Volume 22, Number 3, September 2021 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

EDITORIAL OFFICE **ISSN 1234-7655**

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

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

CONTENTS

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. III–VII

From the Editor

A set of eleven articles written by twenty two authors from ten countries places *Statistics in Transition new series (SiTns)* among the top scientific journals whose aim is the internationalisation of scientific research. This leads to a more general reflection on the nature and place of our publication in the constellation of the existing scientific periodicals. The studies provided by authors from outside of Europe and North America, with experts from African countries, the Middle East and South East Asia being particularly active, complement geographically one of the transition-related components emphasised in *SiTns'* mission objectives; the other one has traditionally meant the permanent process of the inherent development of the discipline. Recognising both directions of the promotion of articles (as the title of the journal suggests) – focusing either on the development of the discipline or on the statistical analysis of real issues in various environments (including developed and developing countries) – *SiTns* has been long pursuing the free-access and free-of-charge (for authors/APC) policy. We strive to continue our work in this very way, believing that we provide an important platform (a niche), essential for the presentation of a wide spectrum of research within both aspects, i.e. the authors' background and research topics, ideally combined and introduced in an innovative manner. The papers contained in the issue are examples of such an approach and policy.

Research articles

The article entitled ***Acceptance sampling plans from a truncated life test based on the power Lomax distribution with application to manufacturing*** by **Amjad D. Al-Nasser** and **Mohammad A. ul Haq** starts with the assumption that the quality characteristic follows the power Lomax distribution. The operating characteristic function values are calculated for the proposed sampling plan, jointly with the optimal sample size and the producer's risk for a selection of distribution parameters. Furthermore, a comparative study with other sampling plans is introduced to demonstrate the advantages of the proposed plan. Finally, a real-life example illustrating the applicability of the proposed sampling plan in a manufacturing company is discussed. Comparisons with other lifetime distributions showed that sampling plans based on PLxD are more efficient than their other counterparts.

In the next paper, ***A calibrated synthetic estimator for small area estimation*****, Matthew J. Iseh** and **Ekaette I. Enang** discuss synthetic estimators used to produce

estimates of population mean in areas where no sampled data are available. Given that such estimates are usually highly biased with invalid confidence statements, the paper presents a calibrated synthetic estimator of the population mean which addresses several problematic issues. Two known special cases of this estimator were obtained in the form of combined ratio and combined regression synthetic estimators, using selected tuning parameters under stratified sampling. In result, their biases and variance estimators were derived. The empirical demonstration of the usage involving the proposed calibrated estimators shows that they provide better estimates of the population mean than the existing estimators discussed in the study. In particular, the estimators were examined through simulation under three distributional assumptions, namely the normal, gamma and exponential distributions. The results show that they provide estimates of the mean displaying less relative bias and greater efficiency. Moreover, they prove more consistent than the existing classical synthetic estimator. The further evaluation carried out using the coefficient of variation provides additional confirmation of the calibrated estimator's advantage over the existing ones in relation to small area estimation. The proposed combined ratio synthetic estimator has shown dominance over the combined regression synthetic estimator suggesting that the latter is not suitable for any real-life data that follow exponential distribution for small domains under stratified sampling.

Andrzej Szymański's and **Agnieszka Rossa's** paper entitled ***The Complex-Number Mortality Model (CNMM) based on orthonormal expansion of membership functions*** deals with a new fuzzy version of the Lee-Carter (LC) mortality model. Mortality rates as well as parameters of the LC model are treated in this model as triangular fuzzy numbers. As a starting point, the fuzzy Koissi-Shapiro (KS) approach is recalled. Based on this approach, a new fuzzy mortality model – CNMM – is formulated using orthonormal expansions of the inverse exponential membership functions of the model components. The paper includes numerical findings based on a case study applying the new mortality model compared to the results obtained with the standard LC model. Moreover, the confidence intervals for log-central mortality rates can also be derived, but they reflect the error term in the random walk model, ignoring the estimation errors of other parameters; thus, the confidence intervals can only be derived for the prediction window.

In the paper ***Bayesian estimation and prediction based on Rayleigh record data with applications*****, Abu Awwad R. R., Bdair O. M.,** and **Abufoudeh G. K.,** consider the problem of estimating the scale and location parameters of the model and predicting the future unobserved record data using a record sample from the Rayleigh model. Maximum likelihood and Bayesian approaches under specific loss functions are used to estimate the model's parameters. The Gibbs sampler and Metropolis-Hastings

methods are applied within the Bayesian procedures to draw the Markov Chain Monte Carlo (MCMC) samples, used in turn to compute the Bayes estimator and the point predictors of the future record data. Two examples of real-life data have been analysed to illustrate the developed procedures.

Vipin Sharma and **Vinod Kumar** in their article entitled ***Trade potential under SAFTA between India and other SAARC countries: the augmented gravity model approach*** attempt to assess the trade potential for the years 1992 to 2019 at annual frequency in general, and for 2004 to 2019 in detail. The findings of this paper show that intra-regional trade volumes between SAARC nations can be increased and encouraged. It is important to undertake structural reforms in order to boost trade with non-member countries. The authors suggest that research on this issue should take into account the effect of locational and infrastructural advantages on transportation costs using a gravity model. Previous research has also shown that an augmented gravity model may help in explaining some key features of South Asian trade which may not be explained by traditional gravity models.

Haitham A. Yousof, M. Masoom Ali, Hafida Goual, and **Mohamed Ibrahim** in their article ***A new Reciprocal Rayleigh Extension: properties, copulas, different methods of estimation and a modified right censored Test for validation*** derive the new reciprocal Rayleigh extension's relevant statistical properties. The authors emphasise the results of their research related to convexity and concavity and discuss their estimation of the model's parameters using different estimation methods such as the maximum likelihood estimation method, the ordinary least squares estimation method, the weighted least squares estimation method, the Cramer-Von-Mises estimation method, and the bootstrapping method. The performances of the proposed estimation methods are investigated through a simulation study. Several bivariate- and multivariate-type models have also been derived based on the Farlie Gumbel Morgenstern copula, the Clayton copula, Renyi's entropy copula and the Ali-Mikhail-Haq copula. The modified Nikulin-Rao-Robson test for right censored validation is applied to a censored real data set.

Agata Boratyńska's paper ***Robust Bayesian insurance premium in a collective risk model with distorted priors under the generalised Bregman loss*** presents a collective risk model relating to insurance claims. The objective is to calculate a premium, which is defined as a functional specified up to unknown parameters. The Bayesian methodology, which combines the prior knowledge about certain unknown parameters with the knowledge in the form of a random sample, has been adopted, along with the generalised Bregman loss function. The results, however, can be applied to numerous loss functions, including the square-error, LINEX, weighted square-error, Brown, entropy loss. Some uncertainty about a prior is assumed by a distorted band class of

priors. A range of collective and Bayes premiums is calculated and posterior regret Γ-minimax premium as a robust procedure has been implemented. Two examples are provided to illustrate the issues considered – the first one with an unknown parameter of the Poisson distribution and the second one with unknown parameters of distributions of the number and severity of claims.

Hajar Bouazzaoui, Mohamed Abdou Elomary, and **My Ismail Mamouni** discuss ***An application of persistent homology and the graph theory to linguistics: The case of Tifinagh and Phoenician scripts*** using tools from within topological data analysis and the graph theory for identifying the similarity between the two scripts (Tifinagh and Phoenician). The clustering of their letter shapes is performed based on the pairwise distances between their topological signatures. The ideas presented in this work can be extended to study the similarity between any two writing systems and as such can serve as the first step for linguists to determine the possibly related scripts before conducting further analysis. For instance, a future work might explore the nature of this relatedness, i.e. whether one script is derived from the other or one was built under the other's influence.

In the paper ***A new count data model applied in the analysis of vaccine adverse events and insurance claims*****, Showkat Ahmad Dar, Anwar Hassan, Peer Bilal Ahmad** and **Sameer Ahmad Wani** present a new probability distribution, created by compounding the Poisson distribution with the weighted exponential distribution. Important mathematical and statistical properties of the distribution are derived and discussed. The paper describes the proposed model's parameter estimation, performed by means of the maximum likelihood method. Finally, a real data set is analysed to verify the suitability of the proposed distribution in modelling a count dataset representing vaccine adverse events counts and insurance claims.

Other articles:

38th Multivariate Statistical Analysis 2019, Łódź. Conference Papers

The last paper, by **Wioletta Grzenda**, entitled ***Modelling the occupational and educational choices of young people in Poland using Bayesian multinomial logit models*** is based on a presentation delivered at the 2019 Multivariate Statistical Analysis conference. A binomial logit model is applied to two states: of being unemployed and employed, or economically inactive and active. The paper focuses on the situation of young people aged 18 to 29 on the labour market in Poland. They were divided into the following groups: the employed and not learning, those combining education with work, the unemployed, learners/students only, and those economically inactive and not at school. All the different models were estimated within the Bayesian approach. The findings show that continuing education by young people may result from their

problems with finding a job; moreover, combining work with education is not the preferred form of professional activity. In addition, the study examines the inequalities observed on the Polish labour market. The paper provided a new insight into how young people enter the labour market in Poland.

Research Communicates and Letters

The Research Communicates and Letters section presents an article entitled ***On improvement of paired ranked set sampling to estimate population mean*** by **Syed Abdul Rehman** and **Javid Shabbir**, who discuss the difficulties involved in the quantification of units in ecological and environmental sampling, relating to time, money, workload, etc. Therefore, the need for efficient and cost-effective sampling methods was identified and addressed by the authors who propose a sampling scheme called Improved Paired Ranked Set Sampling (IPRSS) to estimate the population mean. The performance of the proposed IPRSS is evaluated under perfect and imperfect rankings. A simulation study based on selected hypothetical distributions and a real-life data set show that IPRSS is more precise than RSS, Paired RSS (PRSS) or Extreme

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. IX

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, September 2021
Vol. 22, No. 3, pp. XI–XII

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
- TDNet
- Technische Informationsbibliothek (TIB) - German National Library of Science and Technology
- Ulrichsweb & Ulrich's Periodicals Directory
- WanFang Data
- WorldCat (OCLC)
- Zenodo

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 1–13, DOI 10.21307/stattrans-2021-024
Received – 19.09.2020; accepted – 18.01.2021

Acceptance sampling plans from a truncated life test based on the power Lomax distribution with application to manufacturing

Amjad D. Al-Nasser[1], **Muhammad Ahsan ul Haq**[2]

ABSTRACT

In this research, a new acceptance sampling plan for a truncated life test is presented, assuming that the quality characteristic follows the power Lomax distribution. The operating characteristic function values are calculated for the proposed sampling plan, jointly with the optimal sample size and the producer's risk for a selection of distribution parameters. Furthermore, a comparative study with other sampling plans is introduced to demonstrate the advantages of the proposed plan. Finally, a real-life example illustrating the applicability of the proposed sampling plan in a manufacturing company is discussed.

Key words: acceptance sampling plan, operating characteristic, power Lomax distribution, industry, data analysis.

1. Introduction

Acceptance sampling plans play a very important role in the statistical quality control, especially in the lot production process to decide whether to reject or accept the lot (Al-Nasser and Al-Omari, 2013). The decision on the quality of all entire items in each lot depends on drawing a random sample of size n from a selected lot; after that, within a specific timeframe, testing procedure is initiated to discover the number of failure or defective items included in the sample before the pre-indicated time is terminated (Al-Nasser and Gogah, 2017; Al-Omari et al., 2016; Malathi and Muthulakshmi, 2017).

Then, the problem is to find the optimal sample size n that is necessary to assure a certain average life, when the life test is terminated at a pre-assigned time t. Such that the observed number of failures does not exceed a given acceptance number c. Accordingly, the decision is to reject all entire items in the lot if the number of failures

[1] Department of Statistics, Science Faculty, Yarmouk University, Irbid, Jordan & Vice President for Academic Affairs, Al Falah University, Dubai, United Arab Emirates. E-mail: amjadyu@yahoo.com. ORCID: http://orcid.org/0000-0001-7515-2357.

[2] Quality Enhancement Cell, National College of Arts, Lahore-Pakistan. College of Statistical & Actuarial Sciences, University of the Punjab, Lahore, Pakistan. ORCID: http://orcid.org/ 0000-0002-0902-8080.

in the sample within the timeframe is greater than the pre-assigned acceptance number (c); elsewhere the lot is accepted.

Many authors proposed truncated life test plans for different lifetime distributions. For example, Epstein (1954) was the first who considered truncated life tests in the exponential distribution. The truncated life tests in the Pareto distribution of the second kind are discussed by Baklizi (2003). The Rayleigh model is proposed by Baklizi *et al.*, (2005). The generalized Birnbaum-Saunders Distribution is discussed by Balakrishnan *et al., (*2007). The Marshall-Olkin extended Lomax distribution is given by Rao *et al., (*2008). The generalized exponential distribution is considered by (Rao, 2010). The new Weibull-Pareto distribution is proposed by Al-Omari *et al.*, (2016), weighted exponential distribution is discussed by Gui and Aslam (2017), exponentiated generalized inverse Rayleigh distribution is discussed by Al-Nasser *et al.*, (2017). The inverse gamma model is given by Al-Masri (2018), and Tsallis q-exponential distribution is proposed by (Al-Nasser & Obeidat, 2020).

The purpose of this article is to develop and discuss an acceptance sampling plan (ASP) for a truncated life test on the power Lomax distribution (PLxD) and illustrate the results on manufacturing data. The rest of the paper is organized as follows. Section 2 is based on summaries of PLxD and some of its properties. ASPs and operating characteristic (OC) values and the producer's risk for PLxD are analysed. The analysis and illustrative examples are presented in Section 4. A comparative study between the proposed ASP and other sampling plans based on different distributional assumptions is discussed in Section 5. A real manufacturing data analysis is given in Section 6. The work is concluded in Section 7.

2. The Power Lomax distribution

Power Lomax distribution (PLxD) originally proposed by (Rady, Hassanein, & Elhaddad, 2016) is a lifetime distribution obtained by taking the power of the Lomax distribution random variable. The PLxD distribution is very flexible due to its variable shapes of hazard rate, which accommodates both inverted bathtub and decreasing.

The probability density function (pdf) of PLxD is

$$f(x) = \alpha\beta\lambda^{\alpha}x^{\beta-1}\left(\lambda + x^{\beta}\right)^{-\alpha-1}, \quad x > 0, \qquad \alpha, \beta, \lambda > 0. \tag{1}$$

The corresponding cumulative distribution function (CDF) is

$$F(x) = 1 - \lambda^{\alpha}\left(\lambda + x^{\beta}\right)^{-\alpha} \tag{2}$$

From Figure 1 it can be easily concluded that the shapes of PLxD have a decreasing behaviour for $\beta < 1$, the distribution has an exponentially decreasing behaviour but starting from the y-axis for $\beta=1$. For $\beta > 1$ the pdf curves of the model are unimodal and symmetrical for some combinations of parameters.

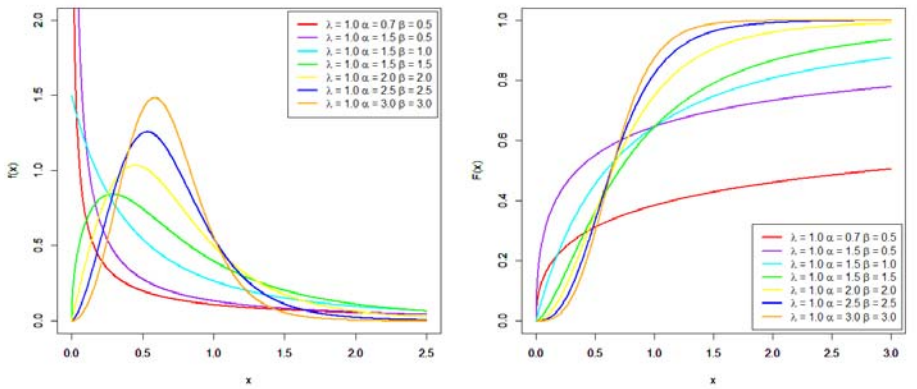


Figure 1. pdf and CDF of PLxD for some given parameter values

The PLxD has the following properties. The rth moments about origin and mean of PLxD are, respectively:

$$\mu_r' = \frac{\alpha\lambda^{\frac{r}{\beta}}\,\Gamma\left[\alpha - \frac{r}{\beta}\right]\Gamma\left[\frac{r+\beta}{\beta}\right]}{\Gamma[1+\alpha]}$$

Therefore, the mean of PLxD is given as:

$$\mu = \frac{\alpha\lambda^{\frac{1}{\beta}}\,\Gamma\left[\alpha-\frac{1}{\beta}\right]\Gamma\left[\frac{1+\beta}{\beta}\right]}{\Gamma[1+\alpha]} \tag{3}$$

The PLxD has an increasing-decreasing hazard rate function, for more details see (Rady, Hassanein, & Elhaddad, 2016).

3. The suggested sampling plans

Suppose that the lifetime of the products being tested follows the PLxD as given in (1) and the specified mean lifetime is μ_0. Now, the ASP problem is to find the optimal sample size that ensures an actual average life (μ) such that no more than c units fail to pass the test period (t). To perform the test according to this plan, a random sample of size n units is selected from a lot. If μ_o can be obtained with a pre-assigned probability, P^*, as specified by the consumer, then the lot is accepted. If not, then it is rejected.

Following Al-Nasser and Obeidat (2020), the ASP-based on truncated life tests consists of the following parameters:

1) The sample size: number of units' n to be drawn from the lot.
2) The test duration t: the maximum test duration time.
3) Acceptable number of defective (d) items: c; if $d \le c$ remains the same until the end of the test period t_0, the lot is accepted.

4) The minimum ratio t/μ_o: where μ_0 is the quality parameter of the product life; and t is the maximum test duration.
5) The ASP parameters will be $(n, c, t/\mu_o)$.

3.1. Optimal sample size of the ASP (n, c, t/μ$_o$)

Let P* be the confidence level where $P^* \in (0,1)$; in the sense that the possibility of rejecting a considered lot having a specified mean less than or equal to the actual mean $(\mu_0 \leq \mu)$ is greater or equal to P*. The shopper's risk, that is the probability of accepting a defective lot, is fixed and less or equal to 1–P*. Also, suppose that the lot size is large enough to use the binomial distribution. Then, the problem of the ASP $(n, c, t/\mu_o)$ is to find the minimum sample size n such that the number of defective units d does not exceed c, to ensure that $\mu > \mu_0$ satisfies the following inequality:

$$\sum_{i=0}^{c} \binom{n}{i} p^i (1-p)^{n-i} \leq 1 - p^* \tag{4}$$

where

$$p = F(t;\mu_0) = 1 - \lambda^{\alpha}\left(\lambda + \left(\frac{t}{\mu_0}\frac{\alpha\lambda^{\frac{1}{\beta}}\Gamma\left[\alpha - \frac{1}{\beta}\right]\Gamma\left[\frac{1+\beta}{\beta}\right]}{\Gamma[1+\alpha]}\right)^{\beta}\right)^{-\alpha}$$

By using the binomial theory, the probability of success in (4), which is used for finding a defective item in each a lot during the test process time t, is $p = F(t; \mu_0)$; this probability in terms of distribution function is a monotonically increasing function of the ratio t/μ_o. Then, for the acceptance sampling ASP $(n, c, t/\mu_o)$ and inequality (4), we assure that $F(t; \mu) \leq F(t; \mu_0)$ with probability P^*, or alternatively $\mu_0 \leq \mu$. The results for this plan when the lifetime distribution is PLxD with $\alpha = 1;\ \beta = 2\ and\ \lambda = 1$ are given in Table 1, under the classical initial values of the ratio t/μ0 = 0.628, 0.942, 1.257, 1.571, 2.356, 3.141, 3.927, 4.712, when P* = 0.75, 0.9, 0.95, 0.99 and c = 0, 1, 2, … , 10 (Gupta and Groll, 1961; Kantam and Rosaiah, 2001; Baklizi, 2003; Baklizi et al., 2005; Al-Nasser et al., 2018; Al-Masri, 2018; Al-Omari et al., 2019).

3.2. Operating characteristic function of the ASP (n, c, t/μ$_o$)

Operating characteristic (OC) function is an important parameter in the ASP, it provides the exact information about the probability of acceptance of a lot. For the *ASP (n, c, t/μ$_o$),* the OC can be computed using binomial distribution as:

$$OC(p) = \sum_{i=0}^{c} \binom{n}{i} p^i (1-p)^{n-i}$$

which can be computed using the incomplete beta function $B_p(a, b)$ as:

$$OC(p) = 1 - B_p(c+1, n-c)$$

where $p = F(t; \mu)$. Table 2 presents the OC function values for the *ASP (n, c, t/μ$_o$)*.

Table 1. Minimum sample size to assert that the mean life exceeds a given value μ_0 with probability P^* and acceptance number *c* based on binomial probabilities when $\alpha = 1;\ \beta = 2\ and\ \lambda = 1$

P^*		t / μ_0							
	c	0.628	0.942	1.257	1.571	2.356	3.141	3.927	4.712
0.75	0	3	2	1	1	1	1	1	1
	1	5	3	3	3	2	2	2	2
	2	7	5	4	4	3	3	3	3
	3	10	7	6	5	4	4	4	4
	4	12	8	7	6	6	5	5	5
	5	14	10	8	8	7	6	6	6
	6	16	11	10	9	8	7	7	7
	7	19	13	11	10	9	9	8	8
	8	21	15	12	11	10	10	9	9
	9	23	16	14	12	11	11	10	10
	10	25	18	15	14	12	12	11	11
0.9	0	4	2	2	2	1	1	1	1
	1	7	4	4	3	3	2	2	2
	2	9	6	5	4	4	4	3	3
	3	12	8	7	6	5	5	4	4
	4	14	10	8	7	6	6	6	5
	5	17	11	9	8	7	7	7	7
	6	19	13	11	10	8	8	8	8
	7	22	15	12	11	10	9	9	9
	8	24	16	14	12	11	10	10	10
	9	26	18	15	13	12	11	11	11
	10	29	20	16	15	13	12	12	12
0.95	0	5	3	2	2	2	1	1	1
	1	8	5	4	4	3	3	3	2
	2	11	7	6	5	4	4	4	4
	3	14	9	7	6	5	5	5	5
	4	16	11	9	8	7	6	6	6
	5	19	12	10	9	8	7	7	7
	6	21	14	12	10	9	8	8	8
	7	24	16	13	12	10	9	9	9
	8	26	18	14	13	11	11	10	10
	9	29	19	16	14	12	12	11	11
	10	31	21	17	15	14	13	12	12
0.99	0	7	4	3	3	2	2	2	2
	1	11	7	5	5	4	3	3	3
	2	14	9	7	6	5	4	4	4
	3	17	11	9	7	6	6	5	5
	4	20	13	10	9	7	7	6	6
	5	23	15	12	10	9	8	8	7
	6	25	17	13	12	10	9	9	8
	7	28	18	15	13	11	10	10	10
	8	30	20	16	14	12	11	11	11
	9	33	22	18	16	13	13	12	12
	10	36	24	19	17	15	14	13	13

Table 2. Operating characteristic function values for the sampling plan $(n, c = 2, t/\mu_0)$ for a given probability P^*

P^*	n	t/μ_0	μ/μ_0					
			2	4	6	8	10	12
0.75	7	0.628	0.85933	0.994463	0.999411	0.999888	0.99997	0.99999
	5	0.942	0.758991	0.985557	0.998274	0.999656	0.999905	0.999967
	4	1.257	0.697065	0.974337	0.996542	0.999275	0.999795	0.999928
	4	1.571	0.51865	0.933522	0.9892	0.997548	0.999276	0.999741
	3	2.356	0.536389	0.901897	0.979068	0.994523	0.998252	0.999345
	3	3.141	0.366443	0.780305	0.934349	0.979078	0.992497	0.996976
	3	3.927	0.259086	0.651114	0.864325	0.948163	0.97906	0.990869
	3	4.712	0.190563	0.536389	0.78025	0.901897	0.955662	0.979068
0.90	9	0.628	0.749567	0.987815	0.998641	0.999736	0.999928	0.999975
	6	0.942	0.639516	0.973686	0.996695	0.999329	0.999813	0.999935
	5	1.257	0.512032	0.945163	0.991985	0.998266	0.999501	0.999824
	4	1.571	0.51865	0.933522	0.9892	0.997548	0.999276	0.999741
	4	2.356	0.222004	0.743325	0.933578	0.980987	0.99364	0.99755
	4	3.141	0.098206	0.518914	0.816848	0.933606	0.974393	0.989217
	3	3.927	0.259086	0.651114	0.864325	0.948163	0.97906	0.990869
	3	4.712	0.190563	0.536389	0.78025	0.901897	0.955662	0.979068
0.95	11	0.628	0.631772	0.978039	0.997433	0.999494	0.999861	0.999952
	7	0.942	0.523698	0.958024	0.994463	0.998854	0.999677	0.999888
	6	1.257	0.35586	0.906065	0.985131	0.996681	0.99903	0.999655
	5	1.571	0.311317	0.867571	0.975906	0.994266	0.998268	0.999372
	4	2.356	0.222004	0.743325	0.933578	0.980987	0.99364	0.99755
	4	3.141	0.098206	0.518914	0.816848	0.933606	0.974393	0.989217
	4	3.927	0.047651	0.341284	0.666449	0.850637	0.933556	0.969202
	4	4.712	0.025325	0.222004	0.518826	0.743325	0.869726	0.933578
0.99	14	0.628	0.464414	0.957378	0.994663	0.99892	0.999699	0.999896
	9	0.942	0.328618	0.916115	0.987815	0.997383	0.99925	0.999736
	7	1.257	0.237226	0.858908	0.975856	0.99444	0.99835	0.999408
	6	1.571	0.174323	0.787955	0.956955	0.989273	0.996684	0.998782
	5	2.356	0.079876	0.572449	0.867672	0.958688	0.985527	0.994273
	4	3.141	0.098206	0.518914	0.816848	0.933606	0.974393	0.989217
	4	3.927	0.047651	0.341284	0.666449	0.850637	0.933556	0.969202
	4	4.712	0.025325	0.222004	0.518826	0.743325	0.869726	0.933578

Table 3. Minimum ratio of the true mean life to specified mean life for the acceptance of a lot with producer's risk of 0.05 with $\alpha = 1;\ \beta = 2\ and\ \lambda = 1$

