 1, …, K) converted into a stimulant, a, b – constants establish arbitrarily (e.g. a = 0 and b = 1).

The third stage is the normalization of variable values. There are many different ways to normalize the value of variables and these methods have different properties. The choice of the best approach for variables of the complex economic phenomena is not simple and requires innovative methods and approaches. In the case of the assessment of a complex phenomenon of units, variables with extreme values or characterized by an asymmetrical distribution of their values are often observed. Therefore, to solve this problem, the modified median standardization was proposed using the spatial median of Oja (cf. Lira, Wysocki and Wagner, 2002). The spatial median of Oja is resistant to variables with strong asymmetry (Oja, 1983, Ronkainen, Oja and Orponen, 2002). Additionally, for limiting the influence of extreme values of variables, threshold values of variables ul_k and ll_k (k = 1, …, K) were applied in the formula of the modified median standardization:

$$z_{ik} = \begin{cases} \dfrac{ll_k - m\tilde{e}d_k}{1.4826 \cdot m\tilde{a}d_k} & \text{for } x_{ik} \le ll_k \\ \dfrac{x_{ik} - m\tilde{e}d_k}{1.4826 \cdot m\tilde{a}d_k} & \text{for } ll_k < x_{ik} < ul_k \\ \dfrac{ul_k - m\tilde{e}d_k}{1.4826 \cdot m\tilde{a}d_k} & \text{for } x_{ik} \ge ul_k \end{cases}, \tag{7}$$

where: x_{ik} – value of the k^{th} variable in the i^{th} object, $m\tilde{e}d_k$ – Oja's median vector ($\boldsymbol{\theta}$) component corresponding to the k^{th} variable, $m\tilde{a}d_k = med_i\left|x_{ik} - m\tilde{e}d_k\right|$ – median absolute deviation of k^{th} variable values from the median component of the k^{th} variable, 1.4826 – a constant scaling factor corresponding to normally distributed data, $\sigma \approx E\left(1.4826 \cdot m\tilde{a}d_k\left(X_1, X_2, ..., X_K\right)\right)$, $\sigma -$ standard deviation (see, e.g. Młodak 2006). An alternative version of the spatial median was given by Weber (1909).

The median standardization is calculated for Winsorized data. Winsorization is the transformation of a variable by limiting its extreme values. In the process of Winsorization a specified number of extreme values of a variable is replaced with a constant (smaller or bigger) value. The constants are established based on the MEF plot. The authors propose to adopt threshold values of variables ul_k and ll_k ($k = 1, 2, \ldots, K$) as the constants in Winsorization.

In the fourth stage, the positive ideal solution (PIS) and the negative ideal solution (NIS) were calculated (Hwang and Yoon, 1981):

$$\text{PIS} \qquad A^{+} = \left(\max_i(z_{i1}), \max_i(z_{i2}), \ldots, \max_i(z_{iK})\right) = \left(z_1^{+}, z_2^{+}, \ldots, z_K^{+}\right), \tag{8}$$

$$\text{NIS} \qquad A^{-} = \left(\min_i(z_{i1}), \min_i(z_{i2}), \ldots, \min_i(z_{iK})\right) = \left(z_1^{-}, z_2^{-}, \ldots, z_K^{-}\right). \tag{9}$$

The PIS are the best values of variables, which are stimulant or are transformed into stimulant, whereas the NIS are the worst values of normalized variables.

Next, Manhattan distances (L_1 distances) for each object from the PIS (A^{+}) and the NIS (A^{-}) were calculated based on (stage 5):

$$d_i^{+} = \sum_{k=1}^{K} \left| z_{ik} - z_k^{+} \right|, \quad d_i^{-} = \sum_{k=1}^{K} \left| z_{ik} - z_k^{-} \right|. \tag{10}$$

The sixth stage involves calculation of values of the synthetic measure with the use the Hwang and Yoon's formula (1981):

$$S_i = \frac{d_i^{-}}{d_i^{-} + d_i^{+}} \quad (i = 1, \ldots, N). \tag{11}$$

The higher the values of the synthetic measure, the better the development level of the complex phenomenon and vice versa.

The calculated values of the synthetic measure are the basis of ranking the objects and identification of their typological classes (stage 7). Class identification can be carried out by different statistical methods or in an arbitrary manner. In the study the arbitrary approach based on a division of synthetic measure into eight classes is proposed, assuming:

Class I (extremely high level)	$S_i \in \langle 0.875, 1.000 \rangle$
Class II (very high level)	$S_i \in \langle 0.750, 0.875)$
Class III (high level)	$S_i \in \langle 0.625, 0.750)$
Class IV (medium-high level)	$S_i \in \langle 0.500, 0.625)$
Class V (medium-low level)	$S_i \in \langle 0.375, 0.500)$
Class VI (low level)	$S_i \in \langle 0.250, 0.375)$
Class VII (very low level)	$S_i \in \langle 0.125, 0.250)$
Class VIII (extremely low level)	$S_i \in \langle 0.000, 0.125)$

The classes of the level of financial self-sufficiency for municipalities were evaluated by statistical criteria. For this purpose, measures of homogeneity were applied. It is a concept related to the degree of similarity of objects in the same class. The idea of the measures is based on distances of objects from the centre of gravity of a class (cf. Młodak, 2006):

$$d_{ic} = \sum_{k=1}^{K} \left| x_{ik} - v_{kc} \right| \quad (c = 1,\ldots, C), \tag{12}$$

$$\bar{d}_c = \frac{1}{N_c} \sum_{i \in P_c} d_{ic}, \tag{13}$$

$$H_{MO} = \frac{1}{r} \sum_{c=1}^{C} \bar{d}_c , \tag{14}$$

where: d_{ic} – intra-class distances for each object in c^{th} class from the centre of gravity of the c^{th} class (c = 1,…, C), C – the number of typological classes, $v_c = (v_{1c}, v_{2c}, ..., v_{Kc}) = \left(\underset{i \in P_1}{\text{med}}(x_{ik}), \underset{i \in P_2}{\text{med}}(x_{ik}), ..., \underset{i \in P_C}{\text{med}}(x_{ik}) \right)$ – the centre of gravity of the c^{th} class (median of its elements), P_c – a set of subscripts of objects belonging to the c^{th} class, $\bar{d}_c$ – the partial mean measure of homogeneity of c^{th} class, N_c – the number of objects in c^{th} class, H_{MO} – the total mean measure of homogeneity, r – the number of non-empty classes.

Also, the total inter-clusters homogeneity measure is based on the idea of Hubert and Lewin (cf. 1976) is proposed by the authors:

$$H_O^c = \frac{\bar{d}_c - \min_i (d_{ic})}{\max_i (d_{ic}) - \min_i (d_{ic})}, \quad H_O^c \in \langle 0,1 \rangle , \tag{15}$$

$$H_O = \frac{1}{r} \sum_{c=1}^{C} H_O^c , \tag{16}$$

where H_O – the total measure of homogeneity. Also, the total mean intra-class distance is a useful measure for assessing homogeneity of clusters: $\bar{d} = \frac{1}{N} \sum_{i=1}^{N} d_{ic}$. The lower the value of the measures $(H_{MO}, H_O, \bar{d})$, the more homogeneous the classes.

3. Results of research

The proposed approach was used to assess the level of financial self-sufficiency of municipalities (N=2412) in Poland in 2016. The study was based on statistical data from 2016 coming from the Central Statistical Office of Poland (*Local Data Bank*). In the first stage of the study, five indicators (variables) were selected based on a substantive and statistical analysis. These were the following indicators: x_1 – own income per capita (in PLN), x_2 – share of own income in total income (in %), x_3 – transfer income (including specific grants and the general subsidy) per capita (in PLN), x_4 – share of tax income (tax bill of agriculture, forestry, real estate, from transport fund of civil law, income from taxation, income from mining fee) in current income (budgetary revenue other than income property) (in %), x_5 – self-financing rate (share of operating surplus/deficit and capital income in capital expenditure).

Table 1. Descriptive statistics and threshold values of the indicators of municipalities in Poland in 2016

Specification	Variables				
	x_1	x_2	x_3	x_4	x_5
Mean	1575.14	38.17	2442.17	15.04	191.10
Median	1382.34	35.98	2441.40	14.01	141.59
Max	45340.71	95.04	4521.20	59.74	18507.88
Min	508.17	13.08	1163.22	2.24	-134.22
St. dev.	123.45	13.20	550.34	6.86	487.88
Mad	352.63	9,30	412.50	3.99	40.65
Range	44832.54	81.96	3357.98	57.50	18642.10
Skewness	20.35	0.58	0.18	1.34	29.95
Ex. kurtosis	667.11	-0.15	-0.37	3.30	1026.09
ll_k	657.275	17.268	1310.765	3.840	73.015
ul_k	2239.219	65.852	3688.543	27.673	270.690

Descriptive statistics of the indicators are presented in Table 1. The greatest volatility, measured by the range, standard deviation and median absolute deviation, was found for x_1, x_3 and x_5. Positive skewness was observed for all indicators, with extremely high skewness noticed for x_1 and x_5. The distributions of three indicators x_1, x_4 and x_5, demonstrated positive kurtosis. This means that extreme values in the indicators appear more frequently that in the normal distribution. In order to limit the influence of extreme values, the limits of extreme values of indicators were established based on analysis of the Mean Excess Function graph (Table 1). The calculations were performed with package *fExtremes* in *R* (Wuertz, Setz and Chalabi, 2017). The results of the analysis indicated the occurrence of very fat right tails of the distribution of indicators x_1 (estimation of shape parameter 0.42) and x_5 (estimation of shape

parameter 0.75). The distribution of Winsorized data demonstrated small skewness. Moreover, kurtosis for indicators x_1, x_4, x_5 was close to normal distribution.

In the second stage, it was assumed that four indicators have a stimulating effect (x_1, x_2, x_4, x_5) and one indicator (x_3) has a de-stimulating effect on the level of financial self-sufficiency of municipalities. The indicator, which was a destimulant, was converted into an opposite indicator type.

In next stage, the values of variables were standardized by the modified Oja's median standardization. The spatial median of Oja was calculated with *OjaNP* package in *R* (Fischer et al., 2015). The standardized values of indicators allowed the authors to calculate the distances of each municipality considered from the PIS and the NIS using the Manhattan distance. In the sixth stage, the values of the synthetic measure were calculated using the positional MEF-TOPSIS method. This allowed the authors to identify eight types of municipal financial self-sufficiency levels in Poland in 2016 (Table 2). The proposed approach (approach I) was compared with the classical TOPSIS (approach II), MEF-TOPSIS (approach III) and positional TOPSIS by Wysocki (approach IV).

Table 2. Typological classification of municipalities in Poland in terms of the level of financial self-sufficiency in 2016

Class	Level of financial self-sufficiency	S_i	Approaches							
			I		II		III*		IV**	
			N_c	%	N_c	%	N_c	%	N_c	%
I	extremely high	⟨0.875, 1.000⟩	32	1.3	0	0.0	11	0.5	0	0.0
II	very high	⟨0.755, 0.875)	194	8.0	0	0.0	112	4.6	2	0.1
III	high	⟨0.625, 0.750)	359	14.9	0	0.0	425	17.6	12	0.5
IV	medium-high	⟨0.500, 0.625)	431	17.9	2	0.1	496	20.6	43	1.8
V	medium-low	⟨0.375, 0.500)	517	21.4	1	0.0	562	23.3	240	10.0
VI	low	⟨0.250, 0.375)	521	21.6	1	0.0	525	21.8	827	34.3
VII	very low	⟨0.125, 0.250)	294	12.2	135	5.6	248	10.3	976	40.5
VIII	extremely low	⟨0.000, 0.125)	64	2.7	2273	94.2	33	1.4	312	12.9

* MEF-TOPSIS with Winsorized data. ** – positional TOPSIS with standardization using Oja's spatial median and pseudo distances for each object from the PIS and the NIS calculated using median absolute deviation.

The rankings obtained by means of the applied methods indicate differences (Wilcoxon rank sum test) in the values of synthetic measures and the arrangement of municipalities into classes (Tables 2, 3). A similar classification was created only for the positional MEF-TOPSIS and the MEF-TOPSIS methods. The values of synthetic measures for these methods did not differ significantly (at the significance level of 0.1,

Wilcoxon rank sum test). These methods are resistant to the occurrence of outliers. The use of the MEF graphs analysis and the determination of the values of PIS and NIS on this basis resulted in greater ranges of variability in the values of synthetic measures. The synthetic measures values fall within the following intervals: ⟨0.003, 0.961⟩ and ⟨0.006, 0.940⟩, respectively. This allowed eight types of municipalities to be determined (including eight levels of financial self-sufficiency of municipalities, from extremely low to extremely high). The application of the classical and positional TOPSIS methods was associated with obtaining the values of the synthetic measures from the intervals ⟨0.011, 0.515⟩ and ⟨0.023, 0.827⟩, respectively. The synthetic measure values in the classical and positional TOPSIS methods are more concentrated and have stronger skewness than in the positional MEF-TOPSIS and the MEF-TOPSIS methods. In the case of the classical TOPSIS method, almost all municipalities (99.8%) were qualified to classes representing an extremely low level or a very low level of financial self-sufficiency. The use of the positional TOPSIS method allowed to distinguish seven levels of financial self-sufficiency of municipalities, from extremely low to very high. In this case, almost all municipalities (almost 98%) were qualified to classes representing levels of financial self-sufficiency from extremely low to medium-low. Despite the indicated differences in the distribution of the values of synthetic measures obtained for the applied methods, the high values of Spearman's and Kendall's rank correlation coefficients of the synthetic measures pointed to a high agreement of the linear ordering results. However, the values of the synthetic measure obtained in classical and positional TOPSIS approaches, especially the values close to zero, do not allow for a meaningful identification of types of financial self-sufficiency of municipalities.