P^*	c	t/μ_0							
		0.628	0.942	1.257	1.571	2.356	3.141	3.927	4.712
	0	7.512	9.181	8.607	10.757	16.132	21.507	26.888	32.263
	1	3.429	3.74	4.991	6.238	6.896	9.194	11.495	13.792
	2	2.567	3.063	3.433	4.291	4.846	6.461	8.077	9.692
	3	2.348	2.744	3.236	3.419	3.908	5.21	6.513	7.815
	4	2.099	2.32	2.744	2.911	4.366	4.47	5.588	6.705
0.75	5	1.937	2.242	2.417	3.021	3.858	3.971	4.964	5.957
	6	1.822	2.018	2.452	2.727	3.49	3.606	4.509	5.41
	7	1.807	1.996	2.248	2.503	3.209	4.278	4.158	4.989
	8	1.732	1.975	2.086	2.325	2.986	3.98	3.877	4.652
	9	1.672	1.847	2.14	2.18	2.802	3.736	3.646	4.375
	10	1.623	1.843	2.018	2.304	2.649	3.531	3.452	4.142
	0	8.684	9.181	12.251	15.311	16.132	21.507	26.888	32.263
	1	4.155	4.5	6.004	6.238	9.354	9.194	11.495	13.792
	2	2.998	3.48	4.087	4.291	6.434	8.578	8.077	9.692
	3	2.636	3.027	3.662	4.044	5.127	6.835	6.513	7.815
	4	2.324	2.767	3.096	3.429	4.366	5.82	7.276	6.705
0.90	5	2.209	2.426	2.721	3.021	3.858	5.143	6.43	7.715
	6	2.054	2.332	2.692	3.065	3.49	4.653	5.817	6.98
	7	2.001	2.261	2.465	2.809	3.753	4.278	5.349	6.418
	8	1.905	2.093	2.468	2.607	3.486	3.98	4.976	5.971
	9	1.828	2.059	2.309	2.442	3.268	3.736	4.671	5.604
	10	1.809	2.031	2.176	2.522	3.086	3.531	4.415	5.297
	0	9.715	11.268	12.251	15.311	22.961	21.507	26.888	32.263
	1	4.473	5.144	6.004	7.504	9.354	12.471	15.591	13.792
	2	3.373	3.85	4.643	5.108	6.434	8.578	10.725	12.868
	3	2.895	3.285	3.662	4.044	5.127	6.835	8.545	10.254
	4	2.529	2.964	3.408	3.869	5.142	5.82	7.276	8.731
0.95	5	2.372	2.596	2.991	3.401	4.53	5.143	6.43	7.715
	6	2.194	2.474	2.911	3.065	4.089	4.653	5.817	6.98
	7	2.121	2.382	2.663	3.081	3.753	4.278	5.349	6.418
	8	2.011	2.309	2.468	2.857	3.486	4.648	4.976	5.971
	9	1.97	2.157	2.465	2.674	3.268	4.357	4.671	5.604
	10	1.895	2.118	2.322	2.522	3.455	4.114	4.415	5.297
	0	11.503	13.025	15.036	18.792	22.961	30.612	38.272	45.922
	1	5.314	6.232	6.864	8.578	11.253	12.471	15.591	18.708
	2	3.867	4.496	5.137	5.803	7.66	8.578	10.725	12.868
	3	3.244	3.745	4.383	4.576	6.065	8.086	8.545	10.254
	4	2.895	3.321	3.692	4.259	5.142	6.855	7.276	8.731
0.99	5	2.669	3.047	3.464	3.738	5.1	6.039	7.55	7.715
	6	2.451	2.854	3.112	3.638	4.596	5.451	6.815	6.98
	7	2.341	2.605	3.017	3.328	4.213	5.003	6.255	7.505
	8	2.209	2.506	2.793	3.084	3.909	4.648	5.81	6.972
	9	2.146	2.425	2.748	3.08	3.662	4.882	5.447	6.536
	10	2.094	2.359	2.587	2.902	3.782	4.606	5.144	6.172

3.3. Producer's risk of the *ASP (n, c, t/μ$_o$)*

Producer's risk (PR) is another important parameter of the acceptance plans, it measures the probability that a consumer rejects a good lot. Based on binomial theory, PR is computed as follows:

$$PR(p) = \sum_{i=c+1}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

or by using the incomplete beta function:

$$PR(p) = B_p(c+1, n-c)$$

Therefore, for the *ASP (n, c, t/μ$_o$)* is solved as an inequality to ensure that the producer's risk is at most equal to a specific small value (i.e. say P^*) such that the ratio of the actual mean to the specified mean (i.e., μ/μ_o) is as specified by the producer. Therefore, the minimum ratio μ/μ_o is specified as a solution of the following inequality:

$$PR(p) > 1 - P^*$$

where $p = F[(t/\mu_o)\ (\mu_o/\mu)]$; μ). The minimum values of the ratio μ/μ_0 for the *ASP (n, c, t/μ$_o$)* are given in Table 3.

4. Explaining the ASP (n, c, t/μ_o) results

In this article, the parameters of the proposed *ASP (n, c,* t/μ_o *),* the smallest sample size, operating characteristic values and the minimum ratio of the true mean life to specified mean life, respectively, based on the PLxD, are given in Table 1 - Table 3.

For example, assume that the researcher aims to ensure that the product's mean lifetime is at least 1000 hours, with probability $P^* = 0.90$ when $c = 2$ such that the experiment will be terminated at t = 942 hours; that is, $\frac{t}{\mu_0} = 0.942$. Then, from Table 1, the optimal sample size for this plan is 6, accordingly, the appropriated *ASP (n, c,* t/μ_o *)* = (*6, 2,* 0.942). Moreover, from Table 2, the $OC(p)$ for the ASP (*6, 2*, 0.942) are given in Table 4:

Table 4. OC and PR for the ASP (6, 2, 0.942)

μ/μ_0	2	4	6	8	10	12
OC(p)	0.639516	0.973686	0.996695	0.999329	0.999813	0.999935
PR	0.360484	0.026314	0.003305	0.000671	0.000187	0.000065

The plan indicates that the lot is accepted if out of 6 items less than or equal to 2 items fail before the time t. Now, if the true mean is four times as the specified mean $\mu/\mu_0 = 4$ then we are assured that the lot will be accepted under this ASP with probability equal to 0.973686 and the producer's risk is about 0.026314. The probability of accepting a lot under the ASP (*6, 2*, 0.942) will be more than 0.97 if and only if the true mean is four times or more than the specified mean.

Furthermore, the results given in Table 3 indicated that when the consumer's risk is 10% ($P^* = 0.90$) and by using the ASP (*6, 2*, 0.942); then, the minimum ratio $\mu/\mu_0 = 3.48$ when the producer risk is equal to 0.05. It implies that a lot with 6 items when c=2 will be rejected with probability less than or equal to 0.05.

5. Comparative Study

The advantages of the proposed ASP based on LPxD are compared with other ASP under various types of distributions assuming that the actual mean is four times of the specified mean and the acceptable number of defectives is equal to two. The comparison criterion will be the cost of inspection based on the sample size of the ASP and the producer's risk (PR). We said that a sampling plan with a smaller sample size is more efficient in reducing the cost of inspection compared to other ASP; at the sometime, we are seeking minimum value of the PR. The proposed ASP is compared with several ASPs that were proposed by Balakrishnan *et al*. (2007) for the generalized Birnbaum–Saunders distribution (GBSD); Sampath and Lalitha (2016) for the hybrid exponential distribution (HED); Al-Nasser & Obeidat (2020) for q-Exponential distribution (QED) and Rao et al., (2008) for Marshall-Olkin extended Lomax distribution (MOELD).

The comparative results are given in Table 5, which indicated that the proposed ASP based on PLxD gave equivalents or smaller sample sizes with smaller PR than the sample sizes and PR that were obtained by all other plans. These encouraging results means the proposed ASP is more efficient than the ASP that considered in these comparisons and it is worth to be used by the decision makers.

Table 5. Comparative ASP $\left(n, c = 2, t / \mu_0\right)$ when μ / μ_0 =4 and $P^* = 0.95$.

t / μ_0	PLxD		QED		GBSD		MOLED		HED	
	n	PR	n	PR	n	PR	n	PR	n	PR
0.628	11	0.0220	10	0.2762	17	0.0351	12	0.2464	16	0.4170
0.942	7	0.0420	7	0.2665	11	0.0657	9	0.2842	11	0.4152
1.257	6	0.0939	6	0.3104	9	0.1159	7	0.2789	9	0.4543
1.571	5	0.1324	5	0.2980	7	0.1289	6	0.2929	7	0.4097
2.356	4	0.2567	5	0.5221	6	0.2699	5	0.3752	6	0.5486
3.141	4	0.4811	4	0.4846	5	0.3332	5	0.5420	5	0.5821
3.927	4	0.6587	4	0.6123	5	0.4896	4	0.4650	4	0.5194
4.712	4	0.7780	4	0.7113	4	0.4106	4	0.5662	4	0.6378

6. Real Data Application

Lifetime data measured in months of 20 small electric carts used by the manufacturing company for internal transportation and delivery services in a large manufacturing facility are used to illustrate the proposed ASP. The data are given as follows (Zimmer et al., 1998; Lio et al., 2010; Al-Omari et al., 2018): 0.9, 1.5, 2.3, 3.2, 3.9, 5.0, 6.2, 7.5, 8.3, 10.4, 11.1, 12.6, 15.0, 16.3, 19.3, 22.6, 24.8, 31.5, 38.1 and 53.0.

First, we need to test whether PLxD can be used or not. Several goodness of fit criteria were used to test if the data fit the model, including the minimum value of the function -log (likelihood) (-2MLL), Cramér-von Mises (CvM), Akaike information criteria (AIC), Bayesian information criteria (BIC), consistent Akaike information criteria (CAIC), Hannan-Quinn information criteria (HQIC) and two distribution tests; K-S and Anderson-Darling (A-D). The goodness of fit results were acceptable. The results indicate an excellent fit with the K-S distance value between the empirical and the theoretical PLxD equal to 0.1579606 with P-value equal to 0.6440496.

Table 6. Information measures and goodness of fit test for the small electric carts

AIC	BIC	W	AD	-log(Likelihood)	KS (P-Value)
158.030	161.017	0.039061	0.261307	76.01396	0.157961 (0.64405)

Moreover, the maximum likelihood estimation (MLE) method is used to estimate the PLxD unknown parameters. The results are given in Table 7:

Table 7. MLE estimates based on the small electric carts data

Estimator	Value	Stdev	95% C.I	
			Lower	Upper
$\hat{\alpha}$	0.7790995	0.5357288	-0.2709095	1.829109
$\hat{\beta}$	1.3513955	0.4189634	0.5302422	2.172549
$\hat{\lambda}$	10.2523672	5.7100538	-0.9391325	21.443867

Therefore, the mean life can be estimated as:

$$\mu = \frac{\alpha\lambda^{\frac{1}{\beta}}\,\Gamma\left[\alpha-\frac{1}{\beta}\right]\Gamma\left[\frac{1+\beta}{\beta}\right]}{\Gamma[1+\alpha]} = \frac{4.360922\ (25.02012)(0.9168207)}{0.9260023} = 108.03$$

assumed T = 100 months. Therefore,

$$\frac{T}{\mu_0} = \frac{100}{108.03} = 0.925$$

based on the estimated values given in Table 7; and for $\frac{T}{\mu_0} = 0.925$; we re-evaluated the minimum sample sizes as given in Table 8:

Table 8. Minimum sample sizes for the small electric carts

c		0	1	2	3	4	5	6	7	8	9	10
P^*	*0.75*	1	2	3	4	5	7	8	9	10	11	12
	0.90	1	2	4	5	6	7	8	9	10	11	13
	0.95	1	3	4	5	6	7	9	10	11	12	13
	0.99	2	3	5	6	7	8	9	10	12	13	14

From the new results, for example, corresponds to P^* = 0.99 and $\frac{T}{\mu_0} = 0.925$, we obtained n = 14 when c = 10, therefore, the optimal acceptance sampling plan will be ASP (*14*, *10*, 0.925). Based on the given data, this means that the manufacturing company can buy only 14 small electric carts in order to complete the manufacturing process within 100 hours, even if 10 out of these 14 electric carts have a mechanical failure within the manufacturing process time; with probability equal to 0.99.

7. Conclusion

In this article, we introduce the lifetime truncated acceptance-sampling plan for the power Lomax distribution. We present the table for the smallest sample size necessary to ensure a certain mean life of the test items. The operating characteristic function values and the associated manufacturer's risks are also discussed. The comparisons results with some other lifetime distribution showed that the proposed sampling plans based on PLxD are better and more efficient to be used when it applies. Therefore, the proposed plans can be used conveniently.

References

Al-Masri A., (2018). *Acceptance sampling plans based on truncated life tests in the inverse-gamma model*, Electronic Journal of Applied Statistical Analysis, 11(1), pp. 397–404.

Al-Nasser, A. D., Obeidat, M., (2020). *Acceptance Sampling Plans from Truncated Life Test Based on Tsallis q-Exponential Distribution*. Journal of Applied Statistics, 47, 4, pp. 685–697.

Al-Nasser A. D., Al-Omari A, Bani-Mustafa A, Jaber K., (2018). *Developing Single-Acceptance Sampling Plans Based on a Truncated Lifetime Test for an Ishita Distribution*. Statistics in Transition New Series, 19 (3), pp. 393–406.

Al-Nasser, A. D., Al-Omari, A., (2013). *Acceptance Sampling Plan Based on Truncated Life Tests for Exponentiated Frechet Distribution*. Journal of Statistics and Management Systems, 16(1), pp. 13–24.

Al-Nasser, A. D., Gogah, F., (2017). *On Using the Median Ranked Set Sampling for Developing Reliability Test Plans under Generalized Exponential Distribution*. Pakistan Journal of Statistics and Operation Research, 13(4), pp. 757–774.

Al-Omari A, Ciavolino E, Al-Nasser AD., (2019). *Economic Design of Acceptance Sampling Plans for Truncated Life Tests using Three-Parameter Lindley Distribution*. Journal of Modern Applied Statistical Methods, 18(2), eP2746.

Al-Omari A., Al-Nasser A. D., Ciavolino E., (2018). *Acceptance Sampling Plans Based on Truncated Life Tests for Rama Distribution.* International Journal of Quality & Reliability Management, 36(7), pp. 1181–1191.

Al-Omari, A., (2018). *Acceptance Sampling Plans Based on Truncated Life Tests for Sushila Distribution*. Journal of Mathematical and Fundamental Sciences, 50(1), pp. 72–83.

Al-Omari, A., Al-Nasser, A. D., Gogah, F., (2016). *Double Acceptance Sampling Plan for Time Truncated Life Tests Based on Transmuted New Weibull-Pareto Distribution.* Electronic Journal of Applied Statistical Analysis, 9(3), pp. 520–529.

Al-Omari, A., Al-Nasser, A. D., Gogah, F., Haq, M. A., (2017). *On the Exponentiated Generalized Inverse Rayleigh Distribution Based on Truncated Life Tests in a Double Acceptance Sampling Plan*. Stochastics and Quality Control, 32(1), pp. 37–47.

Baklizi, A., (2003). *Acceptance Sampling Based on Truncated Life Tests in The Pareto Distribution of the Second Kind*. Advances and Applications in Statistics, 3, pp. 33–48.

Baklizi, A., El Masri, A. and Al-Nasser, A. D., (2005). *Acceptance Sampling Plans in the Rayleigh Model.* The Korean Communications in Statistics, 12(1), pp. 11–18.

Balakrishnan N, Leiva V., López J., (2007). *Acceptance Sampling Plans from Truncated Life Tests Based on the Generalized Birnbaum–Saunders Distribution*. Communications in Statistics - Simulation and Computation, 36(3), pp. 643–656.

Gui, W., Aslam, M., (2017). *Acceptance Sampling Plans Based on Truncated Life Tests for the Weighted Exponential Distribution*. Communications in Statistics-Simulation and Computation, 46(3), pp. 2138–2151.

Gupta, S. S. and Groll, P. A., (1961). *Gamma Distribution in Acceptance Sampling Based on Life Tests*. Journal of American Statistical Association, 56, pp. 942–970.

Kantam, R. R. L., Rosaiah, K. and Rao, G. S., (2001). *Acceptance Sampling Based on Life Tests Log-Logistic Model*. Journal of Applied Statistics, 28, pp. 121–128.

Lio, L. Y., Tsai, T. and Wu, S., (2010). *Acceptance Sampling Plans From Truncated Life Tests Based on the Birnbaum–Saunders Distribution for Percentiles*. Communications in Statistics-Simulation and Computation, 39, pp. 119–136.

Malathi, D., Muthulakshmi, S., (2017). *Economic Design of Acceptance Sampling Plans for Truncated Life Test Using Frechet Distribution*. Journal of Applied Statistics, 44(2), pp. 376–384.

Rady, E.-H. A., Hassanein, W. A., Elhaddad, T. A., (2016). *The Power Lomax Distribution with an Application to Bladder Cancer Data*. Springer Plus, 5(1), p. 1838.

Rao, S. G. A, Ghitany M.E., Kantam R. R. L., (2008). *Acceptance Sampling Plans for Marshall-Olkin Extended Lomax Distribution*. International Journal of Applied Mathematics. 21(2), pp. 315–325.

Rao, S. G. A., (2010). *Group Acceptance Sampling Plans Based on Truncated Life Tests for Marshall-Olkin Extended Lomax Distribution*. Electronic Journal of Applied Statistical Analysis, 3(1), pp. 18–27.

Sampath S, Lalitha S. M., (2016). *Time Truncated Sampling Plan under Hybrid Exponential Distribution*. Journal of Uncertain Systems. 10(3), pp. 181–197.

Zimmer, W. J., Keats, J. B., Wang, F. K., (1998). The Burr XII Distribution in Reliability Analysis. Journal of Quality Technology, 30, pp. 386–394.

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 15–30, DOI 10.21307/stattrans-2021-025
Received – 21.07.2020; accepted – 12.01.2021

A calibrated synthetic estimator for small area estimation

Matthew Joshua Iseh[1], **Ekaette Inyang Enang**[2]

ABSTRACT

Synthetic estimators are known to produce estimates of population mean in areas where no sampled data are available, but such estimates are usually highly biased with invalid confidence statements. This paper presents a calibrated synthetic estimator of the population mean which addresses these problematic issues. Two known special cases of this estimator were obtained in the form of combined ratio and combined regression synthetic estimators, using selected tuning parameters under stratified sampling. In result, their biases and variance estimators were derived. The empirical demonstration of the usage involving the proposed calibrated estimators shows that they provide better estimates of the population mean than the existing estimators discussed in this study. In particular, the estimators were examined through simulation under three distributional assumptions, namely the normal, gamma and exponential distributions. The results show that they provide estimates of the mean displaying less relative bias and greater efficiency. Moreover, they prove more consistent than the existing classical synthetic estimator. The further evaluation carried out using the coefficient of variation provides additional confirmation of the calibrated estimator's advantage over the existing ones in relation to small area estimation.

Key words: auxiliary variable, calibration estimation, simulation, synthetic estimation.

1. Introduction

The theory of small area estimation (SAE) revolves around the use of statistical modelling techniques to produce required estimates for several geographic sub-populations and socio demographic groups when the available survey data are not enough to calculate reliable direct estimates. The inherent challenges facing SAE revolve around finding the best statistical model to be fitted on the available data when a survey is designed for national purposes but preferably used for inferences about small areas to increase the accuracy of sub-national estimates and selecting the best

[1] Department of Statistics, Akwa Ibom State University, Mkpat Enin, Nigeria. E-mail: eeseaglechild@gmail.com. ORCID: https://orcid.org/0000-0003-2696-7319.
[2] Department of Statistics, University of Calabar, Calabar, Nigeria. E-mail: edikanenang@gmail.com. ORCID: https://orcid.org/0000-0002-1273-0436.

estimation method having known that SAE is likely to be used in the survey (Pandey and Tikkiwal 2007).

Authors like Gonzales (1973) and Sarndal (1981, 1984) have made useful contributions on the use of synthetic estimators in domains with zero/small sample sizes. Although the synthetic estimators have been shown to produce estimates for domains without sample units with an attractive property of small mean square error, it has also been noted that these estimators are sometimes characterized with large bias, hence, researchers are advised to apply caution in using this method (Sarndal, Swensson and Wretman 1992; Rao 2003; Rao and Choudhry 1995; Marker 1999).

In progression for improving the performance of small area estimators, a number of estimators have been constructed using weighted linear combination of different statistical principles like: the *empirical Bayes approach* by Fay and Herriot, (1979), *sample dependent (composite) method* by Drew Singh and Choudhry (1982), *Error Prediction approach* by Battesse and Fuller (1984). However, these techniques are identified with non-negligible bias and large MSE in areas with a small to modest sample size, which might constitute an invalid confidence interval.

In a bid to improve on the efficiency of small area estimators in the last two decades, several authors have proposed various types of estimators through the use of the calibration approach. In particular, among them are: Lundstrom and Sarndal (2001), Chambers (2006), Sarndal and Lundstrom (2007), Pfefferman (2013), Hidiroglou and Estavao (2014), Rao and Molina (2015), Clement and Enang (2017). Nevertheless, none of these works has considered improving on the post stratified synthetic estimator whose strength in producing estimates even in areas of no unit is based on the assumption of structural similarities. By keeping in mind the proposition of testability of the assumption of structural similarity of characteristics and a careful choice of auxiliary variable proposed by Rao (2003), this paper seeks to formulate synthetic estimators through calibration techniques in stratified sampling to reduce the bias and improve upon the precision of the synthetic estimators for small area.

2. Notations

Consider a finite population consisting of N units which is divided into D non-overlapping domains $U_d,\ d = 1,2,\dots,D$ with N_d units such that $\sum_d^D N_d = N$. Let the population be further partitioned into G non-overlapping groups (considered to be strata) which are considered to be larger than the domains $U_{.g},\ g = 1,2,\dots,G$ with $N_{.g}$ units such that $\sum_g^G N_{.g} = N$ so that the G groups cuts across the D domains to form a grid of DG cells denoted by U_{dg} with N_{dg} units such that $U = \bigcup_{d=1}^{D} U_d = \bigcup_{g=1}^{G} U_{.g} = \bigcup_{d=1}^{D}\bigcup_{g=1}^{G} U_{dg}$ and $N = \sum_d^D N_d = \sum_g^G N_{.g} = \sum_d^D \sum_g^G N_{dg}$. The sample s is analogously partitioned into domain subsamples s_d, group subsamples $s_{.g}$ and cells subsamples s_{dg}

with corresponding sample sizes n, n_d, $n_{.g}$ and n_{dg} as $s = \cup_{d=1}^{D} s_d = \cup_{g=1}^{G} s_{.g} = \cup_{d=1}^{D} \cup_{g=1}^{G} s_{dg}$ and $n = \sum_d^D n_d = \sum_g^G n_{.g} = \sum_d^D \sum_g^G n_{dg}$. The cells subsamples n_{dg} are assumed to be random. Ordinarily, n_d and $n_{.g}$ are also random but $n_{.g}$would be fixed if the g^{th} group is a stratum from which a fixed number of elements is drawn. Let Y be the study variable whose values y_{dgk} are known for just the element of a sample s, where $k = 1,2,\ldots,N_{dg}$ (the number of population units in the $(dg)^{th}$ cell) and X be the auxiliary variable whose values $x_{dgk} > 0$ may or may not be known *a priori* for all units in U.

Let the population mean $\bar{Y}_d$ for the domain be defined as $\bar{Y}_d = \sum_{g=1}^{G} W_{dg} \bar{Y}_{dg}$, where $\bar{Y}_{dg} = \sum_{k=1}^{N_{dg}} \frac{Y_{dgk}}{N_{dg}}$ is the population mean per $(dg)^{th}$ cell for the small area. A Horvitz-Thompson (1952)-type direct unbiased estimator of the population mean $\bar{Y}_d$ for the domain under stratified sampling is given as:

$$\hat{\bar{y}}_d = \sum_g^G W_{dg} \bar{y}_{dg} \tag{1}$$

where W_{dg} is the stratum weight given as $W_{dg} = N_d^{-1} N_{dg}$ and $\bar{y}_{dg} = \sum_{k=1}^{n_{dg}} \frac{y_{dgk}}{n_{dg}}$ is the sample mean per $(dg)^{th}$ cell for the small area. Equation (1) will perform at its best when n_d is sufficiently large as well as n_{dg}. However, under SAE, even if n_d is large, there is the likelihood that n_{dg} might turn out to be zero for some cells. Consequently, the direct estimator might be very unstable with large variance as well as lead to underestimation in areas with small sample sizes and also impossible to compute where there is no sample observation in the domain of interest. Under the aforementioned conditions, assuming that the groups g's $(g = 1,2,\ldots,G)$ are similar for small area d's $(d = 1,2,\ldots,D)$, Marker (1999) suggested the synthetic estimator of the average of characteristic Y for small area d, as:

$$\hat{\bar{y}}_d^s = \sum_g^G W_{dg} \bar{y}_{.g} \tag{2}$$

And the bias of Eq.2 given as $B(\hat{\bar{y}}_d^s) = \sum_g^G W_{dg} \left(\bar{Y}_{.g} - \bar{Y}_{dg}\right)$ with mean square error (MSE) as:

$$MSE(\hat{\bar{y}}_d^s) = \sum_g^G W_{dg}^2 \left(\frac{1}{n_{.g}} - \frac{1}{N_{.g}}\right) S_g^2 + \left[\sum_g^G W_{dg} \left(\bar{Y}_{.g} - \bar{Y}_{dg}\right)\right]^2,$$

where $\bar{y}_{.g} = \sum_d^D \sum_k^{N_{dg}} \frac{y_{dgk}}{n_{.g}}$ is the sample average of $\bar{Y}_{.g}$ (the population mean for the g^{th} subgroup) across all domains, $\bar{Y}_{dg}$ is the population mean for the d^{th} domain within the g^{th} subgroup and $S_g^2 = \left(N_{.g} - 1\right)^{-1} \sum_d^D \sum_k^{N_{dg}} \left(Y_{dgk} - \bar{Y}_{.g}\right)^2$.

It is assumed that if small areas are similar across the groups, $\bar{Y}_{.g} = \bar{Y}_{dg}$, then the synthetic estimator is almost unbiased. However, in practical terms this assumption of structural similarity of the characteristics within the groups might not hold, hence equation (2) becomes heavily biased.

To further enhance the efficiency of the direct estimator for domain estimation, Clement and Enang (2017) calibrated on equation (1) and obtained combined ratio and regression estimators for domain means under stratified sampling as follows:

$$\hat{\bar{y}}_{dc} = \sum_g^G W_{dg}\bar{y}_{dg} + \frac{\sum_g^G W_{dg}q_{dg}\bar{x}_{dg}\bar{y}_{dg}}{\sum_g^G W_{dg}q_{dg}\bar{x}_{dg}^2}\left(\bar{X} - \sum_g^G W_{dg}\bar{x}_{dg}\right). \quad (3)$$

By extension, setting $q_{dg} = \bar{x}_{dg}^{-1}$ and $q_{dg} = 1$ the authors obtained

$$\hat{\bar{y}}_{dcr} = \frac{\sum_g^G W_{dg}\bar{y}_{dg}}{\sum_g^G W_{dg}\bar{x}_{dg}}\bar{X} \quad (4)$$

as the calibration approach combined ratio estimator and

$$\hat{\bar{y}}_{dcreg} = \hat{\bar{y}}_d + b(\bar{X} - \hat{\bar{x}}_d). \quad (5)$$

as the calibration approach combined regression estimator respectively, where $\hat{\bar{x}}_d = \sum_g^G W_{dg}\bar{x}_{dg}$ is analogous to equation (1) is the domain direct estimator for the auxiliary variable, $\bar{y}_{dg}$ and $\bar{x}_{dg}$ are the cell means for the interest and auxiliary variables respectively and $b = \frac{\sum_g^G W_{dg}\bar{x}_{dg}\bar{y}_{dg}}{\sum_g^G W_{dg}\bar{x}_{dg}^2}$ is the regression coefficient.

Note: Although the estimators in equations (4) and (5) exhibited some level of improvements over that in equation (1), they perform poorly in areas with a small sample size and are impossible to compute in the domains of interest with no sample observation, hence the need for a modified synthetic estimator.

3. Calibrated synthetic estimators.

Consider the Marker (1999) synthetic estimator in equation (2) for a domain of interest with small or no sample observation. If small areas have similar characteristics as large areas (groups), it suffices to borrow strength cross-sectionally (i.e. from larger areas having similar region for the small areas). The idea here is that "the groups (strata) are a powerful factor in explaining the variance of the variables whereas the domains are not". For example, as illustrated in Section 4.1 (Real-life Data Based Evaluation), 'sex' as used in the groupings will often explain a good part of individual variations but beyond that the States (Domains) may be a weak explanatory factor, see Sarndal, Swensson and Wretman 1992. In addition, the idea of calibration allows us to borrow strength from auxiliary variable, hence, a calibrated synthetic estimator of the population mean $\bar{Y}_d$, is obtained as follows:

$$\hat{\bar{y}}_{dc}^{S*} = \sum_g^G W_{dg}^c \bar{y}_{.g} \quad (6)$$

where W_{dg}^c is the new calibration weight chosen such that the distance measure given by:

$$\Phi_1(W^c, W) = \frac{1}{2}\sum_g^G \frac{\left(W_{dg}^c - W_{dg}\right)^2}{W_{dg}q_{dg}} \quad (7)$$

is minimized subject to the calibration constraints;

$$\sum_g^G W_{dg}^c \bar{x}_{.g} = \bar{X}_d \tag{8}$$

and q_{dg} are known positive weights unrelated to W_{dg}^c called the tuning parameter. Minimizing the loss function (7) subject to the calibration constraint (8) yields the calibration weights for small area under stratified sampling as

$$W_{dg}^c = W_{dg} + \frac{(\bar{X}_d - \sum_g^G W_{dg}\bar{x}_{.g})}{\sum_g^G W_{dg} q_{dg} \bar{x}_{.g}^2} W_{dg} q_{dg} \bar{x}_{.g} \tag{9}$$

Substituting (9) into (6) gives

$$\hat{\bar{y}}_{dc}^{s*} = \sum_g^G W_{dg} \bar{y}_{.g} + \frac{\sum_g^G W_{dg} q_{dg} \bar{x}_{.g} \bar{y}_{.g}}{\sum_g^G W_{dg} q_{dg} \bar{x}_{.g}^2} \left(\bar{X}_d - \sum_g^G W_{dg} \bar{x}_{.g}\right). \tag{10}$$

Equation (10) can also be written in the form of GREG estimator:

$$\hat{\bar{y}}_{dc}^{s*} = \hat{\bar{y}}_d^s + (\bar{X}_d - \hat{\bar{x}}_d^s)\hat{B}_d^s \tag{11}$$

where $\hat{\bar{y}}_d^s$ is as defined in Eq.2 and $\hat{\bar{x}}_d^s$ analogously defined as Eq.2 for the auxiliary variable, and

$$\hat{B}_d^s = \left(\sum_g^G W_{dg} q_{dg} \bar{x}_{.g}^2\right)^{-1} \sum_g^G W_{dg} q_{dg} \bar{x}_{.g} \bar{y}_{.g}$$

3.1. Estimator of the variance of $\hat{\bar{y}}_{dc}^{s*}$

*Lemma***:** The variance of the estimator in Eq.10 with one constraint is given as

$$\hat{V}_H(\hat{\bar{y}}_{dc}^{s*}) = \sum_g^G \frac{D_{.g}(W_{dg}^c)^2}{W_{dg}^2} s_{\hat{e}_{.g}}^2 + \frac{\sum_g^G D_{.g}(W_{dg}^c)^2 Q_{dg} s_{.gx}^2}{\sum_g^G D_{.g} W_{dg}^2 Q_{dg} (s_{.gx}^2)^2} s_{\hat{e}_{.g}}^2 \left[V_{st}(\bar{x}_d^s) - \hat{V}_{st}(\bar{x}_d^s)\right].$$

Proof: Consider the estimator of variance of the combined regression estimator under stratified sampling by Sarndal (1996) given as

$$\hat{V}_c(\hat{\bar{y}}_d) = \sum_g^G \frac{D_{.g}(W_{dg}^c)^2}{W_{dg}^2} s_{\hat{e}_{.g}}^2 \tag{12}$$

where $s_{\hat{e}_{.g}}^2 = \left(n_{.g} - 1\right)^{-1} \sum_k^{n_{dg}} \hat{e}_{dgk}^2$ is the g^{th} group (stratum) sample variance, $\hat{e}_{dgk} = y_{dgk} - \bar{y}_{.g} - \hat{B}_d^s\left(x_{dgk} - \bar{x}_{.g}\right)$, W_{dg}^c as given in Eq.9 and $D_{.g} = W_{dg}^2 \gamma_{.g}$ is the initial weight of Eq.12 and $\gamma_{.g} = \frac{1}{n_{.g}} - \frac{1}{N_{.g}}$. Following the procedure by Singh and Arnab (2014), the estimate of variance of the estimator $\hat{\bar{y}}_{dc}^{s*}$ obtained by calibrating on Eq.12 is given as

$$\hat{V}_H(\hat{\bar{y}}_{dc}^{s*}) = \sum_g^G \frac{\Omega_{.g}(W_{dg}^c)^2}{W_{dg}^2} s_{\hat{e}_{.g}}^2 \tag{13}$$

where $\Omega_{.g}$ is the new weights chosen such that the chi-square distance function

$$\Phi_2 = \sum_g^G \frac{\left(\Omega_{.g} - D_{.g}\right)^2}{D_{.g}Q_{dg}} \tag{14}$$

is minimized subject to the calibration constraint

$$\sum_g^G \Omega_{.g} s_{.gx}^2 = V_{st}(\bar{x}_d^s) \tag{15}$$

where Q_{dg} is the tuning parameter unrelated with $\Omega_{.g}$. Hence, the calibration weight is obtained as

$$\Omega_{.g} = D_{.g} + \frac{D_{.g}Q_{dg}s_{.gx}^2}{\sum_g^G D_{.g}Q_{dg}\left(s_{.gx}^2\right)^2}\left[V_{st}(\bar{x}_d^s) - \sum_g^G D_{.g}s_{.gx}^2\right]. \tag{16}$$

Substituting (16) in (13) gives

$$\hat{V}_H(\hat{\bar{y}}_{dc}^{s*}) = \sum_g^G \frac{D_{.g}\left(W_{dg}^c\right)^2}{W_{dg}^2} s_{\hat{e}.g}^2 + \frac{\sum_g^G D_{.g}\left(W_{dg}^c\right)^2 Q_{dg}s_{.gx}^2}{\sum_g^G D_{.g}W_{dg}^2 Q_{dg}\left(s_{.gx}^2\right)^2} s_{\hat{e}.g}^2\left[V_{st}(\bar{x}_d^s) - \hat{V}_{st}(\bar{x}_d^s)\right] \tag{17}$$

where $V_{st}(\bar{x}_d^s) = \sum_g^G D_{.g}S_{.gx}^2$ is assumed to be known variance of $\bar{X}_d$ and $s_{.gx}^2 = \frac{1}{n_{.g}-1}\sum_d^D \sum_k^{n_{dg}}\left(x_{dgk} - \bar{x}_{.g}\right)^2$ is an unbiased of $S_{.gx}^2 = \frac{1}{N_{.g}-1}\sum_d^D \sum_k^{n_{dg}}\left(X_{dgk} - \bar{X}_{.g}\right)^2$.

Eq (17) can further be written as

$$\hat{V}_H(\hat{\bar{y}}_{dc}^{s*}) = \hat{V}_c(\hat{\bar{y}}_d) + \hat{B}_{dc}\left[V_{st}(\bar{x}_d^s) - \hat{V}_{st}(\bar{x}_d^s)\right] \tag{18}$$

where $\hat{V}_c(\hat{\bar{y}}_d) = \sum_g^G \frac{D_{.g}\left(W_{dg}^c\right)^2}{W_{dg}^2} s_{\hat{e}.g}^2$ as earlier defined in Eq.12 and $\hat{B}_{dc} = \frac{\sum_g^G D_{.g}\left(W_{dg}^c\right)^2 Q_{dg}s_{.gx}^2}{\sum_g^G D_{.g}W_{dg}^2 Q_{dg}\left(s_{.gx}^2\right)^2} s_{\hat{e}.g}^2$.

3.2. Combined ratio synthetic estimator

Here, we consider special cases of the estimator in Eq.10.

Case 1: Suppose we set the tuning parameter $q_{dg} = \bar{x}_{.g}^{-1}$ in Eq.10, then

$$\hat{\bar{y}}_{dcr}^{s*} = \frac{\sum_g^G W_{dg}\bar{y}_{.g}}{\sum_g^G W_{dg}\bar{x}_{.g}}\bar{X}_d. \tag{19}$$

The approximate form of the bias of Eq.19 is obtained through Taylor's series approximation method. Equation (19) can be written as

$$\hat{\bar{y}}_{dcr}^{s*} = \frac{\hat{\bar{y}}_d^s}{\hat{\bar{x}}_d^s}\bar{X}_d = \hat{R}_d^s\bar{X}_d \tag{20}$$

$$\hat{\bar{y}}_{dcr}^{s*} - \bar{Y}_d = \bar{X}_d\left(\hat{R}_d^s - R_d\right), \text{where } R_d = \frac{\bar{Y}_d}{\bar{X}_d}.$$

The bias $B(\hat{\bar{y}}_{dcr}^{s*}) = E(\hat{\bar{y}}_{dcr}^{s*} - \bar{Y}_d) = E\left[\bar{X}_d\left(\hat{R}_d^s - R_d\right)\right]$

$$B(\hat{\bar{y}}_{dcr}^{s*}) = \bar{X}_d E\left[\frac{1}{\bar{x}_d^s}(\bar{y}_d^s - R_d\bar{x}_d^s)\right]. \tag{21}$$

But $\frac{1}{\bar{x}_d^s} = \frac{1}{\bar{X}_d}\left[1 + \frac{\bar{x}_d^s - \bar{X}_d}{\bar{X}_d}\right]^{-1}$, such that Taylor's series expansion to the first order approximation gives $\frac{1}{\bar{x}_d^s} = \frac{1}{\bar{X}_d}\left[1 - \frac{\bar{x}_d^s - \bar{X}_d}{\bar{X}_d}\right]$, then Equation (21) becomes $B(\hat{\bar{y}}_{dcr}^{s*}) = \bar{X}_d E\left[\frac{1}{\bar{X}_d}\left(1 - \frac{\bar{x}_d^s - \bar{X}_d}{\bar{X}_d}\right)(\bar{y}_d^s - R_d\bar{x}_d^s)\right]$

$$B(\hat{\bar{y}}_{dcr}^{s*}) = \frac{1}{\bar{X}_d}\sum_g^G W_{dg}^2\gamma_{.g}\left(R_d S_{.gx}^2 - S_{.gxy}\right) \tag{22}$$

where, $S_{.gxy} = \frac{1}{N_{.g}-1}\sum_d^D\sum_k^{N_{dg}}\left(X_{dgk} - \bar{X}_{.g}\right)\left(Y_{dgk} - \bar{Y}_{.g}\right)$ is estimated by

$$s_{.gxy} = \frac{1}{n_{.g}-1}\sum_d^D\sum_k^{n_{dg}}\left(x_{dgk} - \bar{x}_{.g}\right)\left(y_{dgk} - \bar{y}_{.g}\right)$$

and the bias estimator of $\hat{\bar{y}}_{dcr}^{s*}$ is given as;

$$\hat{B}(\hat{\bar{y}}_{dcr}^{s}) = \frac{1}{\bar{X}_d}\sum_g^G W_{dg}^2\gamma_{.g}\left(\hat{R}_d s_{.gx}^2 - s_{.gxy}\right). \tag{23}$$

To obtain the estimator of the variance for Eq.19, we set $q_{dg} = \bar{x}_{.g}^{-1}$, $Q_{dg} = s_{.gx}^{-2}$ and replaced $s_{\hat{e}.g}^2$ by $s_{\hat{e}.gr}^2$ in Eq.17 as;

$$\hat{V}_{st}(\hat{\bar{y}}_{dcr}^{s*}) = \left(\frac{\bar{X}_d}{\bar{x}_d^s}\right)^2\left[\frac{V_{st}(\bar{x}_d^s)}{\hat{V}_{st}(\bar{x}_d^s)}\right]\sum_g^G W_{dg}^2\gamma_{.g}s_{\hat{e}.gr}^2 \tag{24}$$

Equation (24) is in the form of the ratio-type estimator proposed by Wu & Deng (1983) for estimating variance of the combined ratio estimator. The difference here is that it makes use of extra knowledge of known variance of the auxiliary variable at the estimation stage, where

$$s_{\hat{e}.gr}^2 = s_{.gy}^2 + \hat{R}_d^2 s_{.gx}^2 - 2\hat{R}_d s_{.gxy}$$

3.3. Combined regression synthetic estimator

Case 2: Again, we set the tuning parameter $q_{dg} = 1$ in Eq.10, then the combined regression-synthetic estimator in stratified sampling is given as

$$\hat{\bar{y}}_{dcREG}^{s*} = \bar{y}_d^s + \hat{b}_{.g}^*(\bar{X}_d - \bar{x}_d^s) \tag{25}$$

where $\hat{b}_{.g}^* = \frac{\sum_g^G W_{dg}\bar{x}_{.g}\bar{y}_{.g}}{\sum_g^G W_{dg}\bar{x}_{.g}^2}$ is the synthetic regression coefficient of the domain. The estimator in Eq.25 is in the form of Hansen, Hurwitz and Madow (1953) combined regression estimator. The bias of Eq.25 is obtained by replacing R_d by $\hat{b}_{.g}^*$ in Eq.22 such that

$$B(\hat{\bar{y}}_{dcREG}^{s*}) = \frac{1}{\bar{X}_d}\sum_g^G W_{dg}^2\gamma_{.g}\left(\hat{b}_{.g}^* S_{.gx}^2 - S_{.gxy}\right) \tag{26}$$

and its estimator is

$$\hat{B}(\hat{\bar{y}}_{dcREG}^{S*}) = \frac{1}{\bar{X}_d}\sum_g^G W_{dg}^2 \gamma_{.g}\left(\hat{b}_{.g}^{*} s_{.gx}^2 - s_{.gxy}\right). \quad (27)$$

An estimator of variance of the calibration approach combined regression synthetic estimator $\hat{\bar{y}}_{dcREG}^{S*}$ is obtained by setting $q_{dg} = \bar{x}_{.g}^{-1}$ and $Q_{dg} = s_{.gx}^{-2}$ in Eq.17 and replacing $s_{\hat{e}.g}^2$ with $s_{\hat{e}.greg}^2$ as

$$\hat{V}_{st}(\hat{\bar{y}}_{dcreg}^{S}) = \left(\frac{\bar{X}_d}{\bar{x}_d^s}\right)^2 \left[\frac{V_{st}(\bar{x}_d^s)}{\bar{V}_{st}(\bar{x}_d^s)}\right] \sum_g^G W_{dg}^2 \gamma_{.g} s_{\hat{e}.greg}^2 \quad (28)$$

where

$$s_{\hat{e}.greg}^2 = s_{.gy}^2 + \hat{b}_d^2 s_{.gx}^2 - 2\hat{b}_d s_{.gxy} \quad (29)$$

4. Data and methods for empirical evaluation

In this section, data and methods for empirical evaluation of the proposed and existing estimators are discussed. Real-life and simulated data are used. When domains of interest have no sample observation, the existing calibration estimator becomes impossible to compute for the area of interest. This was illustrated using real-life data as shown in Table 1. This will help to validate the theoretical claims. However, on a general note, simulation analysis was done using R-software to ascertain the level of performance of both existing and suggested estimators.

4.1. Real-life data based evaluation

The real-life data used in this analysis were obtained from the population of the household finances and consumptions survey (HFCS) conducted in 2017 by the Statistics Department of the Central Bank of Nigeria. The population comprised of 2986 male and female heads of household. The population was partitioned into two strata of male and female household heads with subpopulation sizes of 1625 and 1361, respectively, across the 37 domains (States).

To illustrate an ideal situation of small area estimation, the study variable y is considered as the household expenditure and the auxiliary variable x as the household income. The object is to estimate the mean of y for all the 37 domains. To compute the estimates for all the domains, the proportional allocation procedure was applied and a sample s of size 10% was drawn from each stratum (group) using simple random sample without replacement (SRSWOR) with the cells averages on bold points as shown in Appendix A. The results are shown in Table 1.

Table 1. Estimators of mean expenditures for domains using existing and proposed estimators

DOMAIN	$\hat{\bar{y}}_{dcr}$	$\hat{\bar{y}}_{dcreg}$	$\hat{\bar{y}}_{d}^{s}$	$\hat{\bar{y}}_{dcr}^{s*}$	$\hat{\bar{y}}_{dcREG}^{s*}$	$\bar{Y}_d$
ABIA	21393.61	21392.2	24061.07	23131.81	23143.91	21366.1
ADAMAWA	21861.68	21856.98	25267.49	36151.67	36035.95	34303.16
AKWA IBOM	19173.4	19176.94	24444.22	34921.18	34791.54	26161.88
ANAMBRA	18401.87	18392.88	24112.91	25195.98	25181.96	23548.92
BAUCHI	12556.76	12557.61	25003.59	15856.92	15959.68	15773.49
BAYELSA	6634.7	6634.7	24343.82	29799.31	29730.82	26115.66
BENUE	31350.08	31345.96	24298.10	37359.16	37194.12	30794.01
BORNO	22672.94	22753.52	25484.83	21287.21	21329.65	20671.39
CROSS R	12685.94	12682.54	23830.67	17343.67	17430.28	17085.43
DELTA	44291.46	44272.31	24021.61	46528.64	46234.19	41404.02
EBONYI	20677.67	20677.67	24343.82	23069.98	23085.97	22713.72
EDO	25419.3	25412.72	24063.92	34683.85	34545.6	30517.07
EKITI	38276.63	38289.31	24821.58	30122.35	30060.72	30346.33
ENUGU	27117.48	27132.43	23628.31	22174.25	22194.02	25256.79
FCT	7508.05	7508.05	25636.96	26374.9	26367.72	24184.29
GOMBE	33280.32	33364.8	24763.67	27232.24	27203.23	29613.17
IMO	39268.29	39190.17	23934.60	30677.41	30588.34	31019.32
JIGAWA	21597.66	21597.22	24824.90	25092.69	25089.58	24196.47
KADUNA	38844.85	38844.88	25463.42	28631.39	28599.19	27648.58
KANO	19557.45	19636.3	25313.68	27801.67	27775.49	30336.33
KATSINA	21793.15	21797.54	24647.22	27130.71	27100.95	28684.67
KEBBI	35688.25	35696.3	24393.48	35026.58	34894.02	33024.15
KOGI	22144.18	22147.94	25059.34	18667.27	18738.29	20948.49
KWARA	27003.13	27003.29	24763.67	31499.15	31420.01	34756.05
LAGOS	31918.62	31948.07	25136.32	43977.99	43771.93	52055.29
NASARA	56889.76	56929.3	23574.10	25923.61	25891.53	28630.6
NIGER	24252.95	24241.96	25095.65	35049.99	34940.2	33504.88
OGUN	47334.04	47321.89	24544.62	35066.09	34937.89	34929.4
ONDO	NA	NA	25239.92	**33167.93**	**33083.13**	**33375.22**
OSUN	14038.37	14038.32	24005.90	14977.6	15095.92	16472.88
OYO	26364.26	26364.43	24120.36	29933.23	29858.07	29499.42
PLATEAU	25735.76	25735.58	24456.47	24168.14	24171.7	24785.42
RIVERS	27626.89	27633.98	25313.11	33789.51	33700.3	35138.01
SOKOTO	10216.44	10216.44	27422.72	16539.78	16591.33	18813.12
TARABA	9306.65	9306.65	25036.58	43455.52	43249.94	46880.2
YOBE	14446.75	14446.75	27251.67	26217.78	26223.21	27139.53
ZAMFARA	20969.1	20968.26	24586.90	19218.79	19283.76	21479.42
AVERAGE	24952.73	24284.21	24765.17	**28574.22**	**28526.87**	**28464.13**

4.2. Simulation Study

Here, the procedure of population generation and sample selection by Hidiroglou and Estevao (2014) was adopted. Bivariate observations (x_{ij}, y_{ij}) were generated to comprise finite population of size 4950 units. The population U considered was created by generating data for three separate subsets of the populations termed *groups* (strata) with different intercepts and slopes. Each group was split into ten domains that are mutually exclusive and exhaustive, as follows: *Group 1;* $U_{11}, U_{21}, \ldots, U_{101}$, *Group 2;* $U_{12}, U_{22}, \ldots, U_{102}$, and *Group 3;* $U_{13}, U_{23} \ldots, U_{103}$. The number of units in each cell N_{dg} was sequentially allocated in a monotonic manner: cell U_{11} with 20 units; cell U_{21} with 30 units; and cell U_{103} with 310 units. The values of x in each group were generated from three different distributions, *Gamma* $(\alpha = 5, \beta = 10)$, *Norm* (5,1) and *Exp* (1.5) distributions. The simulation for the variable of interest y was obtained using the model $y_{dk} = \beta_{og} + \beta_{1g} x_{dk} + v_d + e_{dk}$, where $d = 1,2,\ldots,30;\ k = 1,2,\ldots,N$ and $g = 1,2,3;\ e_{dk} \sim N(0, C_{dk}^2 \sigma_e^2),\ v_d \sim N(0, \sigma_v^2)$. It is assumed that $\sigma_e^2 = \sigma_v^2 = 20^2 = 400$ for the gamma distribution, $\sigma_e^2 = \sigma_v^2 = 1^2 = 1$ for normal and exponential distributions. $c_{dk} = x_{dk}$ is set to reflect the heterogeneity of the model errors for the synthetic and calibration estimators.

4.2.1. Simulation results

The summary of the representation of units in each group across the domains is presented in Table 2 and 3. Table 2 shows how the population was split into the three groups with the respective values of intercepts and slopes for the Gamma, Normal and Exponential distributions. Table 3 illustrates the population under study divided into domains and further partitioned into groups that are larger than the domains and cut across the domains to form grids that are mutually exclusive and exhaustive. The result of the simulation study using R software for selection of independent samples of sizes $n = 248(5\%)$, $n = 495(10\%)$, $n = 744(15\%)$, $n = 990(20\%)$, $n = 1239(25\%)$ drawn using SRSWOR from U and the computation of various estimates is presented in Table 4.

Summary statistics of the simulated data will be done using Average Percent Absolute Relative Bias, Average Percent Relative Efficiency and Average Percent Coefficient of Variation $\%\overline{ARB}$, $\%\overline{RE}$ and $\%\overline{CV}$ respectively, and are obtained as

$$\%\overline{ARB}(\hat{\bar{y}}_{dP}) = \left[\frac{1}{D}\sum_{d=1}^{D} ARB(\hat{\bar{y}}_{dP})\right] \times 100,$$

where $ARB(\hat{\bar{y}}_{dP}) = \left|\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{\bar{y}}_{dP}^{(r)}}{\bar{Y}_d} - 1\right)\right|$

$$\%\overline{RE}(\hat{\bar{y}}_{dP}) = \left[\frac{\overline{MSE}(\hat{\bar{y}}_{dE})}{\overline{MSE}(\hat{\bar{y}}_{dP})}\right] \times 100,$$

where $\overline{MSE}(\hat{\bar{y}}_{dP}) = \frac{1}{D}\sum_{d=1}^{D} MSE(\hat{\bar{y}}_{dP})$ and

$$MSE(\hat{\bar{y}}_{dP}) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{y}}_{dp}^{(r)} - \bar{Y}_d\right)^2$$

$$\%\overline{CV}(\hat{\bar{y}}_{dP}) = \left[\frac{1}{D}\sum_{d=1}^{D} CV(\hat{\bar{y}}_{dP})\right] \times 100, \text{ where } CV(\hat{\bar{y}}_{dP}) = \frac{\sqrt{MSE(\hat{\bar{y}}_{dP})}}{\bar{Y}_d}$$

where $\hat{\bar{y}}_{dP}^{(r)}$ and $\hat{\bar{y}}_{dE}^{(r)}$ denote, say, the proposed and existing estimators respectively, produced for the r^{th} sample, $r = 1,2,\ldots,R$, and for each small area $d = 1,2,\ldots,D$. For each selected sample in each simulation run $= 1,2,\ldots,R$ ($R = 100{,}000$), we shall compute estimates of $\overline{Y}_d$ for the estimators.