Table 3. Descriptive statistics of the synthetic measures of financial self-sufficiency of municipalities in Poland according to approaches

Specification	Approaches			
	I	II	III	IV
Max	0.961	0.515	0.940	0.827
Min	0.003	0.011	0.006	0.023
Range	0.959	0.504	0.935	0.804
Median	0.452	0.082	0.459	0.237
Mean	0.467	0.085	0.468	0.249
Skewness	0.172	3.309	0.076	0.673
Ex. kurtosis	-0.774	40.106	-0.774	0.818

The classes of the level of financial self-sufficiency for municipalities were evaluated by statistical criteria. Measures of homogeneity were calculated for this purpose (Table 4). Values of the calculated measures indicate that the use of the positional MEF-TOPSIS and the MEF-TOPSIS methods allowed to identify municipality classes characterized by better homogeneity than using the classical and positional TOPSIS

methods. It should be added that the highest homogeneity was recorded for classes obtained with the positional MEF-TOPSIS method.

Table 4. Values of homogeneity measures according to approaches

Specification	Approaches			
	I	II	III	IV
H_{MO}	760.9	6775.5	840.5	3734.7
$\overline{d}$	621.2	963.5	622.2	643.4
H_O	0.017	0.215	0.019	0.186

On the basis of the analyses carried out, there is no reason to reject the research hypothesis (the construction of a synthetic measure for complex economic phenomena, described by variables with extreme values, using positional MEF-TOPSIS allows to perform more accurate classifications of objects and to distinguish more homogeneous types than approaches using the classical and positional TOPSIS methods).

4. Conclusion

The proposed positional MEF-TOPSIS method (using Oja's spatial median) of linear ordering of objects reduces the impact of strong asymmetry and extreme values of variables describing objects. The Mean Excess Function to identify extreme values and establish model objects (PIS and NIS) was used in this approach for this purpose. In the case of linear ordering, the occurrence of even one outlier (very large or very small) for an object can significantly affect the assignment of an excessively high or low rank in the final classification of objects. This is particularly evident when the classical TOPSIS method is used. Using the positional TOPSIS with standardization based on Oja's spatial median and pseudo distances for each object from the PIS and the NIS, calculations using median absolute deviation improve the classification of objects. The reason is that in the classical TOPSIS method the squared deviations of each multi-variable object from the PIS and the NIS are calculated and aggregated, whereas in the positional TOPSIS the median from absolute deviations is used, which enables locating the centre of the set of absolute differences between each multi-variable object and the PIS and the NIS. In turn, it makes it possible to limit the impact of outliers on the construction of the synthetic measure. Similar rankings and classifications of objects gave the positional MEF-TOPSIS and the MEF-TOPSIS while in the case of the application of the first method, classes are characterized by greater homogeneity.

The typology of municipalities in Poland in 2016 created on the positional MEF-TOPSIS basis well reflects the inter-class differences in financial self-sufficiency of municipalities. It includes eight classes of municipalities, spanning from an extremely low to extremely high level of financial self-sufficiency.

The research showed that the construction of a synthetic measure for complex economic phenomena, described by variables with extreme values, using the positional MEF-TOPSIS allows to perform more correct classifications of objects and to distinguish more homogeneous types than approaches using the classical and positional TOPSIS methods.

The authors recommend using the positional MEF-TOPSIS in the assessment of the development level of complex economics phenomena for territorial units described by variables with extreme values. In order to establish limits in the procedure of Winsorization, the authors recommend using the Mean Excess Function graphs analysis to determine the threshold of extreme values along with other statistical methods and substantive criteria to avoid mechanical and excessive Winsorization. The Winsorization based on only one criterion can lead to improper placement of objects in classes.

REFERENCES

AFSORDEGAN, A., SÁNCHEZ, M., AGELL, N., ZAHEDI, S., CREMADESL, L. V., (2016). Decision making under uncertainty using a qualitative TOPSIS method for selecting sustainable energy alternatives, *International Journal of Environmental Science and Technology*, 13(6), pp. 1419–1432.

BEHZADIAN, M., OTAGHSARA, K. S., YAZDANI, M., IGNATIUS, J., (2012). A state-of the-art survey of TOPSIS applications, *Expert Systems with Applications*, 39(17), pp. 13051–13069.

CHEN, C.-T., (2000). Extension of the TOPSIS for group decision-making under fuzzy environment, *Fuzzy Sets and Systems*, 114(1), pp. 1–9.

CHEN, S.-M., LEE, L.-W., (2010). Fuzzy multiple attributes group decision-making based on the interval type-2 TOPSIS method, *Expert Systems with Applications*, 37(4), pp. 2790–2798.

CHEN, J., LEI, X., ZHANG, L., PENG, B., (2015). Using extreme value theory approaches to forecast the probability of outbreak of highly pathogenic influenza in Zhejian, China, PLOS ONE, 10(2), [online] Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118521> [Accessed 5 February 2019].

CHEN, T.-Y., TSAO, C.-Y., (2008). The interval-valued fuzzy TOPSIS method and experimental analysis, *Fuzzy Sets and Systems*, 159(11), pp. 1410–1428.

COLES, S., (2001). An Introduction to Statistical Modeling of Extreme Values, Springer, London.

ECHAUST, K., (2014). Ryzyko zdarzeń ekstremalnych na rynku kontraktów futures w Polsce (Risk of extreme events on the futures market in Poland), Poznań University of Economics and Business, Poznan.

FISCHER, D., MÖTTÖNEN, J., NORDHAUSEN, K., VOGEL, D., (2015). Package 'OjaNP'. Multivariate Methods Based on the Oja Median and Related Concepts, Version 0.9-8, [online] Available at: <https://cran.r-project.org/web/packages/OjaNP/>.

HUBERT, L., LEVIN, J., (1976). A general statistical framework for assessing categorical clustering in free recall, *Psychological Bulletin*, 83(6), pp. 1072–1080.

HWANG, C. L., LAI, Y. J., LIU, T. Y., (1993). A new approach for multiple objective decision making, *Computers and Operational Research*, 20, pp. 889–899.

HWANG, C. L., YOON, K., (1981). Multiple attribute decision-making: Methods and applications, Springer, Berlin.

JAHANSHAHLOO, G. R., LOTFI, F. H. IZADIKHAH, M., (2006). An algorithmic method to extend TOPSIS for decision-making problems with interval data, *Applied Mathematics and Computation*, 175(2), pp. 1375–1384.

KOZERA, A., ŁUCZAK, A., WYSOCKI, F., (2016). The application of classical and positional TOPSIS methods to assess financial self-sufficiency levels in local government units, [In:] Palumbo F., Montanari A., Vichi M. (eds.), Data Science. Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Cham, pp. 273–284.

KUSUMAWARDANI, R. P., AGINTIARA, M., (2015). Application of Fuzzy AHP-TOPSIS Method for Decision Making in Human Resource Manager Selection Process, *Procedia Computer Science*, 72, pp. 638–646.

LI, D.-F., (2010). TOPSIS-based nonlinear-programming methodology for multiattribute decision making with interval-valued intuitionistic fuzzy sets, *IEEE Transactions on Fuzzy Systems*, 18(2), pp. 299–311.

LIRA, J., WYSOCKI, F., WAGNER, W., (2002). Mediana w zagadnieniach porządkowania obiektów wielocechowych (Median in the ordering problems of multi-variable objects), [In:] Paradysz, J. (ed.), Statystyka regionalna w służbie samorządu terytorialnego i biznesu, (Regional statistics in the service of local government and business), Academy of Economics in Poznań, Poznan, pp. 87–99.

LIU, S., CHAN, F. T., RAN, W., (2013). Multi-attribute group decision-making with multi-granularity linguistic assessment information: An improved approach based on deviation and TOPSIS, *Applied Mathematical Modelling*, 37(24), pp. 10129–10140.

LOCAL DATA BANK, (2018). Central Statistical Office, Poland, [online] Available at: <www.stat.gov.pl> [Accessed 9 July 2018].

ŁUCZAK, A., (2015). Wykorzystanie rozszerzonej interwałowej metody TOPSIS do porządkowania liniowego obiektów (The use of the extended interval TOPSIS methods for linear ordering of objects.), [In:] Jajuga, K., Walesiak, M. (eds.), Taksonomia, 25, Klasyfikacja i analiza danych. Teoria i zastosowania. Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 385, (Taxonomy No. 25, Classification and Data Analysis. Theory and Applications. Research Papers of Wrocław University of Economics No. 385), Wroclaw University of Economics and Business, Wroclaw, pp. 147–155.

ŁUCZAK, A., JUST, M., KOZERA, A., (2018). Application of the positional POT-TOPSIS method to the assessment of financial self-sufficiency of local administrative units, [In:] Cermakova, K., Mozayeni, S., Hromada, E. (eds.), Proceedings of the 10th Economics & Finance Conference, Rome, International Institute of Social and Economic Sciences and International Society for Academic Studies, z.s., Prague, pp. 610-621, [online] Available at: <http://www.iises.net/proceedings/10th-economics-finance-conference-rome/table-of-content?cid=69&iid=043&rid=10173> [Accessed 5 February 2019].

MARDANI, A., JUSOH, A., ZAVADSKAS, E. K., (2015). Fuzzy multiple criteria decision-making techniques and applications – Two decades review from 1994 to 2014, *Expert Systems with Applications*, 42(8), pp. 4126–4148.

MCNEIL, A. J., (1999). Extreme Value Theory for Risk Management, Department Mathematics ETH Zentrum, Zurich.

MŁODAK, A., (2006). Analiza taksonomiczna w statystyce regionalnej (Taxonomic analysis in regional statistics), Difin, Warsaw.

NĂDĂBAN, S., DZITAC, S., IDZITAC, I., (2016). Fuzzy TOPSIS: A General View, *Procedia Computer Science*, 91, pp. 823–831.

OJA, H., (1983). Descriptive statistics for multivariate distributions, *Statistics and Probability Letters*, 1, pp. 327–332.

RONKAINEN, T., OJA, H., ORPONEN, P., (2002). Computation of the multivariate Oja median, [In:] Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P. J. (eds.), Developments in Robust Statistics, Springer, Heidelberg, pp. 344–359.

TALEIZADEH, A. A., NIAKI, S. T. A., ARYANEZHAD, M-B., (2009). A hybrid method of Pareto, TOPSIS and genetic algorithm to optimize multi-product multi-constraint inventory control systems with random fuzzy replenishments, *Mathematical and Computer Modelling*, 49(5-6), pp. 1044–1057.

VELASQUEZ, M., HESTER, P. T., (2013). An Analysis of Multi-Criteria Decision Making Methods, *International Journal of Operations Research*, 10(2), pp. 56–66.

WANG, P., ZHU, Z., WANG, Y., (2016). A novel hybrid MCDM model combining the SAW, TOPSIS and GRA methods based on experimental design, *Information Sciences*, 345, pp. 27–45.

WEBER, A., (1909). Über den Standort der Industrien, Tubingen.

WUERTZ, D., SETZ, T., CHALABI, Y., (2017). Package 'fExtremes'. Rmetrics – Modelling Extreme Events in Finance, Version 3042.82, [online] Available at: https://cran.r-project.org/web//packages/fExtremes/ fExtremes.pdf.

WYSOCKI, F., (2010). Metody taksonomiczne w rozpoznawaniu typów ekonomicznych rolnictwa i obszarów wiejskich (Taxonomic methods in recognizing economic types of agriculture and rural areas), Poznań University of Life Sciences, Poznan.

WYSOCKI, F., ŁUCZAK, A., (2009). An evaluation of the social and economic development of powiats in the Wielkopolskie province using a fuzzy multi-criteria decision making (FMCDM) method, [In:]: Adamus, W. (ed.), The Analytic Hierarchy & Network. Application in Solving Multicriteria Decision Problems, Jagiellonian University Press, Cracow, pp. 319–329.

YOON, K., (1987). A reconciliation among discrete compromise situations, *Journal of Operational Research Society*, 38, pp. 277–286.

ZAVADSKAS, E. K., MARDANI, A., TURSKIS, Z., JUSOH, A., NOR, K. MD., (2016). Development of TOPSIS Method to Solve Complicated Decision-Making Problems: An Overview on Developments from 2000 to 2015, *International Journal of Information Technology & Decision Making*, 15(03), pp. 645–682.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 173–187, DOI 10.21307/stattrans-2020-019
Submitted – 04.03.2019; accepted for publishing – 31.01.2020

Application of iterated filtering to stochastic volatility models based on non-Gaussian Ornstein-Uhlenbeck process

Piotr Szczepocki [1]

ABSTRACT

Barndorff-Nielsen and Shephard (2001) proposed a class of stochastic volatility models in which the volatility follows the Ornstein–Uhlenbeck process driven by a positive Levy process without the Gaussian component. The parameter estimation of these models is challenging because the likelihood function is not available in a closed-form expression. A large number of estimation techniques have been proposed, mainly based on Bayesian inference. The main aim of the paper is to present an application of iterated filtering for parameter estimation of such models. Iterated filtering is a method for maximum likelihood inference based on a series of filtering operations, which provide a sequence of parameter estimates that converges to the maximum likelihood estimate. An application to S&P500 index data shows the model perform well and diagnostic plots for iterated filtering ensure convergence iterated filtering to maximum likelihood estimates. Empirical application is accompanied by a simulation study that confirms the validity of the approach in the case of Barndorff-Nielsen and Shephard's stochastic volatility models.

Key words: Ornstein–Uhlenbeck process, stochastic volatility, iterated filtering.