Note: In small area estimation, Molina and Rao (2010) suggested a benchmark value for $\%\overline{CV}(\hat{\bar{y}}_{dP})$ at 20-25% as being reliable. As a result, a high value of $\%\overline{CV}(\hat{\bar{y}}_{dP})$ above 25% is considered as unreliable estimates while estimators with values of $\%\overline{CV}(\hat{\bar{y}}_{dP})$ below 25% are considered reliable and suitable for SAE.

Table 2. Partitioning the Population into Groups with their Respective Slopes and Intercepts under Gamma, Normal and Exponential Distributions

Distributions		Gamma		Normal and Exponential	
Group (g)	Cells in groups	β_{0g}	β_{1g}	β_{0g}	β_{1g}
1	U_{d1} for $k = 1,2,\ldots,10$	200	30	5	1.5
2	U_{d2} $for\ k = 11,\ldots,20$	300	20	10	1.0
3	U_{d3} $for\ k = 22,\ldots,30$	400	10	15	0.5

Table 3. Summary of Splitting the Population into Cells, Groups and Domains

Domain number (d)	Group Number (g)			Domains (U_d)
	1	**2**	**3**	
1	U_{11}	U_{12}	U_{13}	U_1
2	U_{21}	U_{22}	U_{23}	U_2
3	U_{31}	U_{32}	U_{33}	U_3
4	U_{41}	U_{42}	U_{43}	U_4
5	U_{51}	U_{52}	U_{53}	U_5
6	U_{61}	U_{62}	U_{63}	U_6
7	U_{71}	U_{72}	U_{73}	U_7
8	U_{81}	U_{82}	U_{83}	U_8
9	U_{91}	U_{92}	U_{93}	U_9
10	U_{101}	U_{102}	U_{103}	U_{10}
Groups ($\boldsymbol{U_{.g}}$)	$U_{.1}$	$U_{.2}$	$U_{.3}$	U

Table 4. Result of Simulation Evaluation for Gamma (5,10), Norm (5,1) and Exp (1.5)

Sample size	Distribution	$\%\overline{ARB}$			$\%\overline{RE}$			$\%\overline{CV}$		
		$\hat{\bar{y}}_d^S$	$\hat{\bar{y}}_{dcr}^{S*}$	$\hat{\bar{y}}_{dcREG}^{S*}$	$\hat{\bar{y}}_d^S$	$\hat{\bar{y}}_{dcr}^{S*}$	$\hat{\bar{y}}_{dcREG}^{S*}$	$\hat{\bar{y}}_d^S$	$\hat{\bar{y}}_{dcr}^{S*}$	$\hat{\bar{y}}_{dcREG}^{S*}$
5%	Gamma	65.7	13.4	13.7	100	2350	2360	65.7	13.4	13.7
	Normal	73.2	5.5	5.5	100	12832.7	12845.9	73.2	5.5	5.5
	Exponential	74.2	14.2	**61.7**	100	2451.9	**145.5**	74.2	15.3	**61.8**
10%	Gamma	65.7	13.3	13.4	100	4150	4110	65.7	13.3	13.4
	Normal	71.9	5.5	5.5	100	12786.6	12846.4	71.9	5.5	5.5
	Exponential	74.2	13.6	**61.7**	100	2383.3	**145.4**	74.2	14.6	**61.8**
15%	Gamma	65.7	13.3	13.4	100	4590	4520	65.7	13.3	13.4
	Normal	71.9	5.5	5.5	100	12770.3	12844.3	71.9	5.5	5.5
	Exponential	74.2	13.7	**61.8**	100	2360.8	**145.3**	74.2	14.4	**61.8**
20%	Gamma	65.7	13.2	13.7	100	4840	4740	65.7	13.2	13.7
	Normal	71.9	5.5	5.5	100	12776.9	12859.7	71.9	5.5	5.5
	Exponential	74.2	13.7	**61.8**	100	2358.7	**145.3**	74.2	14.8	**61.8**
25%	Gamma	65.7	13.2	13.8	100	4750	4750	65.7	13.2	13.8
	Normal	71.9	5.5	5.5	100	12767.3	12853.1	71.9	5.5	5.5
	Exponential	74.2	13.7	**61.8**	100	2353.9	**145.3**	74.2	14.1	**61.8**

5. Discussion of results

From Table 1, it can be seen that the domain called ONDO has no estimate under the existing calibration estimators $\hat{\bar{y}}_{dcr}$ and $\hat{\bar{y}}_{dcreg}$ because there was no sampled unit selected for that domain and the estimates of the population mean could not be computed. This agrees with Purcell and Kish (1979) and Rao (2003), that in areas without sample observations, the direct estimator could not be computed. However, the synthetic estimators (both existing and proposed) $\hat{\bar{y}}_d^S$, $\hat{\bar{y}}_{dcr}^{S*}$ and $\hat{\bar{y}}_{dcREG}^{S*}$ produced estimates (of 25239.92, **33167.93 and 33083.13 naira** respectively) for the average population expenditure (of **33375.22 naira)** for ONDO. This agrees with Rao (2003) proposition on testability of the assumption of structural similarities of characteristics and a careful choice of auxiliary variable in the use of synthetic estimators. It could also be seen that, although the existing synthetic estimator $\hat{\bar{y}}_d^S$ produced estimate for ONDO where there are no sample units, the value obtained underestimated the population mean expenditure of the domain compared to the proposed synthetic estimators that made use of additional supplementary information. Furthermore, on average, the mean expenditures of all the domains produced by the proposed synthetic estimators $\hat{\bar{y}}_{dcr}^{S*}$

and $\hat{\bar{y}}_{dcREG}^{S*}$ as **28574.22** and **28526.87 naira** respectively, are almost the same as the population mean expenditure (**28464.13 naira)** compared to that obtained from the existing synthetic and calibrated estimators. This agrees with the appealing property of the domain estimator (that the sum of the domains estimates will equal the population parameter) as suggested in the literature by Lundstrom and Sarndal (2001). These results confirm the need to borrow strength cross-sectionally in addition to a highly correlated auxiliary variable. In this case, the proposed synthetic estimator gained dominance over the existing synthetic and direct estimators in domains of interest where there are no sample observations.

Results of analysis in Table 4 showed values of $\%\overline{ARB}$ between 5.5% to 14.2% and 5.5% to 61.8% for $\hat{\bar{y}}_{dcr}^{S*}$ and $\hat{\bar{y}}_{dcREG}^{S*}$ respectively, while that of the existing synthetic estimator $\hat{\bar{y}}_{d}^{S}$ was 65.7% to 71.9%. From the result, the proposed synthetic estimators have been found to exhibit a remarkably smaller $\%\overline{ARB}$ than the existing synthetic estimator for all the probability distributions and in all sample sizes under study.

It was further observed that under normal distribution, the proposed estimators $\hat{\bar{y}}_{dcr}^{S*}$ and $\hat{\bar{y}}_{dcREG}^{S*}$ have a constant $\%\overline{ARB}$ of 5.5%, which is regarded as the least for all sample sizes. Under gamma distribution, they recorded between 13.2% to 13.8%. However, under exponential distribution, the $\%\overline{ARB}$ values of $\hat{\bar{y}}_{dcr}^{S*}$ lies between 13.6% to 14.4% while that of $\hat{\bar{y}}_{dcREG}^{S*}$ was seen to be highly biased with 61.7% to 61.8% as indicated in column 5 of Table 4 with bold points. This result conforms to an established fact in the literature by Clement and Enang (2017) that under domain estimation, the calibration approach combined ratio estimator outperforms the combined regression estimator. In addition, this result suggests that for real life data that follow exponential distribution, the proposed combined ratio is more preferred to the combined regression estimator.

From Table 4, the proposed synthetic estimators $\hat{\bar{y}}_{dcr}^{S*}$ and $\hat{\bar{y}}_{dcREG}^{S*}$ were observed to have higher gains in efficiency than the existing estimator $\hat{\bar{y}}_{d}^{S}$ in all sample sizes for all the three probability distributions considered. Contrary to popular claims that the existing synthetic estimator produced estimates for domains without sample units with a very small mean square error, the proposed synthetic estimators have been shown to be more efficient and superior to the existing estimator. The $\%\overline{RE}$ of the proposed synthetic estimators for the three distributions were observed to be between 12770.3% to 12894.3% for normal distribution followed by gamma distribution with 2350% to 4840% and the exponential distribution between 145.5% to 2451.9%. Again, as expected, $\hat{\bar{y}}_{dcr}^{S*}$ was clearly more efficient than $\hat{\bar{y}}_{dcREG}^{S*}$ in all sample sizes under exponential distribution, which supports the results in the literature by Clement and Enang (2017). Furthermore, the gain in efficiency of the proposed $\hat{\bar{y}}_{dcr}^{S*}$ for the three distributions and in all sample sizes considered is an improvement in SAE contrary to

the results in the literature by Rao and Choudhry (1996) on the instability of ratio synthetic MSE.

Again, the result in Table 4 showed that $\hat{\bar{y}}_{dcr}^{S*}$ has the smallest percent average coefficient of variation of 5.5% to 15.3% while $\hat{\bar{y}}_{dcREG}^{S*}$ has 5.5% to 61.8% against $\hat{\bar{y}}_{d}^{S}$ with $\%\overline{CV}$ of 65.7% to 71.9% for all the sample sizes considered in this work. This suggest that the proposed synthetic estimators are highly preferred for small area estimation having met the required benchmark of falling below 25% as suggested by Molina and Rao (2010) but the same could not be said about the existing synthetic estimator. It could be said that synthesizing on the approaches of borrowing strength cross-sectionally and through calibration has been profitable in this study. It was further observed that, for exponential distribution, the combined regression synthetic estimator $\hat{\bar{y}}_{dcREG}^{S*}$ has a constant $\%\overline{CV}$ as high as 61.8% for all sample sizes. This is an indication that $\hat{\bar{y}}_{dcREG}^{S*}$ is not suitable for any real-life data that follow exponential distribution for small domains under stratified sampling. It is convenient to say that the proposed combined ratio has an edge over the combined regression synthetic estimator under exponential distribution in small area estimation. The normal distribution produced a constant value of $\%\overline{CV}$ of 5.5% and is seen as the smallest for all the sample sizes followed by Gamma with 13.2% to 13.8% and exponential with 14.4% to 61.8% for the proposed synthetic estimators. This points to the desired qualities of the Normal distribution in small area estimation under stratified sampling design.

6. Conclusion

The synthetic estimation technique has been shown to be the only remedy if no sampled units are available in some domains of interest as shown by the result of this study. Therefore, it can be concluded that the proposed small area estimators (which borrowed strength cross-sectionally and with auxiliary variable) are an improvement over the Marker (1999) synthetic estimator (that only borrowed strength cross-sectionally) and the calibration approach direct estimators for the estimation of population mean in areas that are characterized by small/no sample sizes. Also, the proposed combined ratio synthetic estimator has shown dominance over the combined regression synthetic estimator suggesting that the latter is not suitable for any real-life data that follow exponential distribution for small domains under stratified sampling.

Acknowledgement

The author is grateful to the anonymous Reviewers who painstakingly read this manuscript and made useful contributions to see the paper through this stage.

References

Battese, G. E., Fuller, W. A., (1984). An Error Components Model for Predictions of County Crop Areas using Survey and Satellite Data. Survey section, Statistics Laboratory Iowa State University, Ames.

Chambers, R. L., (2006). Small Area Estimation for Business Surveys. Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 2803–2809.

Clement, E. P., Enang, E. I., (2017). On the Efficiency of Ratio Estimators over Regression Estimators, *Communication in Statistics - Theory and Methods*, Vol. 46, pp. 5357–5367.

Deville, J. C., Sarndal, C. E., (1992). Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, Vol. 87, pp. 376–382.

Drew, J. D., Singh, M. P., Choudhry, G. H., (1982). Evaluation of Small Area Techniques for the Canadian Labor Force Survey, *Survey Methodology*, Vol. 8, pp. 17–47.

Fay, R. E., Herriot, R. A., (1979). Estimates of Income for Small Places. An Application of James-Stein Procedures to Census Data, *Journal of the American Statistical Association*, Vol. 74, pp. 269–277.

Gonzales, M. E., (1973). Use and Evaluation of Synthetic Estimator, Proceedings of the Section on Social statistics, American statistical Association, pp. 33–36.

Hansen, M. H., Hurwitz, W. N., Madow, W. G., (1953). Sample Survey Methods and Theory, John Wiley and Sons.

Hidiroglou, M. A., Estevao, V. M., (2014). A Comparison of Small Area and Calibration Estimators Via Simulation, *Statistics in Transition*, New Series, Vol. 17, pp. 133–154.

Holt, D., Smith, T. M. F., Tomberlin, T. J., (1979). A Model Based Approach to Estimation for Small subgroups of a Population, *Journal of the American Statistical Association*, Vol. 74, pp. 405–410.

Horvitz, D. G., Thompson, D. J., (1952). A Generalization of Sampling without Replacement from a Finite Universe, *Journal of the American Statistical Association,* Vol. 47, pp. 663–687.

Lundstrom, S., Sarndal, C. E., (2001). Estimation in the presence of Nonresponse and Frame Imperfections, Statistics Sweden.

Marker, D. S., (1999). Organization of Small Area Estimation. Using Generalized Linear Regression Framework, *Journal of Official Statistics*, Vol. 15, pp. 1–24.

Molina, I., Rao, J. N. K., (2010). Small Area Estimation Poverty Indicators, *Canadian Journal of statistics,* Vol. 38, pp. 369–385.

Pfeffermann, D., (2013). New Important Developments in Small Area Estimation, *Statistical Sciences*, Vol. 28, pp. 40–68.

Purcell, N. J., Kish, L., (1979). Estimation for Small Domains, *Biometrics*, Vol. 35, pp. 365–354.

Rao, J. N. K., (2003). Small Area Estimation, Wiley, New York.

Rao, J. N. K., Molina, I. (2015). Small Area, Second Edition. John Wiley, New York.

Sarndal, C. E., Swensson. B., Wretman, J., (1992). Model-Assisted Surveys, New York: Springer-Verlag.

Sarndal, C. E., (1981). When Robust Estimation is not an Obvious Answer: The case of the Synthetic Estimator versus Alternatives for Small Areas, Proceedings of the American Statistical Association, Survey Research Section, pp. 53–59.

Sarndal, C. E., (1984). Design-Consistent Versus Model Dependent Estimation for Domains, Journal *of the American Statistical Association*, Vol. 79, pp. 624–637.

Sarndal, C. E., (1996). Efficient Estimators with Simple Variance in Unequal Probability Sampling, *Journal of the American Statistical Association*, Vol. 91, pp. 1289–1300.

Sarndal, C. E. (2007). The Calibration Approach in Survey Theory and Practice. Survey Methodology, Vol. 33, pp. 99–119.

Singh, S., Arnab, R., (2014). On Calibration of design weights, *International Journal of Statistics*. Vol. 69, pp. 185–205.

Tikkiwal, G. C., Pandey, K. K., (2007). On Synthetic and Composite Estimators for Small Area Estimation under Lahiri-Midzuno Sampling Scheme, *Statistics in Transition-new Series, Poland*, Vol. 8, pp. 111–123.

Wu, C. F., Deng, L. Y., (1983). Estimation of Variance of the Ratio Estimator; An Empirical Study, In: G. E. P. Box et al. (ed.), Scientific Inference, Data Analysis and Robustness. New York, Academic Press.

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 31–57, DOI 10.21307/stattrans-2021-026
Received – 21.11.2019; accepted – 13.04.2021

The Complex-Number Mortality Model (CNMM) based on orthonormal expansion of membership functions

Andrzej Szymański[1], **Agnieszka Rossa**[2]

ABSTRACT

The paper deals with a new fuzzy version of the Lee-Carter (LC) mortality model, in which mortality rates as well as parameters of the LC model are treated as triangular fuzzy numbers. As a starting point, the fuzzy Koissi-Shapiro (KS) approach is recalled. Based on this approach, a new fuzzy mortality model – CNMM – is formulated using orthonormal expansions of the inverse exponential membership functions of the model components. The paper includes numerical findings based on a case study with the use of the new mortality model compared to the results obtained with the standard LC model.

Key words: exponential membership functions, Legendre's polynomials, mortality modelling, orthonormal system.

1. Introduction

In the last four decades several approaches were proposed to model human mortality and to project future mortality evolution. Among the extrapolative methods, a model proposed by Lee and Carter (1992) is one of the most popular approaches, although other mortality models have been also developed, e.g. Heligman and Pollard (1980), Horiuchi and Coale (1990), Milevsky and Promislow (2001), Currie et al. (2004), Bongaarts (2005), Cairns et al. (2006).

The Lee-Carter model (LC) has been extensively used for many real populations and extended in various directions (see, e.g. Renshaw et al. (1996), Tuljapurkar et al. (2000), Booth et al. (2002), (2006), Brouhns et al. (2002), Renshaw and Haberman (2003), De Jong and Tickle (2006), Koissi and Shapiro (2006), Pitacco et al. (2009), Haberman and Renshaw (2012), Danesi et al. (2015)).

The Lee-Carter method (1992) can be treated as a special case of the principal component analysis with a single component (Bozik and Bell (1987)). The focus of this

[1] Institute of Statistics and Demography, University of Lodz, 41 Rewolucji 1905 St., 90-214 Lodz, Poland. .
[2] Institute of Statistics and Demography, University of Lodz, 41 Rewolucji 1905 St., 90-214 Lodz, Poland. E-mail of the corresponding author: agnieszka.rossa@uni.lodz.pl. ORCID: https://orcid.org/0000-0002-0444-4181.

approach is on central age-specific death rates m_{xt} for a range of ages $x = 0,1,2,\dots,X$ and calendar years $t = 1,2,\dots,T$, organized in a two way table with rows referring to one-year age groups and columns referring to one-year period intervals.

The LC method consists of a model of age-specific log-central death rates $y_{xt} = \ln m_{xt}$ with time and age components

$$y_{xt} = a_x + b_x k_t + \varepsilon_{xt}, \qquad x = 0,1,2,\dots X, \;\; t = 1,2,\dots,T, \tag{1}$$

and a model of random walk with a drift to forecast time components k_t for $t > T$

$$k_t = d + \; k_{t-1} + \zeta_t, \tag{2}$$

where $\{a_x\}$ in (1) is a set of age-related effects describing the age profile of mortality, $\{k_t\}$ is a set of the time-related effects representing the general trend of mortality, $\{b_x\}$ is a set of age-related effects describing patterns of deviations from the age profile in response to change of the general trend, d in (2) is a constant (a drift), whereas ε_{xt}, ζ_t in (1) and (2), respectively, are random residuals.

Parameters $\{b_x\}$ show which death rates decline rapidly and which slowly over time in response to change of k_t. For some values of x, b_x may be positive while negative for other values, indicating that log-central death rates $y_{xt} = \ln m_{xt}$ are increasing at some ages while decreasing at other ages.

For the full identification of (1), the following two constraints are imposed

$$\sum_{x=0}^{X} b_x = 1, \; \sum_{t=1}^{T} k_t = 0. \tag{3}$$

Lee and Carter used the SVD method (Singular Value Decomposition) to estimate a_x, b_x, k_t and assumed that error terms ε_{xt} are normally distributed with a small constant variance. This is rather a strong assumption, which is often violated especially in the case of the imprecise input data. Moreover, prediction errors do not account for the estimation errors of the age-specific parameters a_x, b_x, except of incorporating uncertainty from the forecast of the time component k_t.

It is well-known that various kinds of errors can occur in reporting death statistics. This could be e.g. incorrect year, area or age. Moreover, the midyear population data used to calculate period age-specific mortality rates m_{xt} are also the subject of errors. The midyear population size is the population at July 1 and is assumed to be the point at which half of the deaths during the year have occurred. Such estimates can be actually underestimated or overestimated and this affects the resulting death rates. Therefore, exact age-specific mortality rates are seldom known, hence incorporating the data uncertainty into the model structure seems to be a realistic and expected idea.

The new trends in fuzzy analysis are based on the algebraic approach to fuzzy numbers (e.g. Ishikawa (1997), Kosiński et al. (2003), Rossa et al. (2017), Szymański and Rossa (2014), (2017)). The essential idea in such an approach is representing the membership function of

a fuzzy number as an element of the square-integrable function space. We will use this idea to propose a new fuzzy mortality model in the spirit of the Koissi-Shapiro approach.

The log-central mortality rates as well as parameters of the underlying Koissi-Shapiro model are symmetric triangular fuzzy numbers, i.e. numbers with symmetric triangular membership functions. We believe that exponential functions could fit the data better. Therefore, our model is based on exponential membership functions of the model components instead of triangular ones.

The paper is organized as follows. Section 2 recalls the data fuzzification method (Subsection 2.1) and the fuzzy mortality model (Subsection 2.2) as proposed by Koissi and Shapiro. The new complex-number fuzzy mortality model is formulated in Section 3. The concept is presented in six subsections: theoretical backgrounds (Subsection 3.1), formulation of the new mortality model CNMM (Subsection 3.2), estimation of the model parameters (Subsection 3.3), description of the modified fuzzification method (Subsection 3.4), description of the forecasting method (Subsection 3.5) and a case study (Subsection 3.6). Concluding remarks are contained in Section 4. Formal details about orthonormal expansions by means of the Legendre polynomials are included in the Appendix.

2. The Koissi-Shapiro model

2.1. Fuzzification of the input data

In the Koissi-Shapiro model (2006), log-central death rates $y_{xt} = \ln m_{xt}$ are transformed into symmetric triangular fuzzy numbers

$$Y_{xt} = (y_{xt}, e_{xt}), \tag{4}$$

where y_{xt}, e_{xt} are centres and spreads of fuzzy numbers Y_{xt}, respectively.

The addition $\oplus$ and multiplication $\otimes$ of symmetric triangular numbers $A = (a, s_A)$ and $B = (b, s_B)$ defined in the norm T_w are expressed as

$$A \oplus B = (a + b, \max(s_A, s_B)), \tag{5}$$

$$A \otimes B = (ab, \max(s_A|b|, s_B|a|)), \tag{6}$$

and the multiplication of $A = (a, s_A)$ by a scalar $b \in \mathbb{R}$ reduces to

$$A \otimes b = (ab, s_A|b|). \tag{7}$$

Parameters e_{xt} in (4) are also called fuzziness parameters. To determine their values, Koissi and Shapiro postulated using a fuzzy regression model. They assumed existing symmetric triangular fuzzy numbers (c_{0x}, s_{0x}) and (c_{1x}, s_{1x}) satisfying for each age group x the following equalities

$$(y_{xt}, e_{xt}) = (c_{0x}, s_{0x}) \oplus (c_{1x}, s_{1x}) \otimes t, \qquad t = 1,2, \dots, T. \tag{8}$$

This postulate leads to the equalities (9)–(10) of the form

$$y_{xt} = c_{0x} + c_{1x} \cdot t, \qquad t = 1,2, \dots, T. \tag{9}$$

$$e_{xt} = \max(s_{0x}, s_{1x} \cdot t\,), \qquad t = 1,2, \dots, T. \tag{10}$$

To find coefficients in (9) the ordinary least-squares regression is used, i.e. c_{1x} and c_{0x} are found from formulas

$$c_{1x} = \frac{\overline{y_{xt} \cdot t} - \bar{t} \cdot \overline{y_{xt}}}{\overline{t^2} - \bar{t}^2}, \tag{11}$$

$$c_{0x} = \overline{y}_{xt} - c_{1x} \cdot \bar{t}, \tag{12}$$

where $\bar{z}$ means averaging over z_t's.

To find parameters s_{0x}, s_{1x},"the minimum fuzziness criterion" is proposed by minimizing spreads of $Y_{xt} = (y_{xt}, e_{xt})$ and requiring each log-central death rate y_{xt} to fall within the estimated death rates $\hat{y}_{xt}$ at a level $h \in [0,1]$. Since e_{xt} are, by assumption, non-negative numbers and the smallest value they can take is 0, it is necessary to determine such values of s_{0x}, s_{1x}, that at a given x they minimize the sum

$$T \cdot s_{0x} + s_{1x} \cdot \sum_{t=1}^{T} t, \tag{13}$$

subject to the constraints

$$c_{0x} + c_{1x} \cdot t + (1-h)(s_{0x} + s_{1x} t) \geq \ln m_{xt}, \quad t = 1,2,\dots,T, \tag{14}$$

$$c_{0x} + c_{1x} \cdot t - (1-h)(s_{0x} + s_{1x} t) \leq \ln m_{xt}, \quad t = 1,2,\dots,T, \tag{15}$$

where $s_{0x}, s_{1x} \geq 0,\ u \in [0,1)$ and $h \in [0,1]$ is a predetermined value representing the degree of fit of the estimated model to the data. As lower h provides a better fit, we can use $h = 0$. After finding the parameters s_{0x}, s_{1x} for each x, the fuzziness parameters e_{xt} can be determined using formula (10).

2.2. The Koissi-Shapiro model

Let us recall the fuzzy mortality model as proposed by Koissi and Shapiro (2006). The structure of their model is analogous to the Lee-Carter one (1992) and takes the form

$$Y_{xt} = A_x \oplus (B_x \otimes K_t), \tag{16}$$

with the difference that $Y_{xt} = (y_{xt}, e_{xt})$ for $x = 0,1,\dots,X,\ t = 1,2,\dots,T$ are fuzzified log-central death rates expressed as triangular numbers with centres y_{xt} and spreads e_{xt}.

Model parameters are assumed to be symmetric triangular numbers $A_x = (a_x, s_{A_x}),\ B_x = (b_x, s_{B_x}), K_t = (k_t, s_{K_t})$ with unknown centres $a_x, b_x, k_t \in \mathbb{R}$ and spreads $s_{A_x}, s_{B_x}, s_{K_t} \geq 0$, respectively.

To find parameters $a_x, b_x, k_t, s_{A_x}, s_{B_x}, s_{K_t}$, Koissi and Shapiro postulated minimizing the Diamond distance D^2 (Diamond (1988)) between the left and right-

hand sides of (16). This leads to the criterion function defined for each separate x and t as

$$D^2(Y_{xt}, A_x \oplus (B_x \otimes K_t)) = (a_x + b_x k_t - y_{xt})^2 + \left[a_x + b_x k_t - \max\left\{s_{A_x}, |b_x| s_{K_t}, |k_t| s_{B_x}\right\} - (y_{xt} - e_{xt})\right]^2 + \left[a_x + b_x k_t + \max\left\{s_{A_x}, |b_x| s_{K_t}, |k_t| s_{B_x}\right\} - (y_{xt} + e_{xt})\right]^2. \tag{17}$$

Unfortunately, the criterion function contains a max-type operator $\max\left\{s_{A_x}, |b_x| s_{K_t}, |k_t| s_{B_x}\right\}$, which does not allow using standard derivative based solution algorithms for minimization of (17).

3. The Complex-Number Mortality Model CNMM

3.1. Theoretical backgrounds

The new trends in fuzzy analysis are based on the algebraic approach to fuzzy numbers (see, e.g. Ishikawa (1997), Kosiński et al. (2003), Rossa et al. (2017), Szymański and Rossa (2014), (2017)). The essential idea in such an approach is representing the membership function of a fuzzy number as an element of the square-integrable function space.

Let us consider the membership function of the exponential form

$$\mu(z) = \begin{cases} \exp\left\{-\left(\frac{c-z}{\tau}\right)^2\right\}, & z \le c, \\ \exp\left\{-\left(\frac{z-c}{\nu}\right)^2\right\}, & z > c, \end{cases} \tag{18}$$

where $c \in \mathbb{R}$, $\tau, \nu > 0$ are some scalar parameters.

Note that (18) can be decomposed into two parts – strictly increasing and strictly decreasing functions $\Psi(z)$ and $\Phi(z)$, say, of the form

$$\Psi(z) = \exp\left\{-\left(\frac{c-z}{\tau}\right)^2\right\}, \quad z \le c, \tag{19}$$

$$\Phi(z) = \exp\left\{-\left(\frac{z-c}{\nu}\right)^2\right\}, \quad z > c. \tag{20}$$

Then there exist inverse functions

$$\Psi^{-1}(u) = c + \psi(u), \quad u \in [0,1], \tag{21}$$

$$\Phi^{-1}(u) = c + \varphi(u), \quad u \in [0,1], \tag{22}$$

where $\psi(u)$ and $\varphi(u)$ are expressed as

$$\psi(u) = -\tau(-\ln u)^{\frac{1}{2}}, \quad \varphi(u) = \nu(-\ln u)^{\frac{1}{2}}, \quad u \in [0,1]. \tag{23}$$

Denoting $f(u) = \Psi^{-1}(u)$ and $g(u) = \Phi^{-1}(u)$ for $u \in [0,1]$, we can write

$$f(u) = c + \psi(u) = c - \tau(-\ln u)^{\frac{1}{2}}, \quad g(u) = c + \varphi(u) = c + \nu(-\ln u)^{\frac{1}{2}}, \tag{24}$$

Functions f, g are square-integrable, so the ordered pair (f, g) belongs to the Cartesian product $L^2(0,1) \times L^2(0,1)$. The scalar product in the space $L^2(0,1)$ is given by the formula

$$\langle f, g \rangle = \int_0^1 f(u)\, g(u) du. \tag{25}$$

Example 1. Figure 1(a) depicts functions $\Psi(z)$ and $\Phi(z)$ as defined in (19) and (20), while Figure 1(b) shows their inverse counterparts (21) and (22), respectively.

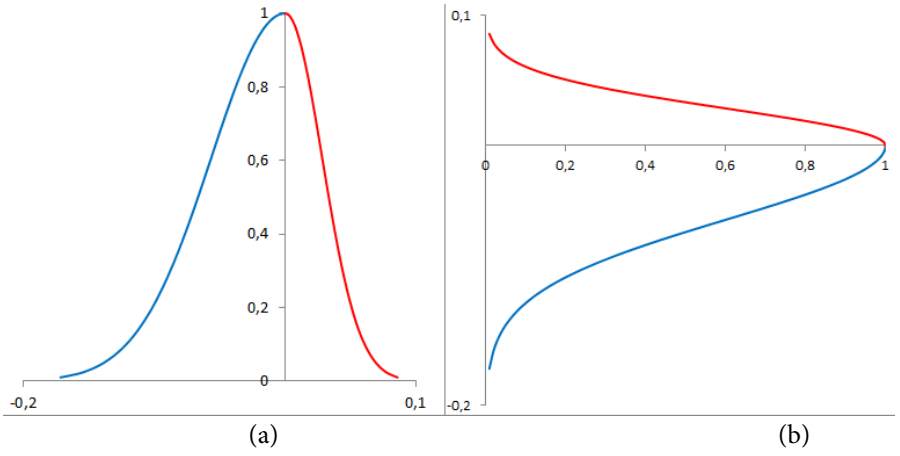


Figure 1. Exponential functions $\Psi(z)$, $\Phi(z)$ and the inverse functions $\Psi^{-1}(u)$, $\Phi^{-1}(u)$ for $c = 0.0$, $\tau = 0.08, \nu = 0.09$.

Source: Developed by the authors.

It is commonly known that a set of vectors $\{P_j\}$ in $L^2(0,1)$ is called an orthonormal set if equalities $\langle P_j, P_k \rangle = 0$ for $j \neq k$ and $\langle P_j, P_j \rangle = 1$ are true.

For any orthonormal set $\{P_j\}$ and $f, g \in L^2(0,1)$ the following relations hold

$$f = \sum_{j=0}^{\infty} \langle P_j, f \rangle P_j, \quad g = \sum_{j=1}^{\infty} \langle P_j, g \rangle P_j. \tag{26}$$

Denoting $\alpha_j = \langle P_j, f \rangle$ and $\beta_j = \langle P_j, g \rangle$, expressions (26) can also be written as

$$f(u) = \sum_{j=0}^{\infty} \alpha_j P_j(u), \quad g(u) = \sum_{j=0}^{\infty} \beta_j P_j(u). \tag{27}$$

Let $A^{(N)}$ be a pair of functions $\left(f^{(N)}, g^{(N)}\right)$, where $f^{(N)}, g^{(N)}$ for $N \in \mathbb{N}$ are some orthonormal expansions of inverse exponential functions (24), i.e.

$$f^{(N)}(u) = \sum_{j=0}^{N} \alpha_j P_j(u), \qquad g^{(N)}(u) = \sum_{j=0}^{N} \beta_j P_j(u), \tag{28}$$

where $\{P_j\}$ is a set of the Legendre polynomials and α_j, β_j are some coefficients of the orthonormal expansion (see Appendix for more details).

Example 2. Let us consider functions $f(u), g(u)$ as depicted in Figure 1(b). Their approximations for $N = 3$ are plotted in Figure 2.

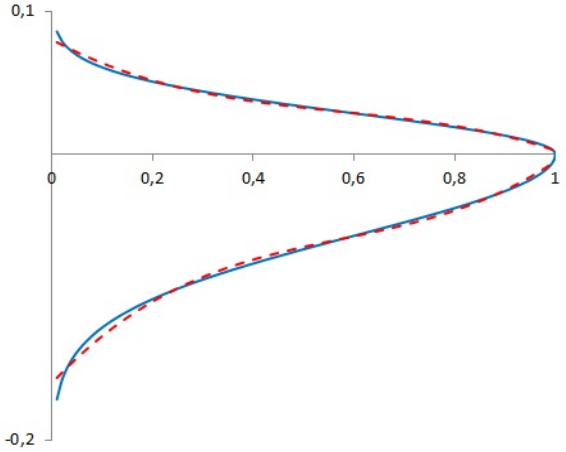


Figure 2. Functions $f(u) = c - \tau(-\ln u)^{\frac{1}{2}}$, $g(u) = c + \nu(-\ln u)^{\frac{1}{2}}$ (solid lines) and their approximations $f^{(3)}(u) = \sum_{j=0}^{3} \alpha_j P_j$, $g^{(3)}(u) = \sum_{j=0}^{3} \beta_j P_j$ (dashed lines).

Source: developed by the authors

Further, we will treat the pairs of functions (f, g) or $\left(f^{(N)}, g^{(N)}\right)$ given in (24), (28), respectively, in terms of the complex analysis. They will be called *complex-valued fuzzy numbers*.

Let the addition, the subtraction and the multiplication of two complex-valued fuzzy numbers $A = (f_A, g_A), B = (f_B, g_B)$ be defined as

$$A \oplus B = (f_A + f_B,\ g_A + g_B), \tag{29}$$

$$A \ominus B = (f_A - f_B, g_A - g_B), \tag{30}$$

$$A \odot B = (f_A f_B - g_A g_B,\ f_A g_B + g_A f_B), \tag{31}$$

while the multiplication of $A = (f_A, g_A)$ by a scalar $d \in \mathbb{R}$ as

$$d \odot A = (d \cdot f_A,\ d \cdot g_A). \tag{32}$$

3.2. The CNMM model formulation

We propose the Complex-Number Mortality Model (CNMM) of the form

$$Y_{xt}^{(N)} = A_x^{(N)} \oplus K_{xt}^{(N)}, \tag{33}$$

where $x = 0,1,\dots,X, t = 1,2,\dots,T$ are age and time indices, respectively, $Y_{xt}^{(N)} = \left(f_{Y_{xt}}^{(N)}, g_{Y_{xt}}^{(N)}\right)$ are complex-valued fuzzy numbers representing fuzzified log-central mortality rates, and $A_x^{(N)} = \left(f_{A_x}^{(N)}, g_{A_x}^{(N)}\right)$, $K_{xt}^{(N)} = \left(f_{K_{xt}}^{(N)}, g_{K_{xt}}^{(N)}\right)$ are some complex-

valued fuzzy numbers with functions $f_{A_x}^{(N)}$, $g_{A_x}^{(N)}$, $f_{K_{xt}}^{(N)}$, $g_{K_{xt}}^{(N)}$ and $f_{Y_{xt}}^{(N)}$, $g_{Y_{xt}}^{(N)}$ being orthonormal expansions (28) of the following functions

$$f_{A_x}(u) = a_x - \tau_{A_x}(-\ln u)^{\frac{1}{2}}, \quad g_{A_x}(u) = a_x + \nu_{A_x}(-\ln u)^{\frac{1}{2}}, \tag{34}$$

$$f_{K_{xt}}(u) = b_x k_t - \tau_{B_x}\omega_t(-\ln u)^{\frac{1}{2}}, \quad g_{K_{xt}}(u) = b_x k_t + \nu_{B_x}\varpi_t(-\ln u)^{\frac{1}{2}}, \tag{35}$$

$$f_{Y_{xt}}(u) = y_{xt} - e_{xt}(-\ln u)^{\frac{1}{2}}, \quad g_{Y_{xt}}(u) = y_{xt} + \upsilon_{xt}(-\ln u)^{\frac{1}{2}}, \tag{36}$$

Coefficients $a_x, b_x, k_t, \tau_{A_x}, \nu_{A_x}\ \tau_{B_x}, \nu_{B_x}, \omega_t, \varpi_t$ in (34)–(36) constitute a set of unknown parameters, $y_{xt} = \ln m_{xt}$ are log-central death rates, and e_{xt}, υ_{xt} represent fuzziness of log-central mortality rates evaluated at the fuzzification stage (see Subsection 3.4).

Let us express the model in terms of complex analysis using an algebraic representation, i.e.

$$Y_{xt}^{(N)} = f_{Y_{xt}}^{(N)} + i \cdot g_{Y_{xt}}^{(N)}, \quad A_x^{(N)} = f_{A_x}^{(N)} + i \cdot g_{A_x}^{(N)}, \quad K_{xt}^{(N)} = f_{K_{xt}}^{(N)} + i \cdot g_{K_{xt}}^{(N)}, \tag{37}$$

where $i = \sqrt{-1}$ is an imaginary unit.

Then, taking into account (28) we can write

$$A_x^{(N)} = \sum_{j=0}^{N} \alpha_{xj} P_j + i \sum_{j=0}^{N} \beta_{xj} P_j = \sum_{j=0}^{N} (\alpha_{xj} + i\,\beta_{xj}) P_j, \tag{38}$$

$$K_{xt}^{(N)} = \sum_{j=0}^{N} \eta_{txj} P_j + i \sum_{j=0}^{N} \lambda_{txj} P_j = \sum_{j=0}^{N} (\eta_{txj} + i\,\lambda_{txj}) P_j. \tag{39}$$

Thus, the right-hand side of (33) can be expressed as

$$A_x^{(N)} \oplus K_{xt}^{(N)} = \sum_{j=0}^{N} \left[(\alpha_{xj} + \eta_{xtj}) + i(\beta_{xj} + \lambda_{xtj}) \right] P_j. \tag{40}$$

By analogy, the left-hand side of (33) can be written in the form

$$Y_{xt}^{(N)} = \sum_{j=0}^{N} \epsilon_{xtj} P_j + i \sum_{j=0}^{N} \theta_{xtj} P_j = \sum_{j=0}^{N} (\epsilon_{xtj} + i\,\theta_{xtj}) P_j. \tag{41}$$

Coefficients $\alpha_{xj}, \eta_{xtj}, \beta_{xj}, \lambda_{txj}$ and $\epsilon_{xtj}, \theta_{xtj}$ in expansions (40), (41), respectively, correspond to parameters $a_x, b_x, k_t, \tau_{A_x}, \nu_{A_x}\ \tau_{B_x}, \nu_{B_x}, \omega_t, \varpi_t$ via relations (42), (43).

For $j = 0$ we have

$$\begin{aligned} \alpha_{x0} &= a_x - \tau_{A_x} c_0, & \beta_{x0} &= a_x + \nu_{A_x} c_0, \\ \eta_{tx0} &= b_x k_t - \tau_{B_x} \omega_t c_0, & \lambda_{tx0} &= b_x k_t + \nu_{B_x} \varpi_t c_0, \\ \epsilon_{xt0} &= y_{xt} - e_{xt} c_0, & \theta_{xt0} &= y_{xt} + \upsilon_{xt} c_0, \end{aligned} \tag{42}$$

and for $j = 1,2, \ldots, N$ there is

$$\begin{aligned} \alpha_{xj} &= -\tau_{A_x} c_j, & \beta_{xj} &= \nu_{A_x} c_j, \\ \eta_{xtj} &= -\tau_{B_x} \omega_t c_j, & \lambda_{xtj} &= \nu_{B_x} \varpi_t c_j, \\ \epsilon_{xtj} &= -e_{xt} c_j, & \theta_{xtj} &= \upsilon_{xt} c_j, \end{aligned} \tag{43}$$

where c_j are some known constants (see Appendix for more details).

For $j = 0,1,2,3$ we get $c_0 = \frac{\sqrt{\pi}}{2}, c_1 = \sqrt{3\pi}\left(\frac{1}{2\sqrt{2}} - \frac{1}{2}\right)$, $c_2 = \sqrt{5\pi}\left(\frac{1}{\sqrt{3}} - \frac{3}{2\sqrt{2}} + \frac{1}{2}\right)$, $c_3 = \sqrt{7\pi}\left(-\frac{5}{\sqrt{3}} + \frac{15}{4\sqrt{2}} - \frac{3}{4\sqrt{2}} + \frac{3}{4}\right)$.

3.3. Estimation of the model parameters

To estimate parameters of the CNMM model we apply the metric in the Hilbert space $L_2(0,1)$ between the left and right-hand sides of (33), i.e. between $Y_{xt}^{(N)}$ and $A_x^{(N)} \oplus K_{xt}^{(N)}$. The estimation problem requires minimizing functional $F^{(N)}$ in the Hilbert space $L_2(0,1)$ of the form

$$F^{(N)} = \sum_{x=0}^{X}\sum_{t=1}^{T}\left\|Y_{xt}^{(N)} \ominus \left(A_x^{(N)} \oplus K_{xt}^{(N)}\right)\right\|^2. \tag{44}$$

Thus, $Y_{xt}^{(N)} \ominus \left(A_x^{(N)} \oplus K_{xt}^{(N)}\right)$ can be expressed as

$$Y_{xt}^{(N)} \ominus \left(A_x^{(N)} \oplus K_{xt}^{(N)}\right) = \\ = \sum_{j=0}^{N}\left[\epsilon_{xtj} - (\alpha_{xj} + \eta_{xtj}) + i(\theta_{xtj} - (\beta_{xj} + \lambda_{xtj}))\right]P_j. \tag{45}$$

After some rearrangements, we get

$$F^{(N)} = \sum_{x=0}^{X}\sum_{t=1}^{T}\left\|Y_{xt}^{(N)} \ominus \left(A_x^{(N)} \oplus K_{xt}^{(N)}\right)\right\|^2 = \sum_{x=0}^{X}\sum_{t=1}^{T}\left\|\sum_{j=0}^{N}\left[\epsilon_{xtj} - (\alpha_{xj} + \eta_{xtj}) + i((\theta_{xtj} - (\beta_{xj} + \lambda_{xtj}))\right]P_j\right\|^2. \tag{46}$$

Using Pythagorean theorem for the Hilbert space of complex functions, i.e.

$$\left\|\sum_{j=0}^{N}\alpha_j P_j\right\|^2 = \sum_{j=0}^{N}\left|\alpha_j\right|^2, \tag{47}$$

the criterion function $F^{(N)}$ takes the form

$$F^{(N)} = \sum_{x=0}^{X}\sum_{t=1}^{T}\sum_{j=0}^{N}\left|\epsilon_{xtj} - (\alpha_{xj} + \eta_{xtj}) + i(\theta_{xtj} - (\beta_{xj} + \lambda_{xtj}))\right|^2 = \\ \sum_{x=0}^{X}\sum_{t=1}^{T}\sum_{j=0}^{N}\left\{\left[\epsilon_{xtj} - (\alpha_{xj} + \eta_{xtj})\right]^2 + \left[\theta_{xtj} - (\beta_{xj} + \lambda_{xtj})\right]^2\right\}. \tag{48}$$

On the basis of relations (42) and (43), we have also

$$F^{(N)} = \sum_{x=0}^{X}\sum_{t=1}^{T}\left[y_{xt} - a_x - b_x k_t + c_0\left(-e_{xt} + \tau_{A_x} + \tau_{B_x}\omega_t\right)\right]^2 + \\ \sum_{x=0}^{X}\sum_{t=1}^{T}\left[y_{xt} - a_x - b_x k_t + c_0(\upsilon_{xt} - \nu_{A_x} - \nu_{B_x}\varpi_t)\right]^2 + \\ D^{(N)}\sum_{x=0}^{X}\sum_{t=1}^{T}\left\{\left(-e_{xt} + \tau_{A_x} + \tau_{B_x}\omega_t\right)^2 + \left(\upsilon_{xt} - \nu_{A_x} - \nu_{B_x}\varpi_t\right)^2\right\}, \tag{49}$$

where $D^{(N)} = \sum_{j=1}^{N} c_j^2$.

The criterion function $F^{(N)}$ can also be written as

$$F^{(N)} = \sum_{x=0}^{X}\sum_{t=1}^{T}\Big[2(y_{xt} - a_x - b_x k_t)^2 + C^{(N)}\big(e_{xt} - \tau_{A_x} - \tau_{B_x}\omega_t\big)^2 + C^{(N)}\big(\upsilon_{xt} - \nu_{A_x} - \nu_{B_x}\varpi_t\big)^2 - 2c_0(y_{xt} - a_x - b_x k_t)\big(e_{xt} - \tau_{A_x} - \tau_{B_x}\omega_t\big) + 2c_0(y_{xt} - a_x - b_x k_t)\big(\upsilon_{xt} - \nu_{A_x} - \nu_{B_x}\varpi_t\big)\Big], \tag{50}$$

where $C^{(N)} = c_0^2 + D^{(N)}$.

To satisfy identifiability of the model, we impose constraints analogous to (3) as well as some additional constraints, i.e.

$$\begin{aligned} &\textstyle\sum_{t=1}^{T} k_t = 0, \quad \sum_{x=0}^{X} b_x = 1, \\ &\textstyle\sum_{x=0}^{X} \tau_{B_x} = 1, \ \sum_{x=0}^{X} \nu_{B_x} = 1, \\ &\textstyle\sum_{t=1}^{T} \omega_t = C, \ \sum_{t=1}^{T} \varpi_t = D, \end{aligned} \tag{51}$$

where $C, D > 0$ are some fixed constants.

Moreover, we impose also boundary constraints of the form

$$\sum_{t=1}^{T} y_{xt} = \sum_{t=1}^{T}(a_x + b_x k_t), \quad \sum_{x=0}^{X} y_{xt} = \sum_{x=0}^{X}(a_x + b_x k_t), \tag{52}$$

$$\sum_{t=1}^{T} e_{xt} = \sum_{t=1}^{T}\big(\tau_{A_x} + \tau_{B_x}\omega_t\big), \quad \sum_{x=0}^{X} e_{xt} = \sum_{x=0}^{X}\big(\tau_{A_x} + \tau_{B_x}\omega_t\big), \tag{53}$$

$$\sum_{t=1}^{T} \upsilon_{xt} = \sum_{t=1}^{T}\big(\nu_{A_x} + \nu_{B_x}\varpi_t\big), \quad \sum_{x=0}^{X} \upsilon_{xt} = \sum_{x=0}^{X}\big(\nu_{A_x} + \nu_{B_x}\varpi_t\big). \tag{54}$$

It follows from requirements (51)–(54) that the following equalities hold

$$a_x = \frac{1}{T}\sum_{t=1}^{T} y_{xt}, \tag{55}$$

$$k_t = \sum_{x=0}^{X}(y_{xt} - a_x), \tag{56}$$

$$\tau_{A_x} = \frac{1}{T}\sum_{t=1}^{T} e_{xt} - \frac{C}{T}\tau_{B_x}, \ \ \nu_{A_x} = \frac{1}{T}\sum_{t=1}^{T} \upsilon_{xt} - \frac{D}{T}\nu_{B_x}. \tag{57}$$

$$\omega_t = \sum_{x=0}^{X}\big(e_{xt} - \tau_{A_x}\big), \quad \varpi_t = \sum_{x=0}^{X}\big(\upsilon_{xt} - \nu_{A_x}\big). \tag{58}$$

Partial derivatives of $F^{(N)}$ with respect to the remaining parameters b_x and τ_{B_x}, ν_{B_x} are of the form

$$\frac{\partial F^{(N)}}{\partial b_x} = -\sum_{t=1}^{T} k_t\big\{4(y_{xt} - a_x - b_x k_t) - 2c_0\big[\big(e_{xt} - \tau_{A_x} - \tau_{B_x}\omega_t\big) - \big(\upsilon_{xt} - \nu_{A_x} - \nu_{B_x}\varpi_t\big)\big]\big\}, \tag{59}$$

$$\frac{\partial F^{(N)}}{\partial \tau_{B_x}} = -2\sum_{t=1}^{T} \omega_t\big[C^{(N)}\big(e_{xt} - \tau_{A_x} - \tau_{B_x}\omega_t\big) - c_0(y_{xt} - a_x - b_x k_t)\big], \tag{60}$$

$$\frac{\partial F^{(N)}}{\partial \nu_{B_x}} = -2\sum_{t=1}^{T} \varpi_t\big[C^{(N)}\big(\upsilon_{xt} - \nu_{A_x} - \nu_{B_x}\varpi_t\big) + c_0(y_{xt} - a_x - b_x k_t)\big]. \tag{61}$$

Setting (59)–(61) equal to zero we receive

$$b_x = \frac{\sum_{t=1}^{T} k_t\left[2y_{xt} - c_0\left(e_{xt} - \upsilon_{xt} - \tau_{B_x}\omega_t + \nu_{B_x}\varpi_t\right)\right]}{2\sum_{t=1}^{T} k_t^2}, \tag{62}$$

$$\tau_{B_x} = \frac{C^{(N)}\sum_{t=1}^{T}\omega_t\left(e_{xt} - \tau_{A_x}\right) - c_0\sum_{t=1}^{T}\omega_t\left(y_{xt} - a_x - b_x k_t\right)}{C^{(N)}\sum_{t=1}^{T}\omega_t^2}, \tag{63}$$

$$\nu_{B_x} = \frac{C^{(N)}\sum_{t=1}^{T}\varpi_t\left(\upsilon_{xt} - \nu_{A_x}\right) + c_0\sum_{t=1}^{T}\varpi_t\left(y_{xt} - a_x - b_x k_t\right)}{C^{(N)}\sum_{t=1}^{T}\varpi_t^2}. \tag{64}$$

The exact solution can be found using an iterative procedure. After choosing a set of starting values for unknown parameters, expressions (57), (58) and (62)–(64) can be computed sequentially using the most recent set of estimates.