1. Introduction

Barndorff-Nielsen and Shephard (2001) proposed a continuous-time stochastic volatility model (BN-S model), in which the logarithm of the asset price $y(t)$ is assumed to be the solution of the following stochastic differential equation:

$$dy(t) = \left(\mu + \beta\sigma^2(t)\right)dt + \sigma(t)dB(t), \tag{1}$$

where $\left(B(t)\right)_{t\geq 0}$ is the Brownian motion, $\mu \in R_+$ is the drift parameter, $\beta \in R_+$ is the risk premium. Latent instantaneous volatility process $\left(\sigma^2(t)\right)_{t\geq 0}$ is determined by the stochastic differential equation

$$d\sigma^2(t) = -\lambda\sigma^2(t)dt + dz(\lambda t), \tag{2}$$

[1] Department of Statistical Methods/Institute of Statistics and Demography/Faculty of Economics and Sociology/ University of Lodz, Poland. E-mail: piotr.szczepocki@uni.lodz.pl. ORCID: https://orcid.org/0000-0001-8377-3831.

where $\lambda \in R_+$ and $\left(z(\lambda t)\right)_{t\geq 0}$ is pure jump Lévy process with stationary, independent and positive increments, and $z(0) = 0$. The process $\left(z(\lambda t)\right)_{t\geq 0}$ is called *Background Driving Lévy Process* (BDPL) of the process $\left(\sigma^2(t)\right)_{t\geq 0}$. Figure 1 presents examples of the pair of the processes $\left(z(\lambda t)\right)_{t\geq 0}$ and $\left(\sigma^2(t)\right)_{t\geq 0}$. There are several important features of such a stochastic volatility process defined by (2), some of which will be outlined in Section 2 on the basis of a series of Barndorff-Nielsen and Shephard papers (Barndorff-Nielsen and Shephard, 2001, 2002, 2003).

A great number of estimation techniques have been proposed to estimate BN-S model. In their introductory paper (Barndorff-Nielsen and Shephard, 2001), Barndorff-Nielsen and Shephard employed a nonlinear least squares estimation and suggested other possible methods: Bayesian inference, quasi-likelihood inference by means of Kalman filter (for more details of Kalman filter implemented for BN-S model, see Szczepocki (2018)), estimation equations (Sørensen, 2000) and indirect estimation (Gourieroux, Monfort and Renault, 1993). In the following years much work on estimation was devoted to the Bayesian Markov Chain Monte Carlo (MCMC) approach: Roberts *et al.* (2004), Griffin and Steel (2006, 2010), Gander and Stephens (2007a,b), Frühwirth-Schnatter and Sögner (2009). Hubalek and Posedel (2006, 2011) proposed an estimator based on martingale estimating functions. Taufer, Leonenko and Bee (2011) introduced a characteristic function-based estimation method. Raknerud and Skare (2011) implemented an indirect inference method based on approximate Gaussian state space representation. Andrieu *et al.* (2010) proposed to use Particle Markov Chain Monte Carlo (PMCMC) estimation method, which combines particle filter with Bayesian inference. Chopin *et al.* (2013) proposed SMC^2 algorithm, which substantially extends PMCMC. James *et al.* (2018) also used PMCMC for *OU-Gamma Time Change* version of BN-S model.

In this paper we propose estimation based on iterated filtering. It is relatively a new class of methods for maximum likelihood inference introduced by Ionides *et al.* (2006) and substantially modified by Ionides *et al.* (2015). It is based on a series of filtering operations which provide a sequence of parameter estimates that converges to the maximum likelihood estimate. In the discussion on (Andrieu *et al.*, 2010) Anindya Bhadra (one of co-authors of Ionides *et al.*, 2011) showed some results from applying the iterated filtering to a single example of BN-S model. However, he applied the initial version of iterated filtering (IF1) from Ionides *et al.* (2006). In this paper we employed the second generation version of iterated filtering (IF2) from Ionides *et al.* (2015).

The paper is organized as follows. Section 2 presents background material on Barndorff-Nielsen and Shephard stochastic volatility model. Section 3 presents iterated filtering. Section 4 contains simulation results on estimation and Section 5 applications to real data. Section 6 gives concluding remarks.

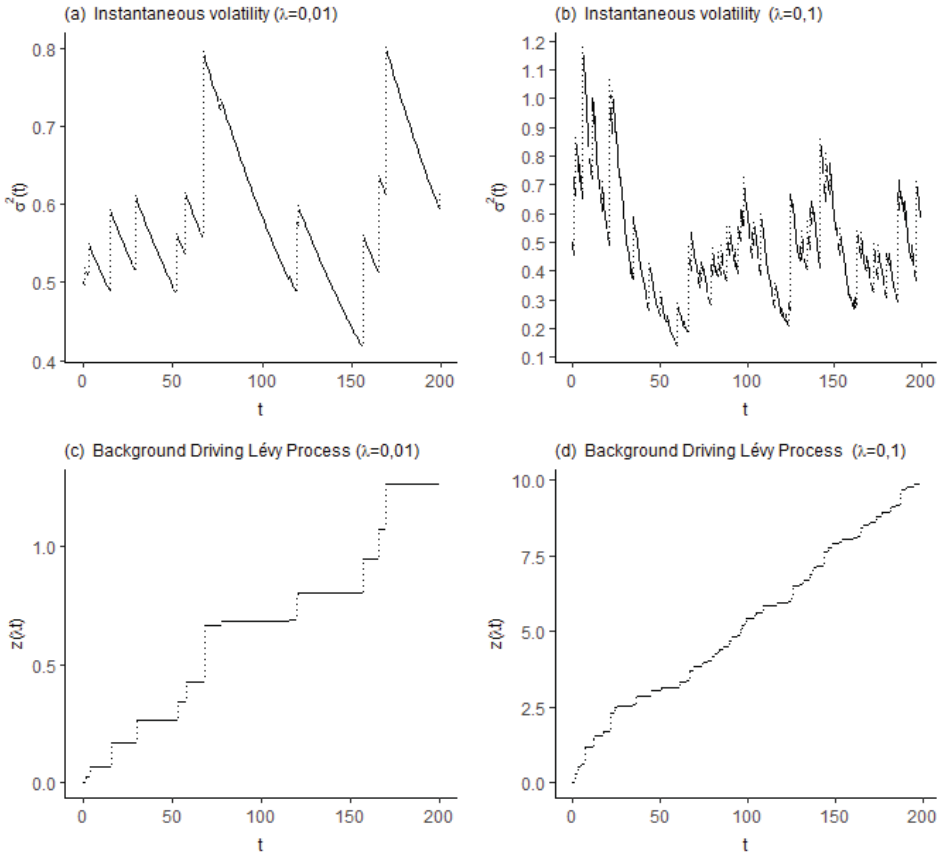


Figure 1. Two simulations of instantaneous volatility process with Gamma marginal (a) and (b), and corresponding Background Driving Lévy Process (c) and (d)

Source: Own work using R software.

2. Barndorff-Nielsen and Shephard stochastic volatility model

BN-S model has several important features which makes it very important for financial modelling. Firstly, instantaneous volatility $\left(\sigma^2(t)\right)_{t\geq 0}$ moves up by jumps according to $\left(z(\lambda t)\right)_{t\geq 0}$ and tails off exponentially at the rate λ. Thus, memory of the volatility process depends strictly on the rate λ. High values of λ result in high jumps, which are quickly discounted. On the contrary, a small value leads to a small jump but the process tails off slowly. Figure 1 shows the impact of λ on the volatility process.

Secondly, the time index of the process $\left(z(\lambda t)\right)_{t\geq 0}$ in (2) is chosen deliberately so that marginal distribution of $\sigma^2(t)$ does not depend on λ. Barndorff-Nielsen and Shephard (2001) proved that for any one-dimensional self-decomposable distribution

D there is a stationary Ornstein-Uhlenbeck process $\left(\sigma^2(t)\right)_{t\geq 0}$ and Lévy process $\left(z(\lambda t)\right)_{t\geq 0}$ satisfying equation (2), for which marginal distribution of $\sigma^2(t)$ is D. The class of self-decomposable distribution includes many distributions important in financial econometrics: gamma, normal-inverse Gaussian, inverse Gaussian, tempered stable, variance gamma, symmetric gamma, the Euler's gamma, Mexiner. (Schoutens, 2003) is a comprehensive reference text on financial application of self-decomposable distributions.

Thirdly, although instantaneous volatility $\left(\sigma^2(t)\right)_{t\geq 0}$ has discontinuous paths (due to jumps), integrated volatility

$$\sigma^{2*}(t) = \int_0^t \sigma^2(u)du \tag{3}$$

has continuous paths. Consequently, the resulting process of the logarithm of the asset price $y(t)$ also has continuous paths. One advantage of stochastic volatility of Ornstein-Uhlenbeck type is that many important process characteristics are analytically tractable. For example, integrated volatility has a simple structure

$$\sigma^{2*}(t) = \frac{1}{\lambda}\left(z(\lambda t) - \sigma^2(t) + \sigma^2(0)\right). \tag{4}$$

Finally, the implication of the formula (1) is that log-returns observed at time $n=1,\ldots, T$ (we assume that time differences $\Delta_n = t_n - t_{n-1}$ are fixed and equal Δ) take the form:

$$y_n = \int_{(n-1)\Delta}^{n\Delta} dy(t) = y(n\Delta) - y((n-1)\Delta) \tag{5}$$

and have conditional Normal distribution

$$y_n = N(\mu\Delta + \beta\sigma_n, \sigma_n^2) \tag{6}$$

where $\sigma_n^2 = \sigma^{2*}(n\Delta) - \sigma^{2*}((n-1)\Delta)$. This discretely observed volatility σ_n^2 ($n=1,\ldots, T$) was called actual volatility by Barndorff-Nielsen and Shephard (2001). Marginal distribution of y_n is a location scale mixture of normals. Thus, returns may capture empirical facts such as skewness and thick tails. Moreover, when $\Delta \to +\infty$ marginal distribution of y_n tends to normal distribution. Hence, non-normality of returns vanishes under temporal aggregation, which is another empirical fact observed in financial data.

BN-S model has attracted much interest and research in mathematical finance and financial econometrics. Nicolato and Venardos (2003) studied equivalent martingale measures and provided closed-form prices for European call options for BN-S model. The minimal entropy martingale measure and numerical option pricing for BN-S

model are investigated in (Benth and Karlsen, 2005) and (Benth and Meyer-Brandis, 2005). Hubalek and Sgarra (2009) provided option pricing by Esscher transform. Benth *et al.* (2003) considered Merton's portfolio optimization problem in a Black and Scholes market with stochastic volatility of BN-S type. Benth *et al.* (2007) provided explicit evaluation of the variance swaps. Hubalek and Sgarra (2011) developed a semiexplicit valuation formula for geometric Asian options.

3. Iterated filtering

3.1. General remarks

Iterated filtering (Ionides *et al.* 2006, 2015) are methods for maximum likelihood inference for state space models (SSMs). These models are also known as partially observed Markov Processes (POMP) or hidden Markov models (HMMs). SSMs consist of a pair of processes: (X_n, Y_n). The former is a Markov process (*state process*) which is not observed directly but may be estimated through the latter (*observation processes*). The observations of Y_n are assumed to be conditionally independent given the X_n (for details, see Durbin And Koopman, 2012). SSMs are very flexible and have been widely applied in economics, medicine, biology, mechanical system monitoring, patter recognition (see Chapter 1 in Cappé *et al.,* 2008 for examples). However, estimation for SSMs is very challenging because likelihood functions are analytically intractable in general.

Iterated filtering is one of the few if not the only available *likelihood-based* (based on the likelihood function for the full data), *simulation*-based (dynamics of the model is captured only via a simulator), frequentist (based on frequency interpretation of probability) methods for SSMs. Iterated filtering has been successfully applied to perform parameter estimation in SSMs, mostly in the context of biological applications (King et al., 2008, He et al., 2009, Bhadra et al., 2011) but also in economic modelling (Bretó, 2014).

The key idea behind iterated filtering is to replace the model we are interested in, which have constant parameters, with a similar model but with parameters that take a random walk in time. This extra variability smooths the likelihood surface and counteracts particle depletion. Over multiple repetitions of the filtering procedure (made by means of a particle filter), the variance of this random walk goes to zero and the augmented model approaches the original one. As an output of iterated filtering, the algorithm provides a sequence of updated parameter estimates that converge to the maximum likelihood estimate. Iterated filtering algorithms use basic sequential Monte Carlo techniques (also known as bootstrap particle filter, Gordon et al., 1993). Thus, they have the property that they do not need to evaluate the transition density of the

latent Markov process. Algorithms with this property have been called plug-and-play (Ionides et al., 2006) or simulation-based. It is vitally important in the case of BN-S model, for which the transition density takes no explicit form. The plug-and-play methodology is relatively recent and have been developing rapidly because of its less restrictive requirements. Examples of plug-and-play methodologies that follow the Bayesian paradigm are Approximate Bayesian Computation (Toni et al., 2009) and Particle Markov Chain Monte Carlo (Andrieu et al., 2010).

There are two generations of iterated filtering which are typically abbreviated by IF1 and IF2. The first was introduced by Ionides et al. (2006) and theoretically justified by Ionides et al. (2011). Later, Lindström et al. (2012) improved numerical performance of IF1 and Doucet et al. (2013) expanded it to include smoothing algorithm. The second generation was initiated by Ionides et al. (2015) and later supported by theoretical study of Nguyen (2016). Although both generations of iterated filtering recursively perform filtering through the augmented model, the theoretical justifications of these algorithms are essentially different. IF1 approximates the Fisher score function, whereas IF2 combines the idea of data cloning (Lele et al., 2007), with convergence of an iterated Bayes map (Nguyen, 2016). Ionides et al. (2015) showed that IF2 outperforms IF1 in empirical examples.