It is worth noting that coefficients $k_t, b_x, \tau_{B_x}, \nu_{B_x}, \omega_t, \varpi_t$ satisfy conditions (51). Indeed, we have

$$\sum_{t=1}^{T} k_t = \sum_{t=1}^{T}\sum_{x=0}^{X}(y_{xt} - a_x) = \sum_{t=1}^{T}\sum_{x=0}^{X} y_{xt} - \sum_{x=0}^{X}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{t=1}^{T} y_{xt}\right) = \sum_{t=1}^{T}\sum_{x=0}^{X} y_{xt} - \sum_{x=0}^{X}\sum_{t=1}^{T} y_{xt} = 0\,, \tag{65}$$

and similarly, there is

$$\sum_{x=0}^{X}\tau_{B_x} = \frac{1}{C^{(N)}\sum_{t=1}^{T}\omega_t^2}\sum_{x=0}^{X}\Big[C^{(N)}\sum_{t=1}^{T}\omega_t\left(e_{xt} - \tau_{A_x}\right) - c_0\sum_{t=1}^{T}\omega_t\left(y_{xt} - a_x - b_x k_t\right)\Big] = \frac{1}{C^{(N)}\sum_{t=1}^{T}\omega_t^2}\Big[C^{(N)}\sum_{t=1}^{T}\omega_t\sum_{x=0}^{X}\left(e_{xt} - \tau_{A_x}\right) - c_0\sum_{t=1}^{T}\omega_t\sum_{x=0}^{X}(y_{xt} - a_x) + c_0\sum_{x=0}^{X} b_x\sum_{t=1}^{T}\omega_t k_t\Big]. \tag{66}$$

From (51), (56), (58) we have $\sum_{x=0}^{X} b_x = 1$, $\sum_{x=0}^{X}(y_{xt} - a_x) = k_t$, $\sum_{x=0}^{X}\left(e_{xt} - \tau_{A_x}\right) = \omega_t$. Thus, we can write

$$\sum_{x=0}^{X}\tau_{B_x} = \frac{1}{C^{(N)}\sum_{t=1}^{T}\omega_t^2}\Big[C^{(N)}\sum_{t=1}^{T}\omega_t^2 - c_0\sum_{t=1}^{T}\omega_t k_t + c_0\sum_{t=1}^{T}\omega_t\, k_t\Big] = 1. \tag{67}$$

We also have

$$\sum_{t=1}^{T}\omega_t = \sum_{t=1}^{T}\sum_{x=0}^{X}\left(e_{xt} - \tau_{A_x}\right) = \sum_{x=0}^{X}\sum_{t=1}^{T}\left(e_{xt} - \tau_{A_x}\right) = \sum_{x=0}^{X}\sum_{t=1}^{T} e_{xt} - T\sum_{x=0}^{X}\tau_{A_x} = \sum_{x=0}^{X}\sum_{t=1}^{T} e_{xt} - \sum_{x=0}^{X}\sum_{t=1}^{T} e_{xt} + C\sum_{x=0}^{X}\tau_{B_x} = C\cdot 1 = C. \tag{68}$$

Similar derivations refer to $\sum_{x=0}^{X}\nu_{B_x}$ and $\sum_{t=1}^{T}\varpi_t$. Hence, there following equalities hold

$$\sum_{x=0}^{X}\tau_{B_x} = \sum_{x=0}^{X}\nu_{B_x} = 1 \quad \text{and} \quad \sum_{t=1}^{T}\omega_t = C, \quad \sum_{t=1}^{T}\varpi_t = D. \tag{69}$$

There is also

$$\sum_{x=0}^{X} b_x = \sum_{x=0}^{X}\frac{\sum_{t=1}^{T} k_t\left[2y_{xt} + c_0\left(e_{xt} - \upsilon_{xt} - \tau_{B_x}\omega_t + \nu_{B_x}\varpi_t\right)\right]}{2\sum_{t=1}^{T} k_t^2} = \frac{1}{2\sum_{t=1}^{T} k_t^2}\Big[2\sum_{t=1}^{T} k_t\sum_{x=0}^{X}(y_{xt} - a_x) + c_0\sum_{t=1}^{T} k_t\sum_{x=0}^{X}\left(e_{xt} - \tau_{A_x}\right) - c_0\sum_{t=1}^{T} k_t\sum_{x=0}^{X}\left(\upsilon_{xt} - \nu_{A_x}\right) - c_0\sum_{t=1}^{T} k_t\left(\omega_t\sum_{x=0}^{X}\tau_{B_x} - \varpi_t\sum_{x=0}^{X}\nu_{B_x}\right)\Big]. \tag{70}$$

Using relations (56), (58) and (51) we obtain

$$\sum_{x=0}^{X} b_x = \frac{2\sum_{t=1}^{T} k_t^2 + c_0\left[\sum_{t=1}^{T} k_t(\omega_t - \varpi_t) - \sum_{t=1}^{T} k_t(\omega_t - \varpi_t)\right]}{2\sum_{t=1}^{T} k_t^2} = 1. \tag{71}$$

The special case. Let us assume that $e_{xt} = \upsilon_{xt}$ for $x = 0,2, \dots, X,\ t = 1,2 \dots, T$, then the criterion function (50) reduces to

$$F^{(N)} = 2\sum_{x=0}^{X}\sum_{t=1}^{T}\left[(y_{xt} - a_x - b_x k_t)^2 + C^{(N)}\left(e_{xt} - \tau_{A_x} - \tau_{B_x}\omega_t\right)^2\right] \tag{72}$$

and formulas (62) and (63) defining parameters b_x and τ_{B_x} simplify to the following ones

$$b_x = \frac{\sum_{t=1}^{T} k_t y_{xt}}{\sum_{t=1}^{T} k_t^2}. \tag{73}$$

$$\tau_{B_x} = \frac{\sum_{t=1}^{T} \omega_t\left(e_{xt} - \tau_{A_x}\right)}{\sum_{t=1}^{T} \omega_t^2}, \tag{74}$$

where $\sum_{x=0}^{X} b_x = 1, \sum_{x=0}^{X} \tau_{B_x} = 1$.

It follows from these derivations that the main parameters a_x, b_x, k_t have similar interpretation as in the standard Lee-Carter model (see Section 1). The age-related effects a_x describe the age profile of mortality, time-related effects k_t describe the overall trend of mortality, and b_x represent the mean change of log-central mortality rate y_{xt} in response to change of the time component k_t. However, the CNMM model also has additional parameters $\tau_{A_x}, \tau_{B_x}, \omega_t$ and $\nu_{A_x}, \nu_{B_x}, \varpi_t$ treated as fuzziness of the model parameters. They will be used to determine the fuzziness boundaries of mortality forecasts.

3.4. Data fuzzification

There are several methods proposed to fuzzify the data. One of them is an approach proposed by Koissi and Shapiro (2006) discussed in Subsection 2.1.

What we propose here is to consider a modified version of the Koissi-Shapiro fuzzification method. Let the fuzziness parameters e_{xt} and υ_{xt} satisfy the following respective equations for each fixed x

$$e_{xt} = s_{0x} + s_{1x}t,\ \ \upsilon_{xt} = r_{0x} + r_{1x}t, \qquad t = 1,2, \dots, T, \tag{75}$$

where $s_{0x}, s_{1x}, r_{0x}, r_{1x}$ are found by solving the following optimization problem

$$\text{minimize } \sum_{t=1}^{T}(e_{xt} + \upsilon_{xt}) = T \cdot (s_{0x} + r_{0x}) + (s_{1x} + r_{1x})\sum_{t=1}^{T} t, \tag{76}$$

subject to the constraints

$$a_x + b_x \cdot k_t + (s_{0x} + s_{1x}t) \geq \ln m_{xt},\ \ t = 1,2, \dots, T, \tag{77}$$

$$a_x + b_x \cdot k_t - (r_{0x} + r_{1x}t) \leq \ln m_{xt},\ \ t = 1,2, \dots, T, \tag{78}$$

where a_x, k_t, b_x are defined in (55), (56) and (73), and $s_{0x}, s_{1x} \geq 0$ as well as $r_{0x}, r_{1x} \geq 0$ are the smallest values satisfying inequalities (77) and (78), respectively. Once, the coefficients $s_{0x}, s_{1x}, r_{0x}, r_{1x}$ are found, the fuzziness parameters e_{xt} and υ_{xt} can be determined from equations (75).

3.5. Mortality prediction

To forecast log-central mortality rates, time component k_t can be viewed, analogously to the Lee-Carter approach, as a stochastic process. The estimated or forecasted values $\hat{y}_{xt}$ of log-central death rates y_{xt} will be derived for from the following formula

$$\hat{y}_{xt} = a_x + b_x k_t, \tag{79}$$

where a_x, b_x are time invariant, and k_t is a time dependent component. For $t > T$, the time component will be forecasted via a time series model of the form

$$k_t = \delta + k_{t-1} + \zeta_t, \tag{80}$$

with δ and ζ_t's denoting, respectively, a drift and independent and identically distributed Gaussian random terms.

Similar approach applies to parameters e_{xt}, υ_{xt} expressing fuzziness of log-central death rates. The estimated or forecasted values $\hat{e}_{xt}, \hat{\upsilon}_{xt}$ will be derived from the following formulas

$$\hat{e}_{xt} = \tau_{A_x} + \tau_{B_x}\omega_t, \qquad \hat{\upsilon}_{xt} = \nu_{A_x} + \nu_{B_x}\varpi_t, \tag{81}$$

where $\tau_{A_x}, \tau_{B_x}, \nu_{A_x}, \nu_{B_x}$ are time invariant, while ω_t, ϖ_t are time dependent model parameters. Thus, for $t > T$, both ω_t and ϖ_t will be forecasted using the following time series models

$$\omega_t = \mu + \omega_{t-1} + \varsigma_t, \quad \varpi_t = \gamma + \varpi_{t-1} + \xi_t, \tag{82}$$

with μ, γ denoting some drifts and ς_t, ξ_t denoting independent and identically distributed Gaussian random terms.

The ML estimates $\hat{\delta}, \hat{\mu}, \hat{\gamma}$ of parameters δ, μ, γ are as follows

$$\hat{\delta} = \frac{k_T - k_1}{T-1}, \qquad \hat{\mu} = \frac{\omega_T - \omega_1}{T-1}, \quad \hat{\gamma} = \frac{\varpi_T - \varpi_1}{T-1}. \tag{83}$$

3.6. The case study

To illustrate theoretical discussions presented in this section, the estimates of a_x, b_x, k_t and $\tau_{A_x}, \tau_{B_x}, \omega_t, \nu_{A_x}, \nu_{B_x}, \varpi_t$ were estimated using the real mortality data. Next, the *ex-post* forecasts from the model (33) were derived and the prediction accuracy with results yielded by the Lee-Carter model compared.

The analysis was based on the central death rates in Poland from the years 1965–2019. For computational reasons, age-specific death rates multiplied by 1000 were used. The necessary data were sourced from the Human Mortality Database (www.mortality.org), separately for males and females. The 2014–2019 death rates served the purpose of evaluating predicted rates and were not used in the estimation. Estimates of the parameters were obtained using scaled central death rates for males and females recorded in the years 1965–2013. To ensure the clarity of data presentation, estimates of a_x, b_x, k_t's vs. x or t are plotted in the separate Figures 3–5.

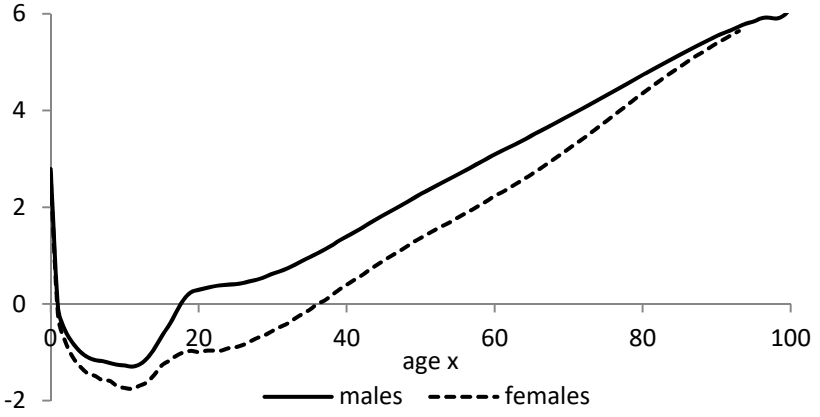


Figure 3. Estimates of parameters a_x, $x = 0,1,2, \dots, X$ (Poland, males and females)

Source: Developed by the authors.

Curves illustrated in Figure 3 show the average profiles of mortality for males and females over the age range $[0,100]$. Both curves exhibit a typical "bath tube" shapes with high values around the infant ages, followed by minimal rates at the childhood ages, higher accidental mortality at young adulthood ages and increasing mortality at adulthood and old ages with nearly constant rate of increase. The "accident hump" at adolescence stands for higher mortality rates due to accidental deaths caused by augmented risk-taking behaviour as well as increased suicide rates. Note that the more demonstrable hump refers to the subpopulation of males.

The arrangement of curves in Figure 4 shows that log-central mortality rates for males in young and old age groups are more sensitive to temporal changes in mortality than analogous rates for females. The reverse relationship applies to other age groups.

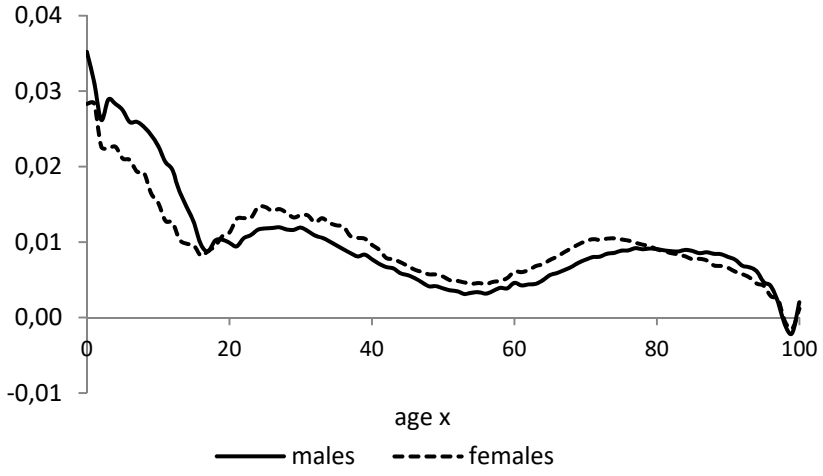


Figure 4. Estimates of parameters b_x, $x = 0,1,2, \dots, X$ (Poland, males and females)

Source: Developed by the authors.

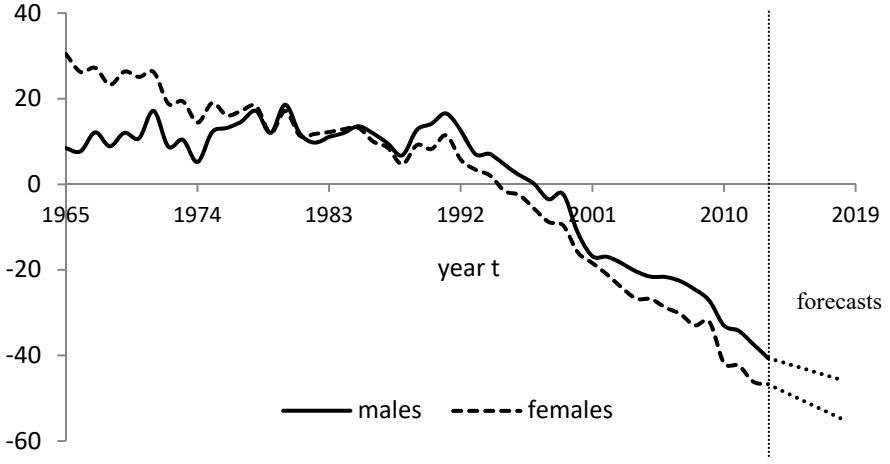


Figure 5. Estimates of parameters $k_t, t = 1,2, \dots, T$ (Poland, males and females) and forecasts up to 2019

Source: Developed by the authors.

Figure 5 illustrates the trend of mortality both for males and females and forecasts up to 2019. It can be seen that curves are generally decreasing, with the decline being faster for women. However, the trend in mortality before 1991 shows a slight flattening, apart from certain fluctuations, which can be explained by the health crisis of the 1970s and 1980s in Poland.

Figures 6−11 exhibit both the real and estimated mortality rates for selected age groups. Estimates of log-central death rates y_{xt} were obtained for males and females by

using formula (79). In this case, as before, the estimation period was 1965–2013 and the period of *ex-post* forecasts spanned the years 2014–2019. Forecasts of k_t after 2013 were determined from the model (80). Similar models (81), (82) were used to estimate and forecast fuzziness parameters e_{xt}, υ_{xt} necessary to determine fuzziness boundaries for mortality forecasts up to 2019.

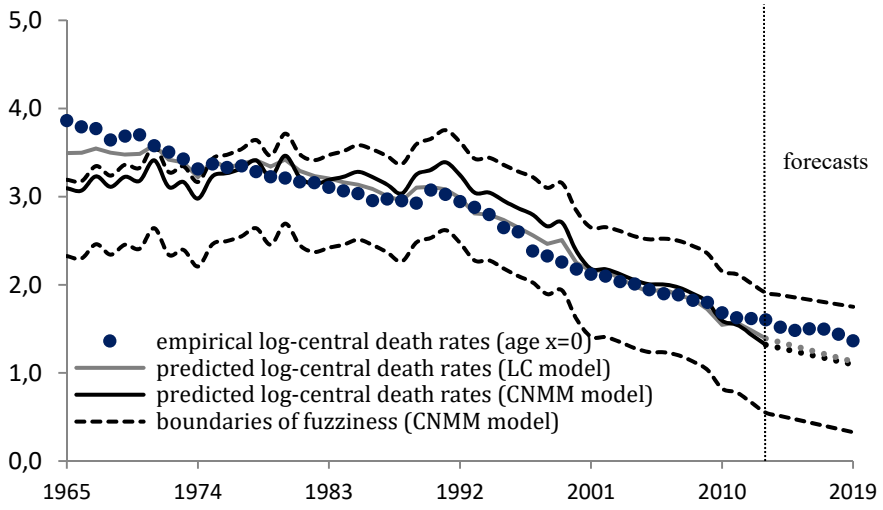


Figure 6. The real and predicted log-central death rates obtained with the LC and CNMM models together with the fuzziness areas (Poland, males aged 0 years)

Source: Developed by the authors.

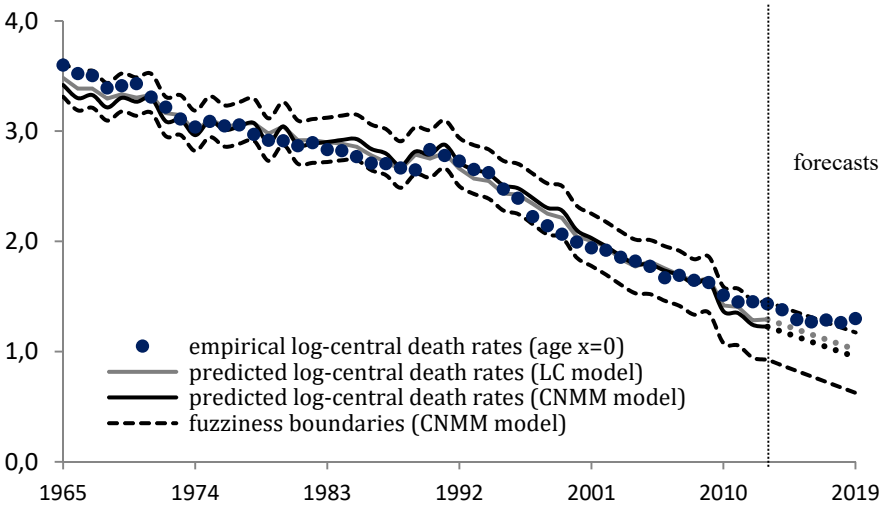


Figure 7. The real and predicted log-central death rates obtained with the LC and CNMM models together with the fuzziness boundaries (Poland, females aged 0 years)

Source: Developed by the authors.

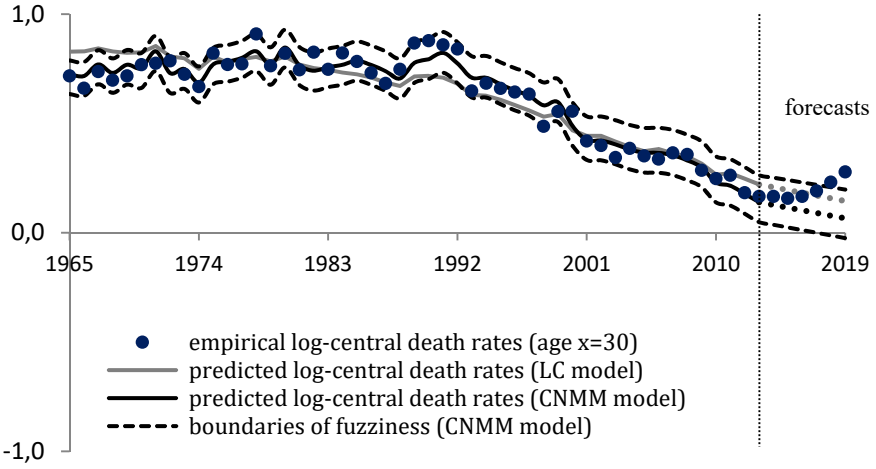


Figure 8. The real and predicted log-central death rates obtained with the LC and CNMM models together with the fuzziness boundaries (Poland, males at the age of 30 years)

Source: Developed by the authors.

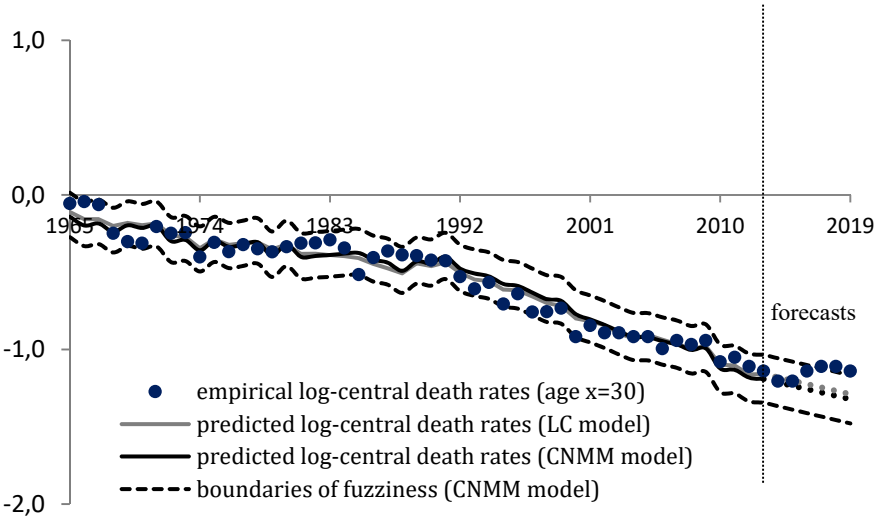


Figure 9. The real and predicted log-central death rates obtained with the LC and CNMM models together with the fuzziness boundaries (Poland, females at the age of 30 years)

Source: Developed by the authors.

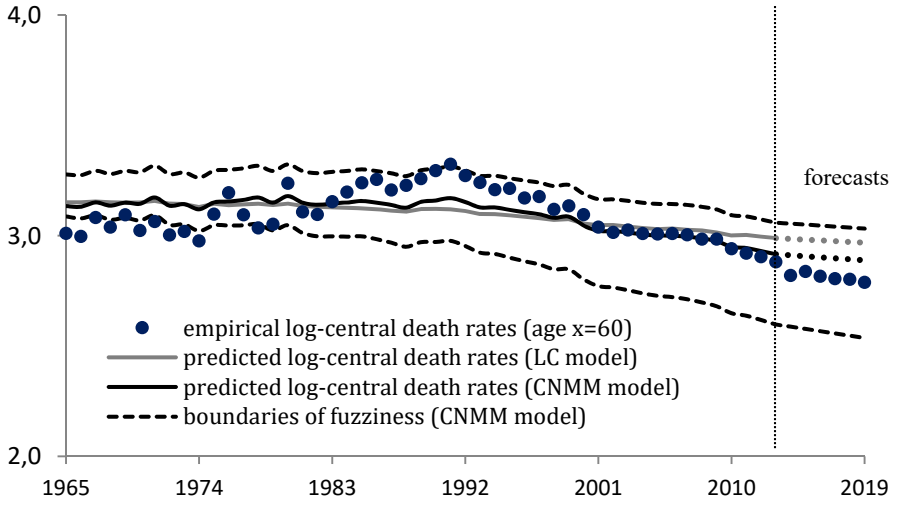


Figure 10. The real and predicted log-central death rates obtained with the LC and CNMM models together with the fuzziness boundaries (Poland, males at the age of 60 years)

Source: Developed by the authors.

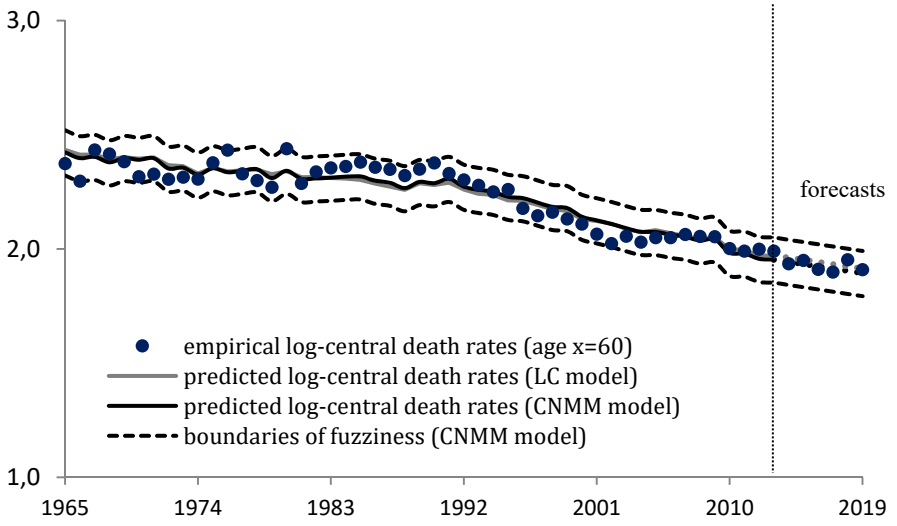


Figure 11. The real and predicted log-central death rates obtained with the LC and CNMM models together with the fuzziness boundaries (Poland, females at the age of 60 years)

Source: Developed by the authors.

The CNMM model as well as the LC model were then compared using *ex-post* mean squared prediction error (MSE) based on the differences between real and estimated log-central mortality rates, i.e.

$$MSE_t = \sqrt{\frac{1}{X+1}\sum_{x=0}^{X}(y_{xt} - \hat{y}_{xt})^2}, \quad t > T, \tag{84}$$

where $\hat{y}_{xt}$ are estimated log-central death rates obtained from the CNMM or LC model.

Table 1. Prediction accuracy of the LC model vs. the CNMM model in terms of the *ex-post* MSE errors

Year	Males		Females	
	LC	**CNMM**	**LC**	**CNMM**
POLAND				
2014	0.166	0.112	0.118	0.116
2015	0.152	0.107	0.111	0.105
2016	0.167	0.116	0.140	0.131
2017	0.174	0.124	0.126	0.125
2018	0.158	0.117	0.150	0.158
2019	0.171	0.129	0.134	0.142
NORWAY				
2014	0.265	0.243	0.305	0.302
2015	0.297	0.273	0.272	0.234
2016	0.294	0.270	0.255	0.248
2017	0.302	0.269	0.340	0.328
2018	0.308	0.283	0.316	0.310
CHECHIA				
2014	0.227	0.220	0.230	0.227
2015	0.281	0.276	0.247	0.238
2016	0.253	0.247	0.235	0.222
2017	0.245	0.242	0.265	0.251

Source: Developed by the authors.

Table 1 summarizes the results of comparisons between the LC and CNMM models in terms of their prediction accuracy for Poland and for two selected European countries. MSE errors were assessed for those years for which the real mortality rates were available.

On the basis of the results obtained, it can be noticed that the CNMM model utilizing complex-valued fuzzy numbers provides comparable or smaller *ex-post* forecast errors, in terms of the MSE measure, than the LC model.

4. Concluding remarks

In the paper an algebraic approach to mortality modelling was introduced. For the formal purposes, the concept of complex-valued fuzzy numbers was also discussed.

The popularity of the widely used Lee-Carter mortality model lies in its simplicity and ease of interpretation. However, due to the uncertainty and imprecision of empirical age-specific mortality rates, it seems justified to use a fuzzy mortality model instead. In our approach, the log-central death rates were viewed as complex-valued fuzzy numbers derived for each age-time cell. The parameters of fuzzified log-central death rates were found in the data fuzzification stage, which was the first step of the model estimation. Next, fuzzy log-central death rates were transformed into complex-valued fuzzy numbers and modelled using the complex analysis.

What makes the CNMM model superior to the standard LC model is that the proposed approach allows for determination of fuzziness boundaries for the mortality trajectories. For the standard LC model, the confidence intervals for log-central mortality rates can also be derived, but they reflect the error term in the random walk model, ignoring the estimation errors of other parameters, so the confidence intervals can only be derived for the prediction window.

Acknowledgements

Authors gratefully acknowledge that their research was supported by a grant from the National Science Centre, Poland, under contract UMO-2015/17/B/HS4/00927.

References

Bongaarts, J., (2005). Long-range trends in adult mortality: Models and projection methods, Demography, 42(1), pp. 23–49.

Booth, H., Hyndman, R. J., Tickle, L., De Jong, P., (2006). Lee-Carter mortality forecasting: A multi-country comparison of variants and extensions models, Demographic Research, 15(9), pp. 289–310.

Booth, H., Maindonald, J., Smith, L., (2002). Applying Lee-Carter under conditions of variable mortality decline. Population Studies, 56 (3), 325–333.

Bozik, J. E., Bell, W. R., (1987). Forecasting Age Specific Fertility Using Principal Components, Bureau of the Census Statistical Research Division Washington D.C., CENSUS/SRD/RR-87/19, https://www.census.gov/srd/papers/pdf/rr87-19.pdf.

Brouhns, N., Denuit, M., Vermunt, J. K., (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables, Insurance: Mathematics and Economics, 31(3), pp. 373–393.

Cairns, A. J. G., Blake, D., Dowd, K., (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration, Journal of Risk and Insurance, 73(4), pp. 687–718.

Currie, I. D., Durban, M., Eilers, P. H. C., (2004). Smoothing and forecasting mortality rates, Statistical Modelling, 4(4).

Danesi, I. L., Haberman, S., Millossovich, P., (2015). Forecasting mortality in subpopulations using Lee-Carter type models: A comparison, Insurance: Mathematics and Economics, 62(4), pp. 151–161.

De Jong, P., Tickle, L., (2006). Extending Lee-Carter mortality forecasting. Mathematical Population Studies, 13(1), pp. 1–18.

Diamond, P., (1988), Fuzzy least-squares, Information Sciences, 46(3), pp. 141–157.

Haberman, S., Renshaw, A., (2012). Parametric mortality improvement rate modelling and projecting, Insurance: Mathematics and Economics, 50(3), pp. 309–333.

Heligman, L., Pollard, J. H., (1980). The age pattern of mortality, Journal of the Institute of Actuaries, 170, pp. 49–80.

Horiuchi, S., Coale, A. J., (1990). Age patterns of mortality for older women: An analysis using the age-specific rate of mortality change with age, Mathematical Population Studies, 2(4), 245–267.

Human Fertility Database. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at www.humanfertility.org.

Ishikawa, S., (1997). Fuzzy inferences by algebraic method, Fuzzy Sets and Systems, 87, pp. 181–200.

Koissi, M.-C., Shapiro, A. F., (2006). Fuzzy formulation of the Lee-Carter model for mortality forecasting, Insurance: Mathematics and Economics, 39, pp. 287–309.

Kosiński, W., Prokopowicz, P., Ślęzak, D., (2003). Ordered Fuzzy Numbers, Bull. Polish Acad. Sci. Math., 51, pp. 327–338.

Lee, R. D., Carter, L., (1992). Modeling and forecasting the time series of U.S. mortality, Journal of the American Statistical Association, 87, pp. 659–671.

Milevsky, M. A., Promislow, S. D., (2001). Mortality Derivatives and the Option to Annuitise,Insurance: Mathematics and Economics, 29, pp. 299–318

Pitacco, E., Denuit, M., Haberman, S., Olivieri, A., (2009). Modelling Longevity Dynamics for Pensions and Annnuity Business, Oxford University Press.

Prokopowicz, P., Czerniak, J., Mikołajewski, D., Apiecionek, Ł., Ślęzak, D. (eds.), (2017). Ordered Fuzzy Numbers: Definitions and Operations, Studies in Fuzziness and Soft Computing, vol. 356, Springer Open.

Renshaw, A. E., Haberman, S., (2003). Lee-Carter mortality forecasting with age specific enhancement, Insurance: Mathematics and Economics, 33(2), pp. 255–272.

Renshaw, A., Haberman, S., Hatzopoulos, P., (1996). The modelling of recent mortality trends in United Kingdom male assured lives. British Actuarial Journal, 2, pp. 449–477.

Rossa, A., Socha, L., Szymański, A., (2011). Analiza i modelowanie umieralności w ujęciu dynamicznym (in Polish), University of Lodz Press, Łódź.

Rossa, A., Socha, L., Szymański, A., (2017). Hybrid Dynamic and Fuzzy Models of Mortality. University of Lodz Press.

Szymański, A., Rossa, A., (2014). Fuzzy mortality model based on Banach algebra, International Journal of Intelligent Technologies and Applied Statistics, 7,pp. 241–265.

Szymański, A., Rossa, A., (2017). Improvement of fuzzy mortality model by means of algebraic methods, Statistics in Transition, 18, pp. 701–724.

Tuljapurkar, S., Li, N., Boe, C., (2000). A universal pattern of mortality decline in the G7 countries, Nature, 405, pp. 789–792.

APPENDIX

A.1. The orthogonal expansion

Two vectors $\varphi, \psi \in L^2(0,1)$ are called orthogonal ($\varphi \perp \psi$) if $\langle \varphi, \psi \rangle = 0$ and parallel if one is multiple of the other. If φ and ψ are orthogonal ($\varphi \perp \psi$), then the Pythagorean theorem is satisfied

$$\| \varphi + \psi \|^2 = \| \varphi \|^2 + \| \psi \|^2.$$

A vector φ is called a unit vector if $\| \varphi \| = 1$.

Suppose φ is a unit vector. Then, the projection of ψ in the direction of φ is given by

$$\psi_{\|} = < \varphi, \psi > \varphi$$

and $\psi_{\perp}$, defined as

$$\psi_{\perp} = \psi - < \varphi, \psi > \varphi,$$

is orthogonal to φ.

It is commonly known that a set of vectors $\{P_j\}$ in $L^2(0,1)$ is called an orthonormal set if $\langle P_j, P_k \rangle = 0$ for $j \neq k$ and $\langle P_j, P_j \rangle = 1$.

A.2. The Legendre polynomials as the basis of the orthonormal expansion

Let us consider the set of orthonormal Legendre polynomials. The first four polynomials take the form

$$P_0(u) = 1,$$

$$P_1(u) = \sqrt{3}(2u - 1),$$

$$P_2 = \frac{\sqrt{5}}{2}[3(2u-1)^2 - 1],$$

$$P_3 = \frac{\sqrt{7}}{2}[5(2u-1)^3 - 3(2u-1)],$$

and recursively

$$P_{n+1}(u) = \frac{\sqrt{(2n+1)(2n+3)}}{(n+1)}(2u-1)P_n(u) - \frac{n}{n+1}\sqrt{\frac{2n+3}{2n-1}}P_{n-1}(u).$$

By putting $j = n + 1$ we have for $n = 2, 3, \ldots$

$$P_j(u) = \frac{\sqrt{(2j-1)(2j+1)}}{j}(2u-1)P_{j-1}(u) - \frac{j-1}{j}\sqrt{\frac{2j+1}{2j-3}}P_{j-2}(u).$$

We will use the recursive formula to find the Legendre polynomials P_4, P_5 orthonormal on the interval $[0, 1]$. We get

$$P_4 = \frac{\sqrt{9}}{8}[35(2u-1)^4 - 30(2u-1)^2 + 3],$$

$$P_5 = \frac{\sqrt{11}}{8}[63(2u-1)^5 - 70(2u-1)^3 + 15(2u-1)].$$

For $j = 3$ there is $P_3 = \frac{\sqrt{7}}{2}[5(2u-1)^3 - 3(2u-1)]$.

Let us calculate the scalar products $\langle P_0, P_3\rangle, \langle P_1, P_3\rangle, \langle P_2, P_3\rangle$.

For $P_0(u) = 1,\ P_3 = \frac{\sqrt{7}}{2}[5(2u-1)^3 - 3(2u-1)]$ there is

$$\langle P_0, P_3\rangle = \frac{\sqrt{7}}{2}\left[5\int_0^1 (2u-1)^3\,du - 3\int_0^1 (2u-1)du\right] = 0.$$

For $P_1(u) = \sqrt{3}(2u-1),\ P_3 = \frac{\sqrt{7}}{2}[5(2u-1)^3 - 3(2u-1)]$ we have

$$\langle P_1, P_3\rangle = \frac{\sqrt{21}}{2}\left[5\int_0^1 (2u-1)^4 - 3\int_0^1 (2u-1)^2 du\right] = 0.$$

For $P_2 = \frac{\sqrt{5}}{2}[3(2u-1)^2 - 1], P_3 = \frac{\sqrt{7}}{2}[5(2u-1)^3 - 3(2u-1)]$ we get

$$\langle P_2, P_3\rangle = \frac{\sqrt{35}}{4}\left[15\int_0^1 (2u-1)^5 du - 14\int_0^1 (2u-1)^3 du + 3\int_0^1 (2u-1)du\right] = 0.$$

Hence, it follows that $P_0 \perp P_3,\ \ P_1 \perp P_3,\ \ P_2 \perp P_3$.

Now, let us verify the normality of the element $P_3 \in L_2(0,1)$, i.e. we will verify the equality $\| P_3 \|^2 = 1$. Note that P_3^2 is equal to

$$P_3^2 = \frac{7}{4}[5(2u-1)^3 - 3(2u-1)]^2 = \frac{7}{4}[25(2u-1)^6 - 30(2u-1)]^4 + 9(2u-1)^2].$$

The squared norm of the element $P_3 \in L_2(0,1)$ is as follows

$$\| P_3 \|^2 = \frac{7}{4}\left[25\int_0^1 (2u-1)^6\,du - 30\int_0^1 (2u-1)^4 du + 9\int_0^1 (2u-1)^2 du\right] = \frac{7}{4}\left(\frac{25}{7} - \frac{30}{5} + \frac{9}{3}\right) = 1.$$

Thus, $P_3 \in L_2(0,1)$ belongs to the orthonormal system $\{P_j\}_{j=0}^3$.

A.3. Orthonormal expansions of inverse triangular functions

Let us assume that f, g take the following forms

$$f(u) = a - s(1-u), \quad g(u) = a + s(1-u), \quad u \in [0,1],$$

and let $\{P_j\}$ be an orthonormal set of Legendre polynomials in $L^2(0,1)$.

First, we will find coefficients α_j, β_j for $j = 0, 1$. We have

$$\alpha_0 = \langle P_0, f\rangle = \int_0^1 f(u)du = a - s\int_0^1 (1-u)du = a - \frac{s}{2},$$

$$\beta_0 = \langle P_0, g\rangle = \int_0^1 f(u)du = a + s\int_0^1 (1-u)du = a + \frac{s}{2},$$

$$\alpha_1 = \langle P_1, f\rangle = \int_0^1 \sqrt{3}(2u-1)[a-(1-u)s]du = \frac{s}{2\sqrt{3}},$$

$$\beta_1 = \langle P_1, g\rangle = \int_0^1 \sqrt{3}(2u-1)[a+(1-u)s]du = -\frac{s}{2\sqrt{3}}.$$

Thus, we obtain

$$f(u) = f^{(1)}(u) = \sum_{j=0}^{1} \alpha_j P_j = a - \frac{s}{2} + s\left(u - \frac{1}{2}\right) = a - s(1-u),$$

$$g(u) = g^{(1)}(u) = \sum_{j=0}^{1} \beta_j P_j = a + \frac{s}{2} + s\left(-u + \frac{1}{2}\right) = a + s(1-u).$$

A.4. Orthonormal expansions of inverse exponential functions

Suppose that f, g are expressed as

$$f(u) = c - \tau(-\ln u)^{\frac{1}{2}}, \quad g(u) = c + \nu(-\ln u)^{\frac{1}{2}}, \quad u \in [0,1]. \tag{A.1}$$

First, we will find coefficients α_j, β_j for $j = 0,1,2,3$. We have

$$\alpha_j = \langle P_j, f\rangle = \langle P_j, c + \psi\rangle = \langle P_j, c\rangle + \langle P_j, \psi\rangle,$$

$$\beta_j = \langle P_j, g\rangle = \langle P_j, c + \varphi\rangle = \langle P_j, c\rangle + \langle P_j, \varphi\rangle.$$

For scalar products $\langle P_j, \psi\rangle$ and $\langle P_j, \varphi\rangle$ we need to calculate the integral $\int_0^1 u^j(-\ln u)^{\frac{1}{2}}du$. After some basic calculations we obtain

$$\int_0^1 u^j(-\ln u)^{\frac{1}{2}}du = \frac{\sqrt{\pi}}{2(j+1)^{\frac{3}{2}}}.$$

For $j = 0$, we get $P_0(u) = 1$ and

$$\langle P_0, c\rangle = \langle 1, c\rangle = \int_0^1 c du = c.$$

Thus,

$$\alpha_0 = \langle P_0, f\rangle = \langle 1, c\rangle + \langle 1, f\rangle = c + \int_0^1 \psi(u)du,$$

$$\beta_0 = \langle P_0, g\rangle = \langle 1, c\rangle + \langle 1, g\rangle = c + \int_0^1 \varphi(u)du.$$

Hence, there is

$$\alpha_0 = c - \tau \int_0^1 (-\ln u)^{\frac{1}{2}}\, du,$$

$$\beta_0 = c + \nu \int_0^1 (-\ln u)^{\frac{1}{2}} du.$$

For $j = 0$, we have $\int_0^1 (-\ln u)^{\frac{1}{2}} du = \frac{\sqrt{\pi}}{2}$, and $\alpha_0,\ \beta_0$ can be reduced to

$$\alpha_0 = \int_0^1 f(u)du = c - \tau\frac{\sqrt{\pi}}{2}, \tag{A.2}$$

$$\beta_0 = \int_0^1 g(u)du = c + \nu\frac{\sqrt{\pi}}{2}. \tag{A.3}$$

Using the recursive formula, we can obtain next orthonormal expansion for $j = 1,2,3,\ldots$.

Let us take $j = 1$, then $P_1(u) = \sqrt{3}(2u-1)$ and

$$\langle P_1, c\rangle = \sqrt{3}c \int_0^1 (2u-1)\, du = \sqrt{3}(c-c) = 0.$$

We have also

$$\langle P_1, \psi\rangle = -\tau \int_0^1 P_1(-\ln u)^{\frac{1}{2}}\, du = -\tau\sqrt{3}\int_0^1 (2u-1)(-\ln u)^{\frac{1}{2}}\, du$$
$$= -\tau\frac{\sqrt{3\pi}}{2}\left(\frac{1}{\sqrt{2}} - 1\right),$$
$$\langle P_1, \varphi\rangle = \nu \int_0^1 P_1(-\ln u)^{\frac{1}{2}}\, du = \nu\sqrt{3}\int_0^1 (2u-1)(-\ln u)^{\frac{1}{2}}\, du$$
$$= \nu\frac{\sqrt{3\pi}}{2}\left(\frac{1}{\sqrt{2}} - 1\right).$$

Thus, we receive

$$\alpha_1 = \langle P_1, f\rangle = -\tau\frac{\sqrt{3\pi}}{2}\left(\frac{1}{\sqrt{2}} - 1\right), \tag{A.4}$$

$$\beta_1 = \langle P_1, g\rangle = \nu\frac{\sqrt{3\pi}}{2}\left(\frac{1}{\sqrt{2}} - 1\right). \tag{A.5}$$

For $j = 2$, there is $P_2 = \frac{\sqrt{5}}{2}[3(2u-1)^2 - 1]$ and

$$\alpha_2 = \langle P_2, f\rangle = \langle P_2, c\rangle + \langle P_2, \psi\rangle, \quad \beta_2 = \langle P_2, g\rangle = \langle P_2, c\rangle + \langle P_2, \varphi\rangle,$$

where

$$\langle P_2, c\rangle = c\frac{\sqrt{5}}{2}\int_0^1 (3(2u-1)^2 - 1)du = 2c\sqrt{5} - 3c\sqrt{5} + \frac{3c\sqrt{5}}{2} - \frac{c\sqrt{5}}{2} = 0$$
$$\langle P_2, \psi\rangle = -\tau\int_0^1 P_2(-\ln u)^{\frac{1}{2}}\, du = -\tau\sqrt{5\pi}\left(\frac{1}{\sqrt{3}} - \frac{3}{2\sqrt{2}} + \frac{1}{2}\right),$$
$$\langle P_2, \varphi\rangle = \upsilon\int_0^1 P_2(-\ln u)^{\frac{1}{2}}\, du = \upsilon\sqrt{5\pi}\left(\frac{1}{\sqrt{3}} - \frac{3}{2\sqrt{2}} + \frac{1}{2}\right).$$

Hence,

$$\alpha_2 = \langle P_2, f\rangle = -\tau\sqrt{5\pi}\left(\frac{1}{\sqrt{3}} - \frac{3}{2\sqrt{2}} + \frac{1}{2}\right),$$

$$\beta_2 = \langle P_2, g\rangle = \upsilon\sqrt{5\pi}\left(\frac{1}{\sqrt{3}} - \frac{3}{2\sqrt{2}} + \frac{1}{2}\right).$$

Let us find coefficients α_3 and β_3, i.e.

$$\alpha_3 = \langle P_3, f\rangle = \langle P_3, c\rangle + \langle P_3, \psi\rangle,$$

$$\beta_3 = \langle P_3, g\rangle = \langle P_3, c\rangle + \langle P_3, \varphi\rangle.$$

We have $P_3 = \frac{\sqrt{7}}{2}[5(2u-1)^3 - 3(2u-1)]$ and

$$\langle P_3, c\rangle = c\frac{\sqrt{7}}{2}\int_0^1 (5(2u-1)^3 - 3(2u-1))du = \frac{c\sqrt{7}}{2}(10 - 20 + 12 - 2) = 0,$$

$$\langle P_3, \psi\rangle = -\tau\int_0^1 P_3(-\ln u)^{\frac{1}{2}}\,du = -\tau\sqrt{7\pi}\left(-\frac{5}{\sqrt{3}} + \frac{15}{4\sqrt{2}} - \frac{3}{4\sqrt{2}} + \frac{3}{4}\right).$$

$$\langle P_3, \varphi\rangle = \upsilon\int_0^1 P_3(-\ln u)^{\frac{1}{2}}\,du = \upsilon\sqrt{7\pi}\left(-\frac{5}{\sqrt{3}} + \frac{15}{4\sqrt{2}} - \frac{3}{4\sqrt{2}} + \frac{3}{4}\right)$$

Hence,

$$\alpha_3 = \langle P_3, f\rangle = -\tau\sqrt{7\pi}\left(-\tfrac{5}{\sqrt{3}} + \tfrac{15}{4\sqrt{2}} - \tfrac{3}{4\sqrt{2}} + \tfrac{3}{4}\right), \tag{A.6}$$

$$\beta_3 = \langle P_3, g\rangle = \upsilon\sqrt{7\pi}\left(-\tfrac{5}{\sqrt{3}} + \tfrac{15}{4\sqrt{2}} - \tfrac{3}{4\sqrt{2}} + \tfrac{3}{4}\right). \tag{A.7}$$

Thus, orthonormal expansions of $f(u)$ and $g(u)$ defined in (A.1) are as follows

$$f(u) \cong f^{(3)}(u) = \sum_{j=0}^{3}\alpha_j P_j\,, \quad g(u) \cong g^{(3)}(u) = \sum_{j=0}^{3}\beta_j P_j,$$

where

$$\alpha_0 P_0(u) = c - \tau\frac{\sqrt{\pi}}{2}, \qquad \beta_0 P_0(u) = c + \nu\frac{\sqrt{\pi}}{2},$$

$$\alpha_1 P_1(u) = -\tau\tfrac{3\sqrt{\pi}}{2}\left(\tfrac{1}{\sqrt{2}} - 1\right)(2u-1), \quad \beta_1 P_1(u) = \nu\tfrac{3\sqrt{\pi}}{2}\left(\tfrac{1}{\sqrt{2}} - 1\right)(2u-1),$$

$$\alpha_2 P_2(u) = -\tau\frac{5\sqrt{\pi}}{2}\left(\frac{1}{\sqrt{3}} - \frac{3}{2\sqrt{2}} + \frac{1}{2}\right)[3(2u-1)^2 - 1],$$

$$\beta_2 P_2(u) = \nu\frac{5\sqrt{\pi}}{2}\left(\frac{1}{\sqrt{3}} - \frac{3}{2\sqrt{2}} + \frac{1}{2}\right)[3(2u-1)^2 - 1],$$

$$\alpha_3 P_3(u) = -\tau\frac{7\sqrt{\pi}}{2}\left(-\frac{5}{\sqrt{3}} + \frac{15}{4\sqrt{2}} - \frac{3}{4\sqrt{2}} + \frac{3}{4}\right)[5(2u-1)^3 - 3(2u-1)],$$

$$\beta_3 P_3(u) = \nu\frac{7\sqrt{\pi}}{2}\left(-\frac{5}{\sqrt{3}} + \frac{15}{4\sqrt{2}} - \frac{3}{4\sqrt{2}} + \frac{3}{4}\right)[5(2u-1)^3 - 3(2u-1)].$$

STATISTICS IN TRANSITION new series, September 2021
Vol. 23, No. 3, pp. 59–79, DOI 10.21307/stattrans-2021-027
Received – 10.03.2020; accepted – 05.12.2020

Bayesian estimation and prediction based on Rayleigh record data with applications

Raed R. Abu Awwad[1], **Omar M. Bdair**[2], **Ghassan K. Abufoudeh**[3]

ABSTRACT

Based on a record sample from the Rayleigh model, we consider the problem of estimating the scale and location parameters of the model and predicting the future unobserved record data. Maximum likelihood and Bayesian approaches under different loss functions are used to estimate the model's parameters. The Gibbs sampler and Metropolis-Hastings methods are used within the Bayesian procedures to draw the Markov Chain Monte Carlo (MCMC) samples, used in turn to compute the Bayes estimator and the point predictors of the future record data. Monte Carlo simulations are performed to study the behaviour and to compare methods obtained in this way. Two examples of real data have been analyzed to illustrate the procedures developed here.

Key words: Bayesian estimation and prediction, Rayleigh distribution, record values, Markov Chain Monte Carlo samples.

1. Introduction

Assume we have a sequence of independent and identically distributed (iid) random variables from Rayleigh distribution. The two-parameter Rayleigh distribution with parameters λ and μ has the cumulative distribution function (CDF) and the probability density function (PDF), respectively

$$F(x;\lambda,\mu) = 1 - e^{-\lambda(x-\mu)^2}, x > \mu, \tag{1}$$

and

$$f(x;\lambda,\mu) = \begin{cases} 2\lambda\,(x-\mu)\,e^{-\lambda(x-\mu)^2} & if\ x > \mu, \\ 0 & if\ x \leq 0\,, \end{cases} \tag{2}$$

where λ and μ are the scale and location parameters, respectively, and $\lambda > 0$ and $\mu > 0$. From now on the Rayleigh distribution with parameters λ and μ will be denoted by $Ra(\lambda,\mu)$. The Rayleigh distribution has many applications in reliability, life testing and survival analysis. More details on the Rayleigh distribution can be found in Johnson et al. (1994).

[1]Department of Mathematics, Faculty of Arts and Sciences, University of Petra, Amman, Jordan. E-mail: raed_abuawwad@yahoo.com.

[2]Faculty of Engineering Technology, Al-Balqa Applied University, Amman 11134, Jordan. Department of Mathematics and Statistics, McMaster University, Hamilton, ON L8S 4L8, Canada. E-mail: bdairmb@bau.edu.jo.

[3]Department of Mathematics, Faculty of Arts and Sciences, University of Petra, Amman, Jordan. E-mail: ghassan_math@yahoo.com.

The random variable X_j is called a record (upper record) if $X_j > X_i$ for all $i = 1, 2, ..., j-1$. By convention X_1 is a record. Then, the record times sequence $\{U(n), n \geq 1\}$ is defined as $U(1) = 1$ with probability one, and for $n \geq 2$, $U(n) = \min\{j : j > U(n-1)\}$. The random variables $X_{U(n)}, n \geq 1$ denote the record values from X-sequence. Naturally, record values appear in many real life situations including data related to weather, sports, economics and life-tests. For more details about the applications of record values, we refer the reader to Arnold et al. (1998), Nevzorov (2000), and Gulati and Padgett (2003). Extensive studies for estimating parameters from the Rayleigh distribution based on different types of ordered data are available in the literature, but no attempt has been made for comparing the performances in estimating and predicting under different types of loss functions based on record values. Many authors in the literature worked on record data, among others; Bdair and Raqab (2016) considered the Bayesian prediction of future records from Weibull distribution when one- and two-sequence are used. Raqab et al. (2007) obtained the maximum likelihood estimator and Bayes estimators for the parameters of the Pareto distribution based on the record data. Raqab et al. (2018) studied the estimation and prediction problem of bathtube-shaped distribution based on record values. Madi and Raqab (2004) studied the problem of temperature records as an application to Pareto Bayesian prediction problem. Based on a set of observed records from the exponential distribution, Ahsanullah M. (1980) discussed the problem of predicting the unseen records. Ahmadi and Doostparast (2006) discussed the Bayesian estimation and prediction based on record values for some distributions like Weibull, Pareto and Burr type XII. Bdair and Raqab (2009) studied the mean residual lifetime of record data and many of its mathematical properties.