Convergence of iterated filtering IF2 to the maximum likelihood estimate has been shown under some regularity conditions (see Ionides et al., 2015 and Nguyen, 2016, for details). The conditions are rather technical so, in practical applications, convergence of algorithm should be assessed via diagnostic plots (Bretó, 2014).

In this paper, we use implementation of iterated filtering provided by the software package POMP (King *et al.*, 2010) written for the R statistical computing environment (R Development Core Team, 2010).

3.2. Implementation of the BN-S model

Barndorff-Nielsen and Shephard (2001) presented their model in state space model representation with y_n as an observation process and actual volatility as a state process. Conditional distribution of observation process given the state process $y_n \mid \sigma_n^2$ is given by the formula (6). The transition density is not available in explicit form. Griffin and Steel (2007) showed that the actual volatility can be written as

$$\sigma_n^2 = \frac{1}{\lambda}\left[\eta_{n,2} - \eta_{n,1} + \left(1 - e^{-\lambda\Delta}\right)\sigma^2\left((n-1)\Delta\right)\right], \tag{7}$$

where

$$\eta_n = \begin{bmatrix} e^{-\lambda\Delta}\int_0^{\Delta} e^{\lambda t}\, dz(\lambda t) \\ \int_0^{\Delta} dz(\lambda t) \end{bmatrix} \tag{8}$$

is a vector of random jumps, which is a pair of stochastic integrals with respect to the BDLP $\left(z(\lambda t)\right)_{t\geq 0}$. The instantaneous volatility process from equation (7) may be discretized by recursion

$$\sigma^2(n\Delta) = \sigma^2\left((n-1)\Delta\right)e^{-\lambda\Delta} + \eta_{n,1}. \tag{9}$$

In this paper, we use the series representation from Barndorff-Nielsen and Shephard (2001) given by

$$\eta_n = \begin{bmatrix} e^{-\lambda\Delta}\sum_{j=1}^{\infty} W^{-1}\left(\frac{a_{i,j}}{\lambda\Delta}\right)e^{-\lambda\Delta} \\ \sum_{j=1}^{\infty} W^{-1}\left(\frac{a_{i,j}}{\lambda\Delta}\right) \end{bmatrix} \tag{10}$$

where for each j ($j=1,2,\ldots$) $a_{i,j}$ are the arrival times of a Poisson process of intensity 1, and $r_{i,j} \sim$ U(0, 1), independent of the $a_{i,j}$. W^{-1} denotes the inverse of the tail mass function

$$W^{+}(x) = \int_x^{+\infty} w(y)dy, \tag{11}$$

where $w(y)$ is a density the Lévy measure of the Lévy-Khintchine representation for z(1) (see chapter 8 in Schoutens (2001) for detailed information of simulation techniques for Lévy processes). The only special case where the sums in (10) have only a finite number of non-zero terms is the gamma marginal distribution of instantaneous volatility. In other cases sums need to be truncated. In the case of the gamma distribution for instantaneous volatility process: $\sigma^2(t) \sim gamma(\nu, \alpha)$ ($\nu > 0$ is the scale parameter and α is the precision parameter) the inverse of the tail mass function W^{-1} takes the form (Barndorff-Nielsen and Shephard, 2001):

$$W^{-1}\left(\frac{a_{i,j}}{\lambda\Delta}\right) = \max\left\{0, \frac{1}{\alpha}\ln\left(\frac{\nu\lambda\Delta}{a_{i,j}}\right)\right\} \tag{12}$$

which is zero for $a_{i,j} \geq \nu\lambda\Delta$.

There is no agreement in the literature on how to choose a marginal distribution. In the rest of the paper we follow Roberts *et al.* (2004), Griffin and Steel (2006), Frühwirth-Schnatter and Sögner (2009), Raknerud and Skare (2011), Chopin *et al.* (2013) and use the gamma marginal distribution.

4. Simulation study

Since convergence of iterated filtering IF2 to the maximum likelihood estimates in the case of BN-S model is difficult to prove analytically, we checked the performance of the method in a simulation study. We considered 4 scenarios with different combinations of parameters. Values of the parameter were taken from Barndorff-Nielsen and Shephard (2002) and Creal (2008). We simulated 500 realizations of each scenario of length T=1000 observations. We run iterated filtering algorithm using J=100 and J=200 iteration with M=5000 particles. Table 1 presents mean errors (MEs) and mean standard errors (MSEs) obtained in the study. For the purpose of comparison, Table 1 reports also MEs and MSEs for the quasi-likelihood inference based on the Kalman filter. Thus, we set $\mu = \beta = 0$ and assessed precision only for volatility parameters: λ – the persistence parameter, ξ – the expected value of marginal distribution ($E(\sigma^2(t)) = \xi = \nu / \alpha$) and the standard deviation of marginal distribution ($\sqrt{Var(\sigma^2(t))} = \omega = \sqrt{\nu} / \alpha$).

Table 1. MEs and MSEs of the estimators

Parameters	KF		IF2 (J=100)		IF2 (J=200)	
	ME	MSE	ME	MSE	ME	MSE
$\lambda = 0.01$	0.066	0.261	0.021	0.163	0.013	0.121
$\xi = 0.5$	0.061	0.166	0.042	0.159	-0.032	0.143
$\omega = 0.25$	0.093	0.13	-0.046	0.186	-0.011	0.012
$\lambda = 0.05$	0.056	0.219	0.015	0.126	0.011	0.109
$\xi = 0.5$	-0.011	0.142	0.045	0.166	0.039	0.132
$\omega = \sqrt{0.4}$	0.072	0.146	0.051	0.232	-0.086	0.123
$\lambda = 0.1$	0.011	0.119	-0.005	0.086	0.019	0.021
$\xi = 0.5$	0.063	0.242	-0.021	0.166	-0.012	0.159
$\omega = 0.25$	0.091	0.246	0.051	0.131	0.013	0.011
$\lambda = 0.1$	0.013	0.145	0.009	0.026	0.019	0.021
$\xi = 0.5$	-0.051	0.171	-0.032	0.146	-0.022	0.169
$\omega = \sqrt{0.4}$	0.093	0.381	0.046	0.322	0.023	0.186

Source: Own work.

The results indicate that the proposed iterated filtering IF2 algorithm is quite reliable. For a smaller number of iterations J=100, the estimators seem to be biased but they become more precise as J increases. Both versions of IF2 outperform quasi-likelihood inference.

5. Empirical example

We estimate models by using Standard & Poor's 500 index (S&P500) daily data for the period 9.10.2012-30.09.2016. S&P500 index is one of the most important American stock market index. It is based on the market capitalizations of 500 large companies listed on the NYSE or NASDAQ. Data consist of 1001 closing values and 1000 log-returns. Table 2 and Figure 2 present data.

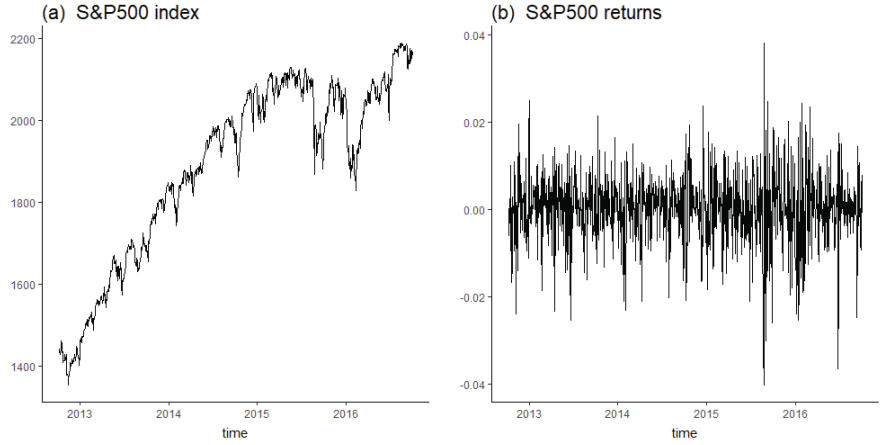


Figure 2. S&P500 daily index (a) and log-returns (b)

Source: Own work using R software.

Table 2. Descriptive statistics of S&P500 daily log-returns

Mean	Standard deviation	Skewness	Kurtosis	Quantiles		
				25%	50%	75%
0.0004	0.0083	-0.3830	5.0486	-0.0036	0.0005	0.0050

Source: Own work.

We run the iterated filtering algorithm with J=200 iteration. Each of iteration uses the bootstrap particle filter with M=5000 particles. Results of estimation are presented in Table 3. The drift parameter μ is close to zero. As may be expected from financial theory, the risk-premium coefficient β is positive. The estimated average actual volatility ξ and standard deviation ω correspond to gamma distribution with the scale

parameter ν=1.571 and the precision parameter α=14.124. Figure 3 presents diagnostic plots for iterated filtering. These plots suggest that the likelihood has in fact been maximized by iterated filtering in our analysis of log-returns of S&P500 index.

Table 3. Estimation results for the log-returns of the S&P500 index

Parameter	μ	β	λ	ξ	ω
Estimates	-0.001	0.051	0.026	0.111	0.089

Source: Own work.

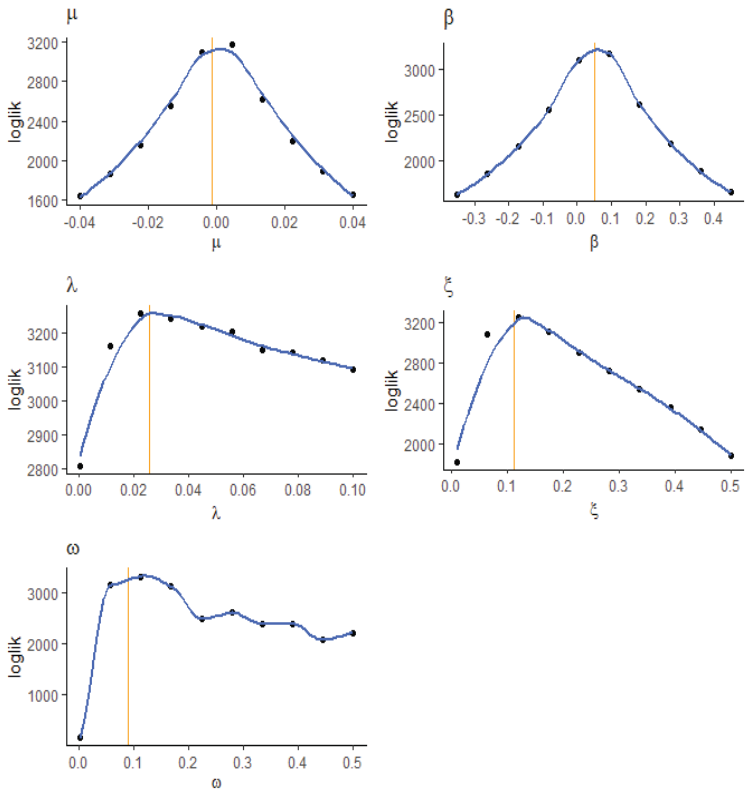


Figure 3. Diagnostic plots for iterated filtering: sliced likelihoods for the inquired parameters. For each plot the likelihood surface is explored along one of the parameters, keeping the other parameters fixed at the point which iterated filtering algorithm converges to. Points show the likelihood estimate obtained with 2,000 particles and the curves result from smoothing the likelihood evaluations with local quadratic regression. The vertical lines show iterated filtering estimates.

Source: Own work using R software.

6. Conclusions

In this article, we presented estimation of a class of stochastic volatility models where the volatility follows an Ornstein–Uhlenbeck process driven by a positive Lévy process via iterated filtration. This class of models, introduced by Barndorff-Nielsen and Shephard (2001), and therefore typically abbreviated to BN-S, has several important features, which aroused great interest in financial modelling for this class of stochastic volatility models.

From a theoretical point of view, the estimation method proposed in this article is convenient because it only requires to simulate the state process and to evaluate conditional density of the observation process given the simulated values of the state process. This feature, also known as plug-and-play property, is crucial for BN-S models, for which transition density is not available as a closed-form expression. Iterated filtration provides likelihood-based inference based on frequentist probability, which may be seen as competitive to plug-and-play methods that are based on Bayesian paradigm such as Particle Markov Chain Monte Carlo or Approximate Bayesian Computation. In this article, we exploited the second generation of iterated filtration IF2 introduced by Ionides *et al.* (2015), which outperforms the first generation IF1 in the rates of convergence to maximum likelihood estimates.

The results of the simulation study confirmed the validity of the approach in the case of BN-S model. In an application of the proposed method to S&P500 daily data, we presented, apart from estimates of parameters, also diagnostic plots for iterated filtering to ensure convergence to maximum likelihood estimates.

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.

ANDRIEU, C., DOUCET, A., HOLENSTEIN, R., (2010). Particle Markov Chain Monte Carlo methods, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 72, No. 3, pp. 269–342.

BHADRA, A., IONIDES, E. L., LANERI, K., PASCUAL, M., BOUMA, M., DHIMAN, R. C., (2011). Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise, *Journal of the American Statistical Association*, Vol. 106, No. 494, pp. 440–451.

BARNDORFF-NIELSEN, O.E. SHEPHARD, N., (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics, *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), Vol. 63, No. 2, pp. 167–241.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N., (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society: Series B (Statistical Method*ology), Vol. 64, No. 2, pp. 253–280.

BARNDORFF-NIELSEN, O. E., SHEPHARD, N., (2003). Integrated OU processes and Non-Gaussian OU-based stochastic volatility models, *Scandinavian Journal of Statistics*, Vol. 30, No. 2, pp. 277–295.

BENTH, F. E., GROTH, M., KUFAKUNESU, R., (2007). Valuing Volatility and Variance Swaps for a Non-Gaussian Ornstein–Uhlenbeck Stochastic Volatility Model, *Applied Mathematical Finance*, 14(4), 347–363.

BENTH, F. E., KARLSEN, K. H., REIKVAM, K., (2003). Merton's portfolio optimization problem in a Black and Scholes market with non-Gaussian stochastic volatility of Ornstein-Uhlenbeck type, *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, Vol. 13, No. 2, pp. 215–244.

BENTH, F. E., KARLSEN, K. H., (2005). A PDE representation of the density of the minimal entropy martingale measure in stochastic volatility markets, *Stochastics an International Journal of probability and Stochastic Processes*, Vol. 77, No. 2, pp. 109–137

BENTH, F. E., MEYER-BRANDIS, T., (2005). The density process of the minimal entropy martingale measure in a stochastic volatility model with jumps, *Finance and Stochastics*, Vol. 9, No. 4, pp. 563–575.

BRETÓ, C., (2014). On idiosyncratic stochasticity of financial leverage effects, Statistics & Probability Letters, Vol. 91, pp. 20–26.

CAPPÉ, O., MOULINES, E., RYDÉN, T., (2009). Inference in Hidden Markov Models, *Springer Series in Statistics*, Springer.

CHOPIN, N., JACOB, P. E., PAPASPILIOPOULOS, O., (2013). SMC2: an efficient algorithm for sequential analysis of state space models, *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), Vol. 75, No. 3, pp. 397–426.

CREAL, D. D., (2008). Analysis of filtering and smoothing algorithms for Lévy-driven stochastic volatility models, *Computational Statistics & Data Analysis*, Vol. 52, No. 6, pp. 2863–2876.

DOUCET, A., JACOB, P. E., RUBENTHALER, S., (2013). Derivative-free estimation of the score vector and observed information matrix with application to state-space models, arXiv preprint, URL: https://arXiv:1304.5768 .

DURBIN, J., KOOPMAN, S. J., (2012). Time series analysis by state space methods (Vol. 38), Oxford University Press.

FRÜHWIRTH-SCHNATTER, S., SÖGNER, L., (2009). Bayesian estimation of stochastic volatility models based on OU processes with marginal Gamma law, Annals of the Institute of Statistical Mathematics, Vol. 61, No. 1, pp. 159–179.

GANDER, M. P. S., STEPHENS, D. A., (2007a). Stochastic volatility modelling in continuous time with general marginal distributions: Inference, prediction and model selection, *Journal of Statistical Planning and Inference*, Vol. 137, No. 10, pp. 3068–3081.

GANDER, M. P. S., STEPHENS, D. A., (2007b). Simulation and inference for stochastic volatility models driven by levy processes, Biometrika, Vol. 94, No. 3, pp. 627–646.

GORDON, N. J., SALMOND, D. J., SMITH, A. F., (1993, April). Novel approach to nonlinear/non-Gaussian Bayesian state estimation, IEE Proceedings F-radar and signal processing, Vol. 140, No. 2, pp. 107–113.

GOURIEROUX, C., MONFORT A., RENAULT E., (1993). Indirect inference, *Journal of Applied Econometrics*, Vol. 8, pp. 85–118.

GRIFFIN, J. E., STEEL, M. F. J., (2006), Inference with non-Gaussian Ornstein–Uhlenbeck processes for stochastic volatility, *Journal of Econometrics*, Vol. 134, No. 2, pp. 605–644.

GRIFFIN, J. E., STEEL, M .F. J., (2010). Bayesian inference with stochastic volatility models using continuous superpositions of non-Gaussian Ornstein–Uhlenbeck processes, *Computational Statistics & Data Analysis*, Vol. 54, No. 11, pp. 2594–2608.

HE, D., IONIDES, E. L., KING, A. A., (2009). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study, *Journal of the Royal Society Interface*, Vol. 7, No. 43, pp. 271–283.

HUBALEK, F., POSEDEL, P., (2006). Asymptotic analysis for an optimal estimating function for Barndorff-Nielsen-Shephard stochastic volatility models, Work in progress, URL: https://arxiv.org/abs/0807.3479 .

HUBALEK, F., POSEDEL, P., (2011). Joint analysis and estimation of stock prices and trading volume in Barndorff-Nielsen and Shephard stochastic volatility models, *Quantitative Finance*, Vol. 11, No. 6, pp. 917–932.

HUBALEK, F., SGARRA, C., (2009). On the Esscher transforms and other equivalent martingale measures for Barndorff-Nielsen and Shephard stochastic volatility models with jumps, *Stochastic Processes and their Applications*, Vol. 119, No. 7, pp. 2137–2157.

HUBALEK, F., SGARRA, C., (2011). On the explicit evaluation of the geometric Asian options in stochastic volatility models with jumps, Journal of Computational and Applied Mathematics, Vol. 235, No. 11, pp. 3355–3365.

IONIDES, E. L., BRETÓ, C., KING, A. A., (2006). Inference for nonlinear dynamical systems, Proceedings of the National Academy of Sciences, Vol. 103, No. 49, pp. 18438–18443.

IONIDES, E. L., BHADRA, A., ATCHADÉ, Y., KING, A., (2011). Iterated filtering, The Annals of Statistics, Vol. 39, No. 3, pp. 1776–1802.

IONIDES, E. L., NGUYEN, D., ATCHADÉ, Y., STOEV, S., KING, A. A., (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps, *Proceedings of the National Academy of Sciences*, Vol. 112, No. 3, pp. 719–724.

JAMES, L. F., MÜLLER, G., ZHANG, Z., (2018). Stochastic Volatility Models based on OU-Gamma time change: Theory and estimation, *Journal of Business & Economic Statistics*, Vol. 36, No. 1, pp. 75-87.

KING, A. A., IONIDES, E. L., PASCUAL, M., BOUMA, M. J., (2008). Inapparent infections and cholera dynamics, Nature, Vol. 454, No. 7206, pp. 877–880.

KING, A. A., IONIDES, E. L., BRETÓ, C., ELLNER, S., KENDALL, B., WEARING, H., FERRARI, M. J., LAVINE, M., REUMAN, D. C., (2010). POMP: statistical inference for partially observed Markov processes (R package), URL http://pomp.r-forge.r-rproject.org.

LELE, S. R., DENNIS, B., LUTSCHER, F., (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods, *Ecology letters*, Vol. 10, No. 7, pp. 551–563.

NICOLATO, E. VENARDOS, E., (2003). Option pricing in stochastic volatility models of the Ornstein-Uhlenbeck type, *Mathematical Finance*, Vol. 13, No. 4, pp. 445–466.

NGUYEN, D., (2016). Another look at Bayes map iterated filtering, Statistics & Probability Letters, Vol. 118, pp. 32–36.

R DEVELOPMENT CORE TEAM, (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org.

ROBERTS, G., PAPASPILIOPOULOS, O., DELLAPORTAS, P., (2004). Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 66, No. 2, pp. 369–393.

RAKNERUD, A., SKARE, Ø., (2012). Indirect inference methods for stochastic volatility models based on non-Gaussian Ornstein–Uhlenbeck processes, *Computational Statistics & Data Analysis*, Vol. 56, No. 11, pp. 3260–3275.

SØRENSEN, M., (2000). Prediction-based estimating functions. The Econometrics Journal, Vol. 3, No. 2, pp. 123–147.

SCHOUTENS, W., (2003). Lévy processes in finance. Wiley.

SZCZEPOCKI, P., (2018). Application of Kalman Filter to Stochastic Volatility Models of the Orstein-Uhlenbeck Type, Acta Universitatis Lodziensis. Folia Oeconomica, Vol. 337, No. 4, pp. 183–201.

TAUFER, E., LEONENKO, N., BEE, M., (2011). Characteristic function estimation of Ornstein–Uhlenbeck-based stochastic volatility models, Computational Statistics & Data Analysis, Vol. 55, No. 8, pp. 2525–2539.

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A., STUMPF, M. P., (2008). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems, *Journal of the Royal Society Interface*, Vol. 6, No. 31, pp. 187–202.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2, pp. 189–200, DOI 10.21307/stattrans-2020-020
Submitted – 26.09.2019; accepted for publishing – 25.03.2020

New linear model for optimal cluster size in cluster sampling

Alok Kumar Shukla[1], **Subhash Kumar Yadav**[2]

ABSTRACT

In this paper, a nonlinear model is proposed for improving the relationship between the size of a cluster and the variance within the cluster. This model describes the most appropriate functional relation between the within-cluster variance and the cluster size. Through this model, we can obtain the optimum size of a cluster and an estimate of the variance between clusters. The proposed model leads to further improvement in the estimation of the optimum size of a cluster, and the formula for the determination of optimum cluster size leads to explicit solution of models.

Key words: Non-linear models, optimum cluster size, four-parameter model, variance function.

1. Introduction

Regression analysis is widely used for better explanation and future prediction of any phenomenon which is assumed to develop in some patterns whether in economics or any other field. In cluster sampling, it is of interest to find the most suitable functional relationship between variance within the cluster (S_w^2) and the cluster size (M) for prediction [Singh and Chaudhary (2009)]. Smith (1938), Jessen (1942), Hansen and Hurwitz (1942), Mahalanobis (1940, 1942), Misra et al. (2010), Tiwari and Misra (2011), Shukla et al. (2013), Shukla and Yadav (2016), Lawson and Skinner (2017) etc., have discussed the problem of determining the optimum cluster size in two important contexts of variance function and cost function respectively. Scarneciu *et al.* (2017) compared various nonlinear models in determining pulmonary pressure in hyperthyroidism. Kaplan and Gurcan (2018) compared different growth curves using non-linear regression function in Japanese quail. Riazoshams *et al.* (2019) described in detail the robust nonlinear regression models with applications using R

[1] Department of Statistics, D. A-V. College, Kanpur-208001, U.P., India. E-mail: alokshukladav@gmail.com. ORCID: https://orcid.org/0000-0001-9797-7894.
[2] Corresponding Author, Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow-226025, U.P., India. E-mail: drskystats@gmail.com. ORCID: https://orcid.org/0000-0002-7181-8075.

software. All the functional relations given by the above authors are of similar functional form describing the relationship between the size of the cluster and variance within the cluster. It is well established in cluster sampling that sampling variance increases as the cluster size increases and it decreases with the number of clusters. The cost also decreases as cluster size decreases and increases as the number of clusters increases. Thus, it becomes important to seek a balancing point through the optimum size of the cluster and the number of clusters in the sample by minimizing the variance for a given/fixed cost or vice-versa.

Let $\bar{y}$ denote the sample mean for the characteristic y under study for the sample size n. We know that in cluster sampling, the variance of the sample mean $\bar{y}$ is given by

$$V(\bar{y}) = \frac{1-f}{n} S_b^2 \tag{1}$$

where $f = \frac{n}{N}$ is finite population correction and $S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i - \bar{\bar{Y}})^2$ is the variance between cluster means,

where $\bar{Y}_i = \frac{Y_i}{M}$ is the mean of the i^{th} cluster for the characteristic under study and $\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i = \frac{Y}{NM}$ is the population mean of NM units for N clusters each of size M with $Y = \sum_{i=1}^{N} Y_i$ as population total for the study variable.

Now, it is of crucial importance to know the behaviour of $V(\bar{y})$ with the cluster size M. This involves knowing the relationship between S_b^2 and M. Through analysis of variance (ANOVA) technique, S_b^2 can be found if we know

(i) The Total variance S^2 between all the NM elements in the population,

(ii) The variance S_w^2 within all M elements of the same cluster for all N clusters,

where S^2 and S_w^2 are respectively given by

$$S^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij} - \bar{\bar{Y}})^2 \text{ and } S_w^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{ij} - \bar{Y}_i)^2 .$$

Thus, the total variance S^2 of all population units can be written in the form of S_b^2 and S_w^2 as

$$S^2 = [M(N-1)S_b^2 + N(M-1)S_w^2]/(NM-1) \tag{2}$$

If N is large, we express the above equation (2) as

$$S^2 = S_b^2 + (M-1)S_w^2/M \tag{3}$$

Hence,

$$S_b^2 = S^2 - (M-1)S_w^2/M\,. \tag{4}$$

Thus, S_b^2 also depends on S_w^2.

We are considering the problem of determining the best relationship between variance function and the size of the cluster.

Jessen (1942) suggested the following relationship for S_w^2 and M by a non-linear form of model as

$$S_w^2 = \alpha M^{\beta},\ M > 1 \tag{5}$$

where α and β are the parameters of the above non linear model.

Misra *et al.* (2010) established the relationship between S_w^2 and M through an asymptotic regression model given as

$$S_w^2 = \alpha + \beta \rho^M,\ M > 1 \tag{6}$$

where α, β and ρ are the parameters of the asymptotic regression model.

Tiwari and Misra (2011) suggested a three-parameter linear regression model for the relationship between S_w^2 and M as

$$S_w^2 = \alpha + \beta M + \rho/M,\ M > 1 \tag{7}$$

where α, β and ρ are the parameters to be estimated for the above linear regression model.

2. Suggested model

In the present paper, we have proposed the following four-parameter model for the most appropriate relationship between S_w^2 and M as

$$S_w^2 = \alpha + \beta M + \delta M^2 + \rho/M,\ M > 1 \tag{8}$$

Expression (8) is a linear model in which its parameters are appearing linearly and there is no problem in assuming an additive error term in model (8).

The above model can be postulated as a statistical model as

$$S_w^2 = \alpha + \beta M + \delta M^2 + \rho / M + e_i,\ M > 1,\ i = 1,2,3,.....n \tag{9}$$

where the random variable $e_i's$ are assumed to be independently and identically normally distributed with mean zero and fixed variance σ^2 and α, β, δ and ρ are the parameters of the model (9).

2.1. Fitting of models

Draper and Smith (1998) have classified the above models in two groups; one as intrinsically linear and another as intrinsically non-linear models. The model (5) is an intrinsically linear model as it can be transformed into a linear model. Model (6) is purely nonlinear model as it cannot be transformed by means of any transformation into a linear model. The OLS method is not directly applied for estimating the parameters of model (6). The parameters of model (6) are estimated through the iterative procedure as Levenberg-Marquardt's method. Model (7) and the proposed model (8) are linear in parameters so their parameters are estimated by the method of least squares.

2.2. Goodness of fit of different models

Coefficient of Determination - R^2

The assessment of the regression model is to observe how much of the total sum of squares (TSS) has fallen into the sum of squares due to the regression (SSR).

$$R^2 = \frac{SSR}{TSS}$$

Adjusted Coefficient of Determination - R^2_{Adj}

Montgomery *et al.* (2012) have described R^2_{Adj} considering good for model comparison when the number of parameters is not equal in two models.

$$R^2_{Adj} = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R^2\right)$$

Residual Mean Square- s^2

The residual mean square is defined as

$$s^2 = \frac{SSE}{n-p}$$

where n is the number of observations and p is the number of the model parameters used and SSE is the sum of squares due to errors. A small value of s^2 reflects the appropriateness of the fitted model.

***Mean Absolute Error* (M.A.E)**

The Mean Absolute Error is defined as

$$M.A.E. = \frac{\sum_{i=1}^{n} |errors|}{n}$$

where n is the number of observations. A smaller $M.A.E.$ is preferred in fitting of various models.

***Akaike Information Criterion* (*A.I.C.*)**

Gujarati and Sangeetha (2007) have given a lot of importance to Akaike Information Criterion (A.I.C.), defined as,

$$A.I.C. = Exp\left(\frac{2p}{n}\right)\frac{RSS}{n}$$

where n is the number of observations and p is the number of parameters. RSS is residual or error sum of square.

2.3. Examination of Residuals

Analysis of the residuals (errors) is strongly recommended to decide about the suitability of a model by Draper and Smith (1998). Three important assumptions of the model are:

(i) Errors are not auto correlated.

(ii) Errors are independent.

(iii)Errors are normally distributed.

The assumptions can be verified by examining the residuals.

Test for auto correlation of errors (Durbin-Watson Test)

We test H_0 : Errors are not auto correlated (if DW test values > d_U)

Against H_1 : Errors are auto correlated (if DW test values < d_L)

where d_L and d_U are given in Draper and Smith (1998). DW Test values greater than 1.72 times d_U confirm that there is no problem of auto-correlation.

Test for independence of errors (Run Test)

We test H_0 : Errors are independent.

Against H_1 : Errors are not independent.

Test for normality (Shapiro-Wilk Test,* $n < 50$*)

We test H_0 : Errors are normally distributed.

Against H_1 : Errors are not normally distributed.

2.4. Determination of Variance Function

If the total population is considered as a single cluster containing *NM* elements, the S_w^2 will be equal to the total variance S^2. Thus, for the proposed model, we have

$$S^2 = \alpha + \beta NM + \delta (NM)^2 + \rho / NM \tag{10}$$

The between-cluster variance for the proposed model is obtained by putting (8) and (10) in (4) as,

$$\hat{S}_b^2 = [\hat{\alpha} + \hat{\beta} NM + \hat{\delta}(NM)^2 + \hat{\rho} / NM] - \frac{(M-1)}{M}[\hat{\alpha} + \hat{\beta} M + \hat{\delta} M^2 + \hat{\rho} / NM] \tag{11}$$

where $\hat{\alpha}$, $\hat{\beta}$, $\hat{\delta}$ and $\hat{\rho}$ are the estimated values of the parameters of the suggested model.

The variance of the sample mean of the characteristic under study through the suggested model in cluster sampling can be obtained as

$$V(\bar{y}) = \frac{1-f}{n} \hat{S}_b^2 \tag{12}$$

3. Empirical study

The appropriateness and model adequacy of various models have been examined by using two natural data sets from Sukhatme *et al.* (1984) and Govindarajulu (1999) respectively. The S_w^2 have been calculated for different sizes of the clusters in (acre)², with the study variable as the area under wheat crop. We have computed the estimated values of parameters, goodness of fit and residuals analysis for the models (5)-(8) given in Table-1(a) and Table-1(b). The Estimated values of S_w^2 and S_b^2 are given in Table-1(c) and Table-1(d). These values are given in Table-2(c) and Table-2(d) for the models (5)-(8) respectively. The above values have been obtained using SPSS 17.0 Statistical software.

Table-1(a). Parameter estimates for various models

Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\rho}$
Model (5)	78.886	0.0473	-	-
Model (6)	108.171	31.530	-	0.948
Model (7)	93.813	0.012	-	-32.888
Model (8)	88.1502	0.4272	-0.0003	-21.8678

Table-1(b). Goodness of fit of models and Residuals Analysis

	Model (5)	Model (6)	Model (7)	Model (8)
R^2	0.941	0.983	0.984	0.999
R^2_{Adj}	0.921	0.966	0.978	0.998
s^2	10.479	4.410	2.756	0.0248
M.A.E.	2.078	1.208	0.9	0.055
A.I.C.	13.6643	5.8573	3.6594	0.0245
DW#	1.3457	2.2104	2.3238	3.4161
R*	0.001 (1.000)	0.109 (0.913)	0.001 (1.000)	1.200 (0.230)
W^	0.9917 (0.9784)	0.8994 (0.4037)	0.9713 (0.8749)	0.9952 (0.9926)

is Durbin and Watson Test values, * is Run test values, ^ is Shapiro-Wilk test values, the p-values are given in parentheses.

Table-1(c). Estimated S_w^2 for various models

M	Observed value of S_w^2	Estimated value of S_w^2 for model (5)	Estimated value of S_w^2 for model (6)	Estimated value of S_w^2 for model (7)	Estimated value of S_w^2 for model (8)
2	78.10	81.53	79.84	77.39	78.0695
4	84.28	84.25	82.71	85.64	84.3868
8	88.92	87.05	87.60	89.80	88.8127
16	93.50	89.95	94.75	91.96	93.5308
NM =1176	108.33	110.22	108.17	108.34	108.33

Table-1(d). Estimated S_b^2 from equation (4) for various models

M	Observed value of S_b^2 from equation (4)	Estimated value of S_b^2 for model (5)	Estimated value of S_b^2 for model (6)	Estimated value of S_b^2 for model (7)	Estimated value of S_b^2 for model (8)
2	69.28	69.45	68.25	69.64	69.2952
4	45.12	47.03	46.13	44.11	45.0399
8	30.52	34.05	31.53	29.76	30.6188
16	20.69	25.89	19.34	22.12	20.6448

Table-2(a). Parameter estimates for various models

Model	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\delta}$	$\hat{\rho}$
Model (5)	0.072	0.182	-	-
Model (6)	0.372	-0.566	-	0.966
Model (7)	0.3163	0.0000061	-	-4.311
Model (8)	-0.3039	0.0132	-0.000001	2.3414

Table-2(b). Goodness of fit of models and Residuals Analysis

	Model (5)	Model (6)	Model (7)	Model (8)
R^2	0.768	0.982	0.959	0.996
R^2_{Adj}	0.710	0.971	0.932	0.991
s^2	0.0039	0.0004	0.0009	0.0001
M.A.E.	0.0433	0.0116	0.0183	0.0045
A.I.C.	0.0051	0.0004	0.0013	0.00006
DW#	0.9083	1.8057	1.6865	3.5122
R*	-1.369	0.001	0.001	1.369
	(0.171)	(1.000)	(1.000)	(0.171)
W^	0.9648	0.9577	0.9447	0.9759
	(0.8577)	(0.8127)	(0.7175)	(0.9217)

is Durbin and Watson Test values, * is Run test values, ^ is Shapiro-Wilk test values, the p-values are given in parentheses.

Table-2(c). Estimated S_w^2 for various models

M	Observed value of S_w^2	Estimated value of S_w^2 for model (5)	Estimated value of S_w^2 for model (6)	Estimated value of S_w^2 for model (7)	Estimated value of S_w^2 for model (8)
15	0.05	0.1176	0.0361	0.0289	0.0501
20	0.08	0.1239	0.0898	0.1008	0.0770
25	0.11	0.1290	0.1349	0.1440	0.1193
30	0.18	0.1334	0.1728	0.1727	0.1694
35	0.22	0.1372	0.2046	0.1933	0.2239
NM =8820	0.37	0.3747	0.3717	0.3727	0.3699

Table-2(d). Estimated S_b^2 from equation (4) for various models

M	Observed value of S_b^2 from equation (4)	Estimated value of S_b^2 for model (5)	Estimated value of S_b^2 for model (6)	Estimated value of S_b^2 for model (7)	Estimated value of S_b^2 for model (8)
15	0.320	0.265	0.338	0.346	0.3231
20	0.294	0.257	0.286	0.277	0.2967
25	0.264	0.251	0.242	0.234	0.2553
30	0.196	0.246	0.205	0.206	0.2061
35	0.156	0.241	0.173	0.185	0.1523

4. Results and discussion

From Table-1(b) and Table-2(b), it is easily evident that the value of R^2 for the competing models ranges from [0.941 0.984] and [0.768 0.982] respectively for Data Set-1 and Data Set-2, while that for the suggested model is 0.999 and 0.996 respectively. The value of R_{Adj}^2 for the competing models lies between [0.921 0.978] and [0.710 0.971] respectively while that for the suggested model between 0.998 and 0.991 respectively. The values of the s^2 for the competing models range from [2.756 10.479] and [0.0039 0.0004] while for the proposed model are 0.0248 and 0.0001 for Data Set-1 and Data Set-2 respectively. The values of M.A.E. are between [0.9 2.208] and [0.0183 0.0433] for the models in comparison while for the suggested models they are 0.055 and 0.0045 for Data Set-1 and Data Set-2 respectively. The values of A.I.C. lie between [3.6594 13.6643] and [0.0004 0.0051] for the competing models while these of proposed models are 0.0245 and 0.00006 for Data Set-1 and Data Set-2 respectively. Other measures are also better for the suggested model as compared to competing models.

Figure-1 and Figure-2 show the graph of R^2 and R_{Adj}^2 and s^2, M.A.E. and A.I.C. for the suggested and the competing models respectively for Data Set-1 while Figure-3 and Figure-4 for Data Set-2.

Figure-1. R^2 and R^2_{Adj} for data set-1 for various models

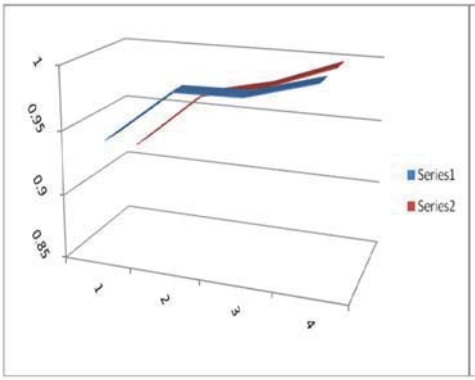


Figure-2. s^2, MAE and AIC for data set-1 for various models

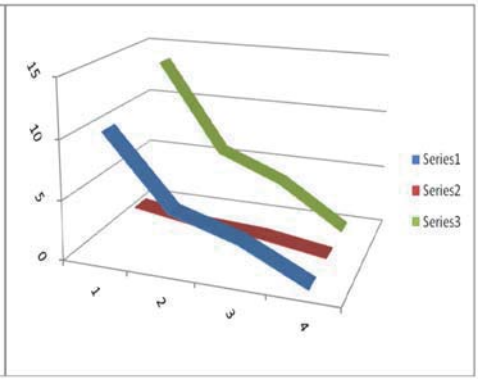


Figure-3. R^2 and R^2_{Adj} for data set-2 for various models

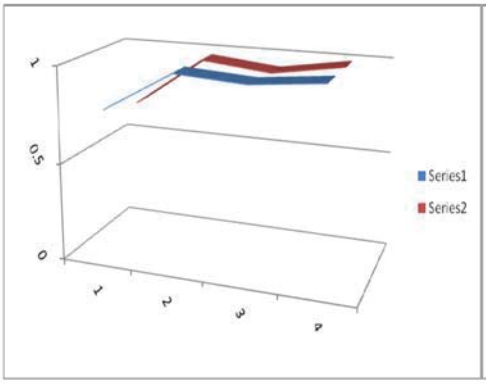
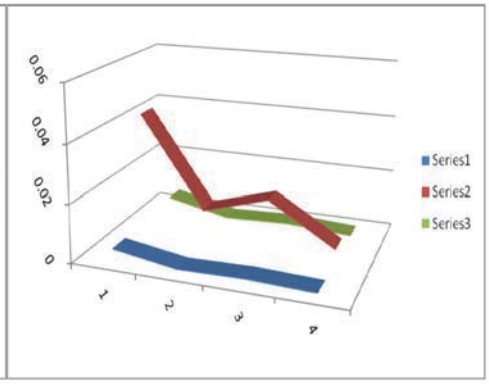


Figure-4. s^2, MAE and AIC for data set-2 for various models



5. Conclusion

In the present manuscript, we have proposed a four-parameter linear regression model for enhanced estimation of the variance function for clustered data. The parameters of the proposed model have been estimated through a well-known method of least squares. The proposed model and many other linear and nonlinear models have been fitted for the real data sets. The suggested model is compared with the competing linear and non-linear models. It has been shown that the proposed model fits well in comparison with other models for variance function in cluster sampling as it has

lesser mean residual error and other good measures of adequacy. Thus, it is recommended to use the proposed model for improved estimation of variance function, the relation between within-cluster variance and cluster size, between cluster variance and the cluster size in cluster sampling.

Acknowledgment

The authors expresses their heartfelt gratitude to the Editor and Editior-in-Chief along with the learned referees for critically examining the manuscript which improved the earlier draft.

REFERENCES

DRAPER, N. R., SMITH, H., (1998). Applied regression analysis. 3rd Ed., John Wiley & Sons,.

GUJARATI, D. N., SANGEETHA, (2007). Basic Econometrics, 4th Ed, Tata McGraw-Hill.

HANSEN, M. H., HURWITZ, W. N., (1942). Relative efficiencies of various sampling units in population enquiries, *Journal of American Statistical Association*, 37, pp. 89–94.

JESSEN, R. J., (1942 (). Statistical investigation of sample survey for obtaining farm facts, *Iowa Agricultural Experiment Station, Research Bulletin*, p. 304.

KAPLAN, S., GURCAN, E. K., (2018). Comparison of growth curves using non-linear regression function in Japanese quail, *Journal of Applied Animal Research*, 46, 1, pp. 112–117.

LAWSON, N., SKINNER, C., (2017). Estimation of a cluster-level regression model under nonresponse within clusters, *Metron*, 75, pp. 319–331.

MAHALANOBIS, P. C., (1940). A sample survey of acreage under jute in Bengal, *Sankhya*, 4, pp. 511–530.

MAHALANOBIS, P. C., (1942). General report on the sample census of area under jute in Bengal, *Indian Central Jute Committee*.

MISRA, G. C., YADAV, S. K., SHUKLA, A. K., RAJ, B., (2010). Use of a non-linear model for improved estimation in cluster sampling, *Journal of Reliability and Statistical Studies*, 3, 2, pp. 73–78.

MONTGOMERY D. C., PECK E. A., VINING G. C., (2012). Introduction to Linear Regression Analysis, 5th Ed., Wiley.

RIAZOSHAMS, H., MIDI, H., GHILAGABER, G., (2019). Robust Nonlinear Regression: with Applications using R, 1st Ed, Wiley.

SCARNECIU, C. C., SANGEORZAN, L., RUS, H., SCARNECIU, V. D., VARCIU, M. S., ANDREESCU, O., SCARNECIU, I., (2017). Comparison of linear and non-linear regression analysis to determine Pulmonary Pressure in Hyperthyroidism, *Pakistan Journal of Medical Sciences*, 33, 1, pp. 111–120.

SINGH, D., CHAUDHARY, F. S., (2009). Theory and Analysis of Sample Survey Designs, New Age International.

SMITH, H. F., (1938). An empirical law describing heterogeneity in the yields of agricultural crops, *Journal of Agricultural Science*, 28, pp. 1–23.

SUKHATME, P. V., SUKHATME, B. V., SUKHATME, S., ASOK, C., (1984). Sampling theory of surveys with applications, *Indian Society of Agricultural Statistics*.

SHUKLA, A.K., YADAV, S. K., (2016). Asymptotic Non-Linear Models for Uniformity Trial Experiments, *Elixir International Journal,* 94, 1, pp. 40042–40044.

SHUKLA, A. K., YADAV, S. K., MISRA, G. C., (2013). A Linear Model for Uniformity Trial Experiments, *Statistics in Transition-new series*, 14, 1, pp. 161–170.

TIWARI, R. B., MISRA, G. C., (2011). Estimation of optimum cluster size, IFRSA's *International Journal of Computing*, l, 4, pp. 717–722.

ZAKKULA G., (1999). Elements of Sampling Theory and Methods, Printice Hall Publication.

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2 pp. 201–207, DOI 10.21307/stattrans-2020-021

Report

The XXXVIII Conference on Multivariate Statistical Analysis 4–6 November 2019, Łódź, Poland

The 38th edition of the International Conference on Multivariate Statistical Analysis (MSA) was held in Łódź, Poland, on November 4-6, 2019. The MSA conference was organized by the Department of Statistical Methods of the University of Łódź, the Institute of Statistics and Demography of the University of Łódź, the Polish Statistical Association, Branch in Łódź and the Committee on Statistics and Econometrics of the Polish Academy of Sciences. The conference organization was financially supported by the National Bank of Poland, the Polish Academy of Sciences and the Ministry of Science and Higher Education[1].

The Scientific Committee was headed by Professor Czesław Domański and the Organizing Committee consisted of members: Aleksandra Baszczyńska, Assistant Professor from the Department of Statistical Methods of the University of Łódź and Katarzyna Bolonek-Lasoń, Assistant Professor from the Department of Statistical Methods of the University of Łódź.

The Multivariate Statistical Analysis conference constituted a forum for discussion and exchanging opinions about development of statistics. Participants presented the latest theoretical achievements in the field of the multivariate statistical analysis, its practical aspects and applications. The scientific programme covered a wide range topics of statistical mathematics and multivariate statistical methods including multivariate distributions, statistical tests, nonparametric inference, factor analysis, cluster analysis, discrimination analysis, Bayesian methods, stochastic analysis and application of statistical methods in finance, economy, capital market and risk management.

The conference was attended by 72 participants from many academic centres in Poland (Gdańsk, Katowice, Kraków, Łódź, Poznań, Rzeszów, Szczecin, Warszawa, Wrocław) and from abroad (Germany, Italy). Representatives of Statistics Poland and Statistical Office in Łódź, Statistical Office in Poznań and Statistical Office in Rzeszów

[1] Organization of the international conference "Multivariate Statistical Analysis 2019 (MSA 2019)" – task financed under contract 712 / P-DUN/202019 from the funds of the Minister of Science and Higher Education allocated to the dissemination of science.

were also participants of the 2019 MSA conference. In 15 sessions (plenary and parallel) 42 papers were presented including 4 invited lectures.

The conference was opened by the Head of the Scientific Committee, Professor Czesław Domański. The subsequent speakers at the conference opening were Professor Antoni Różalski, Rector of the University of Łódź, and Professor Michał Przybyliński, the Vice Dean – Education and Student Affairs of the Faculty of Economics and Sociology of the University of Łódź.

The chairman of the session Czesław Domański opened the first plenary session with two papers. The first one was an invited lecture entitled "Optimal sample allocation in stratified sampling schemes – linear algebra methods and algorithms" and it was presented by Professor Jacek Wesołowski (Statistics Poland, Warsaw University of Technology). The description how methods of linear algebra (eigenvectors and eigenvalues of a population based matrix) can be used in order to determine such sample allocations in stratified schemes which are domains-wise optimal was presented. The second paper "Kernel discriminant coordinates in the case of geographically weighted temporal-spatial data with variable selection" presented by Professor Mirosław Krzyśko (Adam Mickiewicz University Poznań) where the extension of a method developed by Mika et al. (1999) as well as Baudat and Anouar (2000) in kernel discriminant coordinates analysis for fixed vector data is used.

The second session (chairman Professor Mirosław Krzyśko) was a historical one with papers devoted to two important statisticians: Jakub Kazimierz Haur (a paper presented by Professor Czesław Domański) and Marcin Kromer (a paper presented by Professor Jerzy T. Kowaleski, University of Łódź).

In the third reminiscent session (chairman Professor Czesław Domański), the conference participants recalled outstanding statisticians who died last year. Professor Krystyna Katulska was commemorated by Professor Mirosław Krzyśko, Professor Mirosław Krzysztofiak was commemorated by Ewa Wycinka (University of Gdańsk), Professor Józef Kolonko by Professor Janusz Wywiał (University of Economics in Katowice), Professor Stanisław Wydmus and Professor Michał Major by Professor Stanisław Wanat (Cracow University of Economics).

During the conference other invited lectures were presented:

- "Selected aspects of households' well-being measurement" by Professor Józef Dziechciarz (Wroclaw University of Economics and Business), where an attempt to review problems and methodological proposals for measuring households' well-being was presented.
- "Advances in learning from contaminated datasets" by Professor Francesca Greselin (University of Milan-Bicocca, Italy), with an introduction into a robust

and adaptive version of the Discriminant Analysis rule, capable of handling situations in which one or more of the afore-mentioned problems occur.

- "A new virtual library containing interactive learning objects for statistics education" by Professor Hans-Joachim Mittag (University of Hagen, Germany), with presentation of project activities aiming at developing interactive learning objects for statistics education.

Papers presenting the latest theoretical achievements in the field of the multivariate statistical analysis are the following:

- Andrzej Bąk, "Methods of imputation of missing data using the R program on the example of the Local Data Bank", with results of attempts to apply supplementing missing data using methods proposed in the literature and packages of the R program.
- Katarzyna Budny, "Multivariate Chebyshev's inequality – some bounds on the probability of a random vector taking values in the Euclidean ball", where some multivariate generalizations of Chebyshev's inequality with the bounds on the probability of a random vector taking values in the Euclidean ball, expressed by the moments of a random vector based on the definition of the power of a vector are proposed.
- Anna Denkowska, Stanisław Wanat, "Linkages and systemic risk in the European insurance sector: Some new evidence based on dynamic spanning trees", with presentation of the analysis results of linkage dynamics and systemic risk in the European insurance sector, which are obtained using correlation networks.
- Czesław Domański, "Some remarks about normality tests based on characteristics of stochastic processes", with some results on normality tests.
- Wojciech Gamrot, „On Likert scale and regression coefficient", where an approach of using the Likert scale variables in statistical surveys with closed questions is considered.
- Grzegorz Kończak, "On permutation multivariate extension of McNemar test", with the proposal of the extension of the well-known McNemar test based on data from k ($k > 2$) samples.
- Jerzy Korzeniewski, "Determining semantic relatedness of concepts – modifications proposals", with presenting the modification of the Leacock and Chodorow method in determining the semantic relatedness of concepts.
- Małgorzata Krzciuk, "On EBLUP under some linear mixed model with correlated random effects", with considerations on the problem of small area prediction under a linear mixed model with presenting results of the Monte Carlo simulation analyses based on real data from the Local Data Bank of Statistics Poland.

- Dominika Polko-Zając, "On permutation tests for comparing multidimensional populations", with presentations of a permutative, simultaneous procedure for identifying differences between the vectors of average values and the variance-covariance matrices in two studied populations.
- Dominik Sieradzki, Wojciech Zieliński, "Sample allocation in estimation of proportion in finite populations", where comparison of precision of estimation depending on chosen sample allocation for new proposed method and Neyman allocation and proportional allocation is presented.
- Agnieszka Stanimir, "Multivariate statistical methods in the analysis of multiple responses questions", with presentation of the possibility of using multivariate statistical methods in the analysis of questions with multiple choices of responses.
- Piotr Sulewski, "Recognizing distributions rather than goodness-of-fit testing", where the idea of recognizing distributions rather than carrying out classic goodness-of-fit tests based on the measure of discrepancy is considered.
- Krzysztof Szymoniak-Książek, "Properties of nonparametric isotropy tests" focusses on the discussion of properties of nonparametric significance tests verifying random field isotropy hypothesis.
- Janusz L. Wywiał, Grzegorz Sitek, "On variance of sample matrix eigenvalue", where the estimator being a function of simple random sample variances and covariances of a multidimensional random variable whose distribution is not necessarily normal is regarded.
- Artur Zaborski, "Triads or tetrads? Comparison of incomplete methods for measuring similarity in preferences" with the comparison of the two incomplete methods for measuring the similarity of preferences, i.e. the triad method and the tetrad method.
- Tomasz Żądło, "On generalization of Quatember's bootstrap", where a generalization of the Quatember algorithm is proposed with the study on the properties of the proposal with recent competitors.

Papers presenting practical aspects as well as theoretical ones in the field of the multivariate statistical analysis are the following:

- Maciej Beręsewicz, Katarzyna Zadroga, "Estimation of the number of illegally residing foreigners in Poland in 2017–2018 using Bayesian non-linear mixed count regression models" focuses on estimating the number of foreigners residing illegally in Poland in 2017–2018, where the Bayesian non-linear mixed model for count data was proposed, depending solely on the aggregated data reported by the Border Guards, the Police and in the PESEL register.

- Michał Bernardelli, "Identification of turning points in time series from the cryptocurrency market" with the investigation of the possibility of using the hidden Markov models and Viterbi paths for the analysis of one-dimensional price series from the cryptocurrency market.
- Jacek Białek, "Chain drift problem in the CPI measurement based on scanner data" with presentation of some simulation results which show the situations on the market leading to the biggest chain drift bias if the index differs from unity when prices revert back to their base level.
- Beata Bieszk-Stolorz, "Selected models of recurrent events in the assessment of the risk of re-registration in the labour office" with the analysing of the risk of subsequent registrations in the labour office depending on selected characteristics of the unemployed: gender, age, education and seniority.
- Second Bwanakare, Marek Cierpiał-Wolan, "Generalised Cross-Entropy Econometrics vs conflicting cross-border (Big) data sources. National accounts updating", where the proposal of an efficient approach to combining data from various sources and a comparison of the results with the traditional technique applied in official statistics are presented.
- Grażyna Dehnel, Marek Walesiak, "An assessment of social cohesion of Poland's provinces based on classic and interval-valued data" focusses on the description of a comparative analysis of results of assessing social cohesion with two assessment criteria: cluster analysis to identify similarities and differences in the ranking of provinces, and the analysis of the degree to which different rankings of objects with respect to specific variables correspond to those obtained by using the aggregate measure for 4 datasets.
- Małgorzata Graczyk, Bronisław Ceranka, "Some remarks about highly D-efficient spring balance weighing designs", with consideration of a new construction method of determining highly D-efficient spring balance weighing designs in classes in which D-optimal design does not exist.
- Małgorzata Graczyk, Bronisław Ceranka, "New results regarding the construction method of D-optimal chemical balance weighing designs", where the study of the experiment in that determination of the unknown measurements of p objects in n weighing operations according to the model of the chemical balance weighing design is presented.
- Wioletta Grzenda, "Bayesian multinomial logit models for disordered categories in the analysis of the situation of young people in the labour market in Poland" focusses on the binomial logit model used in the analysis of the situation of respondents in the labour market with special attention paid to inequalities in the labour market in Poland and the problem of saturation of this market with university graduates.

- Stanisław Jaworski, "Some remarks about estimation of Polish unemployment rate", where the discussion on the estimation of unemployment rate by using structural time series model is presented.
- Alina Jędrzejczak, Kamila Trzcińska, "Application of the Zenga Distribution to the analysis of household income in Poland by socio-economic group" with the results of the calculations confirming that the Zenga distribution is a good income distribution model, which can be successfully applied to income inequality analysis and income distribution comparisons.
- Adam Juszczak, "Application of web-scrapping in inflation measurement", where both positive and negative aspects of web-scraping usage in the Consumer Price Index Calculation (CPI) are considered.
- Marta Małecka, "Asymptotic Properties of Duration-Based VaR Backtests" focusses on applying the non-standard likelihood ratio properties, especially a generalized geometric VaR test, with presenting its asymptotic distribution.
- Iwona Markowicz, Paweł Baran, "Divergences in intra-Community trade: the case of Poland" deals with the analysis of data discrepancies in Polish trade in relations: Poland–EU country (bilateral relations) and Poland–EU countries (country–countries relationship, called an aggregate).
- Aneta Ptak-Chmielewska, "Application of multidimensional classification to prediction of SME", with a comparison of the effectiveness of linear discriminant analysis with multidimensional discrimination, such as support vector machines.
- Elżbieta Roszko-Wójtowicz, Maria M. Grzelak, "Innovation activities and competitiveness of manufacturing divisions in Poland in the years 2009–2017", where measuring and assessing the impact of innovative activity on the competitiveness of manufacturing divisions are presented using both static lagged panel models and dynamic panel models.
- Grażyna Trzpiot, "Seniors in cities and senior friendly cities analysis for selected Polish cities" focusses on the results of a study assessing selected Polish cities as senior-friendly cities, using the robust taxonomic approach.
- Łukasz Wawrowski, "Impact of dependent variable transformation on poverty rate estimates in poviats" presenting the results of the estimation of headcount ratio at LAU 1 level in Poland that was possible through the use of data from the EU-SILC and The Polish Census of Population and Housing and indirect estimation methods.
- Ewa Wycinka, Beata Jackowska, "Competing risks models in estimation of companies life time" focuses on the proposal of the use of estimators (the naive Kaplan-Meier estimator, the Aalen-Johansen estimator and the IPCW estimator)

which take into account the type of event in modelling the distribution of enterprise existence time.

- Łukasz Ziarko, "On the possibility of using association analysis to describe the behaviour of contractors in public tenders" with the presentation of the application of association analysis (basket analysis) described in the literature to identify illegal agreements concluded between the applicants for public procurement and evaluation of the proposed approach.

The XXXVIII conference on Multivariate Statistical Analysis 2019 was closed by the Head of the Scientific Committee, Professor Czesław Domański, who summarized the conference and thanked the guests for arriving and taking active participation in the conference. The next edition of MSA 2020 conference is planned for November 16–18, 2020 and will be held in Łódź, Poland.

Prepared by
Aleksandra Baszczyńska
Katarzyna Bolonek-Lasoń
Department of Statistical Methods, University of Łódź

STATISTICS IN TRANSITION new series, June 2020
Vol. 21, No. 2 pp. 209–213,

About the Authors

Alabid Abdelhakim is an Associate Professor at the Department of Statistics and Informatics, Faculty of Commerce and Economics, University of Sana'a. His main areas of interest include multivariate statistical analysis, analysis of multivariate functional data and correlation and regression studies. Simultaneously he is the deputy head of the Central Statistical Organization. Dr. Abdelhakim Alabid has published over 3 research papers in international journals. He has also published two books.

Chaudhuri Arijit is a honorary Visiting Professor in Indian Statistical Institute (ISI), after serving there as a Professor and later as CSIR Emeritus Scientist. His main area of research has been sample surveys and for a brief period reliability and life testing. He has guided successfully 10 PhD students. He has published about 150 papers in peer-reviewed journals, a few jointly with students and colleagues, and 11 books on survey sampling with Marcel Dekker, Taylor & Francis, CRC Press, North Holland, LAP (Germany) and Prentice Hall of India as the publishers. PhD from Calcutta University, post-doctoral research for 2 years in Sydney University and visited with Academic assignments in universities/Statistical Offices in USA, Canada, England, Germany, the Netherlands, Sweden, Israel, Cuba, Cyprus, Bangladesh, Turkey, Japan and South Africa, Australia intermittently during 1973–2007.

Chwila Adam is a PhD student at the Department of Statistics, Econometrics and Mathematics at the University of Economics in Katowice. He is employed as a market risk management specialist in the ING BSK. His research interests focus on small area estimation, data transformations, mixed models and metaheuristic optimization algorithms.

Hurairah Ahmed is an Associate Professor at the Department of Statistics and Informatics, Faculty of Commerce and Economics, University of Sana'a. Simultaneously he holds the position of the Dean of Graduate Studies at the Azal University for Human Development. His main areas of interest include statistical analysis, statistical inference and data analysis. Dr Ahmed Hurairah has published over 15 research papers in international journals and conferences. He has also published two books.

Just Małgorzata is an Assistant Professor at the Department of Finance and Accounting, Faculty of Economics and Social Sciences, Poznań University of Life Sciences, Poland. Her research interests include financial econometrics, volatility

modelling, dependence structure modelling in commodity and financial markets, modelling of extreme events, investment, risk management. She is a member of the International Institute of Social and Economic Sciences

Łuczak Aleksandra is an Associate Professor at the Department of Finance and Accounting, Faculty of Economics and Social Sciences, Poznań University of Life Sciences, Poland. Her main research interests include multicriteria quantitative methods and their applications in economics and finance. She is especially interested in the taxonomic methods and the decision making methods and their applications in solving problems related to local and regional development planning. She is also a member of the Polish Statistical Association and a member of editorial team of Journal of Agribusiness and Rural Development.

Motoryn Ruslan is a Professor at the Department of Statistics and Econometrics in Kyiv National University of Trade and Economics, Ukraine. His research interests are international statistics, economic statistics and the system of national accounts. Professor Motoryn has published over 160 research papers in international/national journals and conferences. He has also published eight books/monographs. Professor Motoryn is an active member of many scientific professional bodies including Ordinary (elected 1998) Member of International Statistical Institute, International Association for Official Statisticians, International Association for Statistical Education, Member of the Committee for Glossary of Statistical Terms of ISI, National correspondent of Ukraine (IASE).

Pal Sanghamitra is an Assistant Professor at the Department of Statistics, West Bengal State University, India. Her main areas of interest include survey sampling, development on unequal probability sampling scheme, randomized response technique, resampling methods, adaptive sampling, small area estimation theory, etc. She has received her PhD degree from Indian Statistical Institute, Kolkata, India. Dr Pal has published twenty research papers in peer-reviewed journals. She attended many international conferences as an invited speaker and organized invited sessions in international/national conferences. She was the guest editor of the journal statistics and applications: Special issue on randomized response technique, 2017, 15(1&2). She is the reviewer of many international journals.

Patra Dipika is a Research Scholar at the Department of Statistics, West Bengal State University, India. Her main areas of interest include survey sampling, randomized response, adaptive sampling, application of Kalman Filtering, etc. She has presented papers in international conferences.

Prykhodko Kateryna is an Associate Professor at the Department of Statistics and Demography, Faculty of Economics, Taras Shevchenko National University of Kyiv, Ukraine. Her main areas of interest include international comparisons of

macroeconomic indicators, labour migration and education quality. Currently, she is a regular member of the International Statistical Institute and a member of the Association of Statisticians of Ukraine.

Safari-Katesari Hadi is a PhD candidate at the Department of Mathematics, Southern Illinois University, Carbondale, IL, USA. His research interest is multivariate statistical analysis, copula and dependency models and biostatistical data. He has published two research papers in international journals.

Shukla Alok Kumar is an Assistant Professor at the Department of Statistics, D.A-V., College, CSJM University, Kanpur, India. He has earned his MSc and PhD from CSJM University, Kanpur, India. He has 14 years of teaching and research experience. His research interests are regression analysis, sampling and econometrics. Dr Shukla has published over 35 research papers in international/national journals of repute.

Singh S. K. is currently working as a Professor in the Department of Statistics, Banaras Hindu University Varanasi, India. He did his MSc in Statistics at the Gorakhpur University in 1985 and doctorate from Banaras Hindu University in 1989 in the area of statistical inference. He has an excellent academic record and holds many administrative as well as educational posts in the university. He is a prominent researcher in the field of classical and Bayesian inference, reliability analysis with censored data and distribution theory, and published over 100 papers in very reputed journals of statistics. He is a reviewer in many statistical and applied science journals and has reviewed many articles related to his area expertise.

Singh Umesh is currently working as a Professor in the Department of Statistics, Banaras Hindu University, Varanasi, India. He did his MSc in statistics in 1973 at the Allahabad University, Allahabad, India, and PhD in statistics in 1978 in the University of Rajasthan, Rajasthan, India. His academic career is excellent and he has occupied several administrative positions. He has extensive knowledge of several areas of statistics and contributed over 120 research papers in national and international journals of repute. He reviewed over 100 research papers for various well-reputed international journals. He is a member of many editorial boards of several national and international journals. He is a renowned researcher and has a good number of contributions in almost all fields of statistics, viz., distribution theory, reliability theory, Bayesian inference, record statistics, analysis with different types of censoring schemes, etc.

Szczepocki Piotr is an Assistant Professor at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz. His main areas of interest include sequential Monte Carlo methods and volatility modelling.

Ślusarczyk Bogusław is the employee of the Department of Economics and International Economic Relations at the Institute of Economics and Finance, University of Rzeszów (Poland). His research interests focus on issues related to competitiveness of Polish market in comparison with other European countries. His interests are also in the development of competitiveness in the micro- and macroeconomic scale before and after Poland's accession to the EU as well as in the light of the ongoing changes in the global economy. He is the author and co-author of over 280 scientific papers and expertise. He is the member of many editorial boards of scientific journals in Poland and abroad. He acts as a chairman and a member of domestic and international scientific conferences.

Tharshan Ramajeyam is a lecturer at the Department of Mathematics and Statistics, University of Jaffna, Sri Lanka. He has received his undergraduate degree from University of Jaffna, Sri Lanka and MSc degree in statistics from Wright State University, USA. His research interests are in the areas of finite mixture models, models for over-dispersed count data, statistical inference and data analysis. Currently, he is an MPhil student in Statistics at the Postgraduate Institute of Science, University of Peradeniya, Sri Lanka.

Wijekoon Pushpakanthie has received her PhD degree from the University of Dortmund, Germany. She is currently a Senior Professor at the Department of Statistics and Computer Science, University of Peradeniya, Sri Lanka. The scope of her research interests include the biased estimation, Stein-rule estimation and preliminary test estimation in the linear regression model, misspecified linear models, binomial and Poisson mixture models, improved methods in estimation in the exponential family of distributions and multivariate statistical methods. She has published over 50 research papers in journals and conferences throughout her professional career. She has also served as a referee for various reputed national and international journals. Professor Wijekoon is also an active member of many scientific professional bodies.

Yadav Abhimanyu Singh is currently working as an Assistant Professor in the Department of Statistics, Banaras Hindu University Varanasi, India. He has also worked for the Department of Statistics, Central University of Rajasthan, Rajasthan, and PUC Mizoram University Aizawl, India. He did his MSc in statistics (2011) in the Banaras Hindu University, Varanasi, and PhD in statistics (2015) in the same University. He has received a merit scholarship in MSc first year from the Department of Statistics, Banaras Hindu University. He is a young researcher and has published 28 research papers in the area of statistical inference and distribution theory in national and international journals of repute. He has reviewed over 15 research papers for various well-reputed national/international journals.

Yadav Subhash Kumar is an Assistant Professor at the Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow, India. He has earned his MSc and Ph.D. from University of Lucknow, Lucknow, India. He has 14 years of teaching and research experience. His research interests are regression analysis, sampling, econometrics and operations research. Dr Yadav has published over 40 research papers in Scopus/WOS indexed international/national journals of repute and two books from an international publisher. He is a referee for over 20 reputed international journals. He has been awarded Young Scientist Award in 2016 for the contribution in the field survey sampling by Venus International Research Foundation, Chennai, India. He has presented papers in over 18 national and international conferences and also delivered invited talks in several conferences. He was awarded Best Paper Award in 2018 in MTMI International Conference at Virginia Beach Resort Hotel, Virginia Beach, Virginia, USA.

Zaroudi Samira received her PhD in 2018 at the Department of Statistics, Azad University, Tehran, Iran. Currently, she is pursuing her master's degree in Mathematics at Southern Illinois University, Carbondale, IL, USA. Her research interest is multivariate statistical analysis, copula and dependency models and actuarial data. He has published three research papers in international journals.

Żądło Tomasz is employed as an Associate Professor at the Department of Statistics, Econometrics and Mathematics at the University of Economics in Katowice. His research interests focus on small area estimation, survey sampling, mixed models and bootstrap methods. He is an elected member of the International Statistical Institute and a country representative of the International Association of Survey Statisticians.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).