Bayesian estimation of the distribution's parameters as well as prediction of future records are of natural interest in this context. For estimating θ by a decision δ, we consider three types of loss functions. The first one is a symmetric quadratic loss function, which is given by

$$LF_1(\theta, \delta) = (\theta - \delta)^2.$$

The second one is an alternative to the squared loss function, namely the absolute loss function and it is given by

$$LF_2(\theta, \delta) = |\theta - \delta|.$$

Varian (1975) proposed the LINEX loss function which is more commonly used form of asymmetric loss. LINEX loss function can be defined by

$$LF_3(\theta, \delta) = e^{a^*(\delta-\theta)} - a^*(\delta - \theta) - 1, a^* \neq 0.$$

To perform a Bayesian estimates of the Rayleigh distribution parameters, their prior distributions should be specified. When both parameters λ and μ are unknown, we assume that λ has the gamma prior distribution. The gamma prior distribution of λ denoted by

$Gamma(a,b)$ and is given by

$$\pi_1(\lambda|a,b) = \begin{cases} \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda} & \text{if } \lambda > 0, \\ 0 & \text{if } \lambda \leq 0. \end{cases} \tag{3}$$

Here, the hyper-parameters $a > 0$, $b > 0$, and $\Gamma(a)$ is the gamma function, *i.e.* $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$. The prior of μ $(\pi_2(\mu))$ is assumed with support $(0, x_{U(1)})$. For more details, one may refer to Kundu (2008) and Abu Awwad et al. (2018) and the reference therein.

The remaining sections of the paper are organized as follows. In Section 2, we propose the maximum likelihood estimator (MLE) of the parameters of Rayleigh distribution. In Section 3, we use the Metropolis-Hastings method with normal proposal (see Metropolis, et al. (1953)) and the Gibbs sampling approach to compute the Bayes estimators (BEs) of λ and μ under different loss functions LF_1, LF_2 and LF_3. The implementation of Gibbs sampling and Metropolis-Hastings methods to compute sample-based estimators for the predictive density functions of the future record values based on some current available records is discussed in Section 4. In Section 5, we show numerical data analyses for illustrative purpose. For this, we employ Monte Carlo simulation to compare the BEs with the corresponding maximum likelihood estimates as well as to predict and compare between the predicted values based on the suggested types of loss functions. We conclude the results obtained in this work in Section 6.

2. Maximum likelihood estimation

Let $x_{U(1)}, x_{U(2)}, ..., x_{U(n)}$ be a sequence of n Rayleigh upper record values with respective PDF and CDF given in Eq. (1) and Eq. (2). The likelihood function of this sample, see for example Arnold et al. (1998) and Ahsanullah (2004), is given by

$$\begin{aligned} L(\lambda,\mu|data) &= \prod_{i=1}^{n-1} \frac{f(x_{U(i)}|\lambda,\mu)}{1-F(x_{U(i)}|\lambda,\mu)} f(x_{U(n)}|\lambda,\mu) \\ &= 2^n\lambda^n \prod_{i=1}^{n}(x_{U(i)-\mu})\, e^{-\lambda(x_{U(n)}-\mu)^2}. \end{aligned} \tag{4}$$

The natural logarithm of the likelihood function is

$$\ln L(\lambda,\mu|data) = n\ln 2 + n\ln\lambda + \sum_{i=1}^{n}\ln(x_{U(i)}-\mu) - \lambda(x_{U(n)}-\mu)^2. \tag{5}$$

By equating the partial derivatives of Eq. (5), $\frac{\partial \ln L(\lambda,\mu|data)}{\partial\lambda}$ and $\frac{\partial \ln L(\lambda,\mu|data)}{\partial\mu}$, to zero, we readily conclude the following two normal equations

$$\frac{n}{\lambda} - (x_{U(n)}-\mu)^2 = 0, and \tag{6}$$

$$-\sum_{i=1}^{n}(x_{U(i)}-\mu)^{-1}-2\lambda(x_{U(n)}-\mu)=0. \tag{7}$$

Equations (6) and (7) cannot be solved explicitly to obtain exact solutions for λ and μ, hence fixed point iteration method is employed for that. From Eq. (6), we can find the MLE of λ as a function of μ, say $\hat{\lambda}(\mu)$, as follows

$$\hat{\lambda}(\mu)=\frac{n}{(x_{U(n)}-\mu)^2}. \tag{8}$$

Substituting Eq. (8) in Eq. (5), without adding the constant term, we obtain the natural logarithm of the likelihood function of μ as

$$g(\mu)=-n\ln(x_{U(n)}-\mu)^2+\sum_{i=1}^{n}\ln(x_{U(i)}-\mu). \tag{9}$$

By maximizing Eq. (9) with respect to μ, we get the MLE of μ, say $\hat{\mu}_{MLE}$. Applying the fixed point solution method on Eq.'s (10) and (11) below, we can directly obtain the maximum of Eq. (9).

$$h(\mu)=\mu, \tag{10}$$

where

$$h(\mu)=x_{U(n)}+2n\left(\sum_{i=1}^{n}(x_{U(i)}-\mu)^{-1}\right)^{-1}. \tag{11}$$

Very simple iterative procedure $h(\mu^{(j)})=\mu^{(j+1)}$, where $\mu^{(j)}$ is the j-th iterative, can be used to solve Eq. (10). Once $\hat{\mu}_{MLE}$ is obtained, the MLE of λ, say $\hat{\lambda}_{MLE}$, can be calculated from Eq. (8) as $\hat{\lambda}_{MLE}=\hat{\lambda}(\hat{\mu}_{MLE})$.

3. Bayesian estimation and corresponding CIs

Let us first consider the case when the location parameter μ is known. Based on the n observed upper record data $x_{U(1)},x_{U(2)},...,x_{U(n)}$, and by combining the likelihood function Eq. (4) and the prior density Eq. (3), the marginal density of λ given μ and data can be obtained to be $Gamma(a+n,\ b+(x_{U(n)}-\mu)^2)$ of the form

$$\pi_1(\lambda|\mu,data)=\frac{(b+(x_{U(n)}-\mu)^2)^{a+n}}{\Gamma(a+n)}\,\lambda^{a+n-1}\,e^{-\lambda(b+(x_{U(n)}-\mu)^2)}. \tag{12}$$

Under the squared error loss function LF_1, the BE $\hat{\lambda}_{B_1}$ of λ is the posterior mean which is given by

$$\hat{\lambda}_{B_1}=E_{posterior}(\lambda|\mu,data)=\frac{a+n}{b+(x_{U(n)}-\mu)^2}.$$

Clearly, the BE under LF_1 loss function, $\widehat{\lambda}_{B_1}$ is the same as the the corresponding MLE of λ when Jeffrey's prior ($a = b = 0$) is employed. The median of the posterior density, $\hat{\lambda}_{B_2}$, is the BE of λ in case of absolute error loss function LF_2. Since the median of the posterior density cannot have an explicit expression, a numerical solution is required by solving the following equation in w:

$$\Gamma\left(a+n,\ (b+(x_{U(n)}-\mu)^2)w\right) - \frac{\Gamma(a+n)}{2} = 0,$$

where

$$\Gamma(a,c) = \int_c^{\infty} x^{a-1}\, e^{-x}\, dx,\ a>0, c>0,$$

is the incomplete gamma function. Under the LINEX loss function LF_3 and for any given $a^* \neq 0$, the BE $\hat{\lambda}_{B_3}$ of λ can be computed using the PDF of the gamma distribution as follows:

$$\begin{aligned}
\hat{\lambda}_{B_3} &= -\frac{1}{a^*}\ln\left[E_{posterior}[e^{-a^*\lambda}|data]\right] \\
&= -\frac{1}{a^*}\ln\left[\int_0^{\infty} e^{-a^*\lambda}\ \pi_1(\lambda|\mu,data)\, d\lambda\right] \\
&= -\frac{a+n}{a^*}\ln\left[\frac{b+(x_{U(n)}-\mu)^2}{a^*+b+(x_{U(n)}-\mu)^2}\right].
\end{aligned}$$

Since the posterior distribution of λ given μ and data follows a gamma distribution, a credible interval of λ can be easily obtained using the percentiles from the gamma distribution. In particular, if a is positive integer, then the chi-square table values can be easily used for constructing credible interval for λ.

Now, we consider the case when both parameters λ and μ are unknown. By using the prior distributions $\pi_1(\lambda|a,b)$ and $\pi_2(\mu)$, the joint posterior function of λ and μ is given by

$$\pi(\lambda,\mu|data) = \frac{L(\lambda,\mu|data)\,.\,\pi_1(\lambda|\mu,a,b)\pi_2(\mu)}{\int_0^{\infty}\int_0^{\infty} L(\lambda,\mu|data)\,.\,\pi_1(\lambda|\mu,a,b)\pi_2(\mu)\, d\lambda\, d\mu}. \tag{13}$$

The marginal density of μ is obtained to be

$$\pi(\mu|\lambda,data) \propto \prod_{i=1}^{n}(x_{u(i)}-\mu)\ e^{\lambda(x_{U(n)}-\mu)^2}\ \pi_2(\mu), \tag{14}$$

where $\pi_2(\mu)$ is a prior distribution with support $(0, x_{U(1)})$. Here, we follow the approach suggested by Berger and Sun (1993) that no specific form of prior $\pi_2(\mu)$ on μ is assumed. For more details about this type of prior, the reader is referred to Abu Awwad et al. (2018). Under the squared error loss function LF_1, the BE of $\theta = g(\lambda,\mu)$, a function λ and μ, can

be presented as

$$\hat{\theta}_{B_1} = E_{posterior}(\theta|data) = \int_0^\infty \int_0^\infty \theta\, \pi(\lambda,\mu|data)\, d\lambda\, d\mu.$$

The BE of θ ($\hat{\theta}_{B_2}$), when the absolute error loss function LF_2 is used, is just the median of the posterior distribution, *i.e.*

$$\hat{\theta}_{B_2} = Med_{posterior}(\theta|data).$$

The BE $\hat{\theta}_{B_3}$ of θ, under the LINEX loss function LF_3, can be obtained as

$$\hat{\theta}_{B3} = -\frac{1}{a^*} \ln \left[E_{posterior}(e^{-a^*\theta}|data) \right] = -\frac{1}{a^*} \ln \left[\int_0^\infty \int_0^\infty e^{-a^*\theta}\, \pi(\lambda,\mu|data)\, d\lambda\, d\mu \right].$$

Here, the Bayes point estimators $\hat{\theta}_{B_1}$, $\hat{\theta}_{B_2}$ and $\hat{\theta}_{B_3}$ cannot be obtained in closed forms. It can be easily checked that λ can be generated directly using Eq. (12), while μ cannot be generated directly from Eq. (14). For this, we implement the Metropolis-Hastings (M-H) method (see Metropolis, et al. (1953)) with normal proposal distribution to generate random values of μ from Eq. (14). The MLEs of λ and μ can be considered as initial values. We can apply this method of generation, M times, to obtain $\{(\lambda_i,\mu_i); i = 1,...,M\}$. We use these MCMC samples to obtain the Bayes estimates of $\theta = g(\lambda,\mu)$ and the corresponding credible intervals. The M-H algorithm proceeds as follows.

M-H algorithm for prediction problem:

1. Start with an initial values $(\lambda^{(0)},\mu^{(0)})$ and set k = 1;
2. Given $\mu^{(k-1)}$, generate μ from $\pi(\mu|\text{data})$ appeared in Eq. (14) with the $N(\mu^{(k-1)},S_\mu^2)$ proposal distribution, where S_μ^2 is the variance of μ. The values of μ can be generated as follows:
   a. Generate ζ_k from $\Omega(.|\mu^{(k-1)},S_\mu^2) = N(\mu^{(k-1)},S_\mu^2)$ and u from the uniform distribution $U(0,1)$
   b. If $u < \min(1,\nu)$ then let $\mu^{(k)} = \zeta_k$, else go to (a), where
$$\nu = \frac{\pi(\zeta_k|\text{data})}{\pi(\mu^{(k-1)}|\text{data})} \frac{\Omega(\mu^{(k-1)}|\zeta_k,S_\mu^2)}{\Omega(\zeta_k|\mu^{(k-1)},S_\mu^2)}.$$
3. Given μ, generate λ from $Gamma(a+n,\ b+(x_{U(n)}-\mu)^2)$;
4. Set $k = k+1$.
5. Repeat steps 2-4, M times.

The BE of $\theta = g(\lambda,\mu)$ under the squared error loss function LF_1 is obtained as

$$\widehat{\theta}_{B_1} = \frac{1}{M}\sum_{i=1}^{M} g(\lambda_i,\mu_i).$$

To obtain the BE of $\theta = g(\lambda,\mu)$, under the absolute error loss function LF_2, we compute $\theta_i = g(\lambda_i,\mu_i)$, $i = 1,2,...,M$ and order $\theta_1,\theta_2,...,\theta_M$ as $\theta_{(1)},\theta_{(2)},...,\theta_{(M)}$, then the BE of $\theta = g(\lambda,\mu)$ is $\widehat{\theta}_{B_2} = Median\{\theta_{(1)},\theta_{(2)},...,\theta_{(M)}\}$.

We evaluate the Bayes estimator of $\theta = g(\alpha,\lambda)$ with respect to the LINEX loss function LF_3 with $a^* \neq 0$ as

$$\widehat{\theta}_{B_3} = -\frac{1}{a^*}\ln\left[\frac{1}{M}\sum_{i=1}^{M} e^{-a^* g(\lambda_i,\mu_i)}\right].$$

Obtain the posterior variance of $\theta = g(\lambda,\mu)$ as

$$\hat{V}ar(\theta|data) = \frac{1}{M}\sum_{i=1}^{M}(\theta_i - \hat{\theta}_{B_{1,2,or3}})^2$$

To compute the CI of $\theta = g(\lambda,\mu)$, we order $\theta_1,\theta_2,...,\theta_M$ as $\theta_{(1)},\theta_{(2)},...,\theta_{(M)}$. Then, $(1-\gamma)100\%$ symmetric CI of θ is given by $\left(\theta_{([\frac{M\gamma}{2}])},\theta_{([\frac{M(1-\gamma)}{2}])}\right)$.

4. Bayesian prediction for future records and corresponding PIs

Here, we predict the future unseen records based on a sequence of observed records, under different loss functions LF_1, LF_2 and LF_3 with $a^* \neq 0$, when the one-sample prediction problem is used. Naturally, we can notice the prediction problems in many real life situations such as the prediction of extremes of rainfall, water levels and sea surface. In the past two decades, many improvements have been done to this field. The readers may refer to Ahsanullah (1980) and Nagaraja (1984). Al-Hussaini and Ahmad (2003) studied the Bayesian prediction interval for the future generalized order statistics.
Suppose that we can only notice the first m upper records $\underset{\sim}{x} = (x_{U(1)},x_{U(2)},...,x_{U(m)})$. Our goal is to obtain the Bayes point prediction of unobserved records under different loss functions LF_1, LF_2 and LF_3, as well as to construct the Bayes predictive interval for the nth future upper record $X_{U(n)}$, where $1 \leq m < n$. The posterior predictive density of $X_{U(n)}$ at any point $y > x_{U(m)}$ is given by

$$f^P_{X_{U(n)}|\underset{\sim}{x}}(y|\alpha,\lambda) \quad = \quad E_{posterior}\left[f_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu)\right],$$

where $f_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu)$ is the conditional PDF of $X_{U(n)}$ given the records data $\underset{\sim}{x}$. Applying the Markovian property on the record values, then $f_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu) = f_{X_{U(n)}|x_{U(m)}}(y|\lambda,\mu)$ and

the posterior predictive density of $X_{U(n)}$ at any point $y > x_{U(m)}$ is obtained as

$$\begin{aligned} f^{P}_{X_{U(n)}|\underset{\sim}{x}}(y|\alpha,\lambda) &= E_{posterior}\left[f_{X_{U(n)}|x_{U(m)}}(y|\lambda,\mu)\right] \\ &= \int_0^{\infty}\int_0^{\infty} f_{X_{U(n)}|x_{U(m)}}(y|\lambda,\mu)\,\pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda\,d\mu \\ &= \int_0^{\infty}\int_0^{\infty} \frac{[H(x_{U(n)})-H(x_{U(m)})]^{n-m-1}}{(n-m-1)!}\,\frac{f(x_{U(n)})}{1-F(x_{U(m)})}\,\pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda\,d\mu \end{aligned}$$

where $H(x) = -\ln(1-F(x))$. Using Eq's (1) and (2) and the binomial expansion, we have

$$\begin{aligned} f^{P}_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu) &= \int_0^{\infty}\int_0^{\infty} \frac{2\lambda^{n-m}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i}\,e^{\lambda(x_{U(m)}-\mu)^2} \\ &\quad \times (y-\mu)^{2(n-m-i-\frac{1}{2})}e^{-\lambda(y-\mu)^2}\,\pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda d\mu,\ y > x_{U(m)}. \end{aligned}$$

Under LF_1, the BP of $Y = X_{U(n)}$ can be evaluated as

$$\begin{aligned} X^{BP1}_{U(n)} &= E_{f^P}(Y|\underset{\sim}{x}) \\ &= \int_{x_{U(m)}}^{\infty} y\left[\int_0^{\infty}\int_0^{\infty} \frac{2\lambda^{n-m}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i}\,e^{\lambda(x_{U(m)}-\mu)^2}\right. \\ &\qquad \left.\times\,(y-\mu)^{2(n-m-i-\frac{1}{2})}e^{-\lambda(y-\mu)^2}\,\pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda d\mu\right] dy. \\ &= \int_0^{\infty}\int_0^{\infty} \frac{e^{\lambda(x_{U(m)-\mu})^2}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i} \\ &\times \left[\lambda^{i-\frac{1}{2}}\Gamma\left(n-m-i+\frac{1}{2},\lambda(x_{U(m)}-\mu)^2\right)+\mu\lambda^i\Gamma\left(n-m-i,\lambda(x_{U(m)}-\mu)^2\right)\right] \\ &\quad \pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda d\mu. \end{aligned}$$

Based on the MCMC samples $\{(\lambda_j,\mu_j);\ j=1,2,...,M\}$ obtained in Section 3, a simulation predictor $\hat{X}^{BP_1}_{U(n)}$ of $Y=X_{U(n)}$ can be computed as

$$\begin{aligned}\hat{X}^{BP_1}_{U(n)} &= \frac{1}{M}\sum_{j=1}^{M}\frac{e^{\lambda_j(x_{U(m)}-\mu_j)^2}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu_j)^{2i}\,\lambda_j^{\,i-\frac{1}{2}} \\ &\times \left[\Gamma\left(n-m-i+\frac{1}{2},\lambda_j(x_{U(m)}-\mu_j)^2\right)+\mu_j\lambda_j^i\Gamma\left(n-m-i,\lambda_j(x_{U(m)}-\mu_j)^2\right)\right].\end{aligned} \tag{15}$$

Usually, it is important to predict the first unseen record value $X_{U(m+1)}$, the simulation-based consistent predictor of the first unseen record value can be evaluated by submitting $n=m+1$ in Eq. (15) as

$$\hat{X}^{BP_1}_{U(n)}=\frac{1}{M}\sum_{j=1}^{M}e^{\lambda_j(x_{U(m)}-\mu_j)^2}\left[\lambda_j^{-\frac{1}{2}}\Gamma\left(\frac{3}{2},\lambda_j(x_{U(m)}-\mu_j)^2\right)+\mu_j\Gamma\left(1,\lambda_j(x_{U(m)}-\mu_j)^2\right)\right].$$

Under LF_2, the BP of $Y=X_{U(n)}$ is given by

$$X^{BP2}_{U(n)}=Med_{f^P}(Y|\underset{\sim}{x}),$$

which is obtained by solving the equation

$$\int_{x_{U(m)}}^{X^{BP2}_{U(n)}} f^P_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu)\,dy=\frac{1}{2},$$

or the equation

$$\int_{X^{BP2}_{U(n)}}^{\infty} f^P_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu)\,dy=\frac{1}{2}.$$

Which are equivalent to the simultaneous equations

$$\int_{X^{BP2}_{U(n)}}^{\infty}\left[\int_0^{\infty}\int_0^{\infty}\frac{2\lambda^{n-m}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i}\,e^{\lambda(x_{U(m)}-\mu)^2}\right.$$
$$\left.\times\,(y-\mu)^{2(n-m-i-\frac{1}{2})}e^{-\lambda(y-\mu)^2}\,\pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda\,d\mu\right]dy=\frac{1}{2},$$

and

$$\left[\int_0^\infty\int_0^\infty \frac{1}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i}\,e^{\lambda(x_{U(m)}-\mu)^2}\right.$$
$$\left.\times\ \lambda^i\Gamma\left(n-m-i,\lambda(X_{U(n)}^{BP2}-\mu)^2\right)\ \pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda\,d\mu\right]=\frac{1}{2}.$$

Based on the MCMC samples $\{(\lambda_j,\mu_j);\ j=1,2,...,M\}$ obtained in Section 3, a simulation predictor $\hat{X}_{U(n)}^{BP_2}$ of $Y=X_{U(n)}$ can be obtained by solving the equation, for $\hat{X}_{U(n)}^{BP_2}$

$$\frac{1}{M}\sum_{j=1}^{M}\left[\frac{1}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu_j)^{2i}\,e^{\lambda_j(x_{U(m)}-\mu_j)^2}\right.$$
$$\left.\times\ \lambda_j^i\Gamma\left(n-m-i,\lambda_j(\hat{X}_{U(n)}^{BP_2}-\mu_j)^2\right)\right]=\frac{1}{2}.$$

In the similar way of the BP under the square error loss function, the BP of $Y=X_{U(n)}$ under the the LINEX loss function LF_3 can be obtained as

$$\begin{aligned}
X_{U(n)}^{BP3} &= -\frac{1}{a^*}\ln\left[E_{f^P}(e^{-a^*Y}|\underset{\sim}{x})\right]\\
&= -\frac{1}{a^*}\ln\left[\int_{x_{U(m)}}^\infty e^{-a^*y}\int_0^\infty\int_0^\infty\frac{2\lambda^{n-m}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i}\right.\\
&\quad\left.\times\ e^{\lambda(x_{U(m)}-\mu)^2}(y-\mu)^{2(n-m-i-\frac{1}{2})}e^{-\lambda(y-\mu)^2}\ \pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda\,d\mu\,dy\right]\\
&= -\frac{1}{a^*}\ln\left[\int_0^\infty\int_0^\infty\frac{\lambda^{n-m-1}e^{-a^*\mu+\frac{a^{*2}}{4\lambda}}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i}\right.\\
&\quad\times e^{\lambda}(x_{U(m)}-\mu)^2\sum_{k=0}^{2(n-m-i-\frac{1}{2})}\binom{2(n-m-i-\frac{1}{2})}{k}(-1)^k(\frac{a^*}{2\lambda})^k\\
&\quad\left.\times\frac{\Gamma\left(n-m-i-k,\lambda((x_{U(m)}-\mu)+\frac{a*}{2\lambda})\right)}{\lambda^{n-m-i-k-1}}\ \pi(\lambda,\mu|\underset{\sim}{x})\,d\lambda\,d\mu\right]. \qquad (16)
\end{aligned}$$

Based on the MCMC samples $\{(\lambda_j,\mu_j);\ j=1,2,...,M\}$, a simulation predictor $\hat{X}^{BP_3}_{U(n)}$ of $Y=X_{U(n)}$ can be obtained as

$$\begin{aligned}\hat{X}^{BP_3}_{U(n)} &= -\frac{1}{a^*}\ln\Bigg[\frac{1}{M}\sum_{j=1}^{M}\frac{\lambda_j^{n-m-1}e^{-a^*\mu_j+\frac{a^{*2}}{4\lambda_j}}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu_j)^{2i}\\ &\quad\times e^{\lambda_j(x_{U(m)}-\mu_j)^2}\sum_{k=0}^{2(n-m-i-\frac{1}{2})}\binom{2(n-m-i-\frac{1}{2})}{k}(-1)^k(\frac{a^*}{2\lambda_j})^k\\ &\quad\times\frac{\Gamma\Big(n-m-i-k,\lambda_j((x_{U(m)}-\mu_j)+\frac{a*}{2\lambda_j})\Big)}{\lambda_j^{n-m-i-k-1}}\Bigg].\end{aligned}\tag{17}$$

The method for obtaining prediction intervals for the nth record value $Y=X_{U(n)}$, $1\le m<n$ under different loss functions depends on the predictive survival function of $Y=X_{U(n)}$ at any point $y>x_{U(m)}$, which is defined as follows:

$$S^P_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu) = E_{posterior}\left(S_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu)\right),$$

where $S_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu)$ is the survival function of $Y=X_{U(n)}$. Based on the Markovian property of record values, we have

$$\begin{aligned}S_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu) &= S_{X_{U(n)}|x_{U(m)}}(y|\lambda,\mu)\\ &= \int_y^\infty f_{X_{U(n)}|x_{U(m)}}(z|\lambda,\mu)\\ &= \frac{2\lambda^{n-m}}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i}\,e^{\lambda(x_{U(m)}-\mu)^2}\\ &\quad\times\int_y^\infty(z-\mu)^{2(n-m-i-\frac{1}{2})}e^{-\lambda(z-\mu)^2}\,dz\\ &= \frac{1}{(n-m-1)!}\sum_{i=0}^{n-m-1}\binom{n-m-1}{i}(-1)^i\,(x_{U(m)}-\mu)^{2i}\,e^{\lambda(x_{U(m)}-\mu)^2}\\ &\quad\times\frac{\Gamma\big(n-m-i,\lambda(y-\mu)^2\big)}{\lambda^{-i}}\end{aligned}\tag{18}$$

The predictive survival function of $Y = X_{U(n)}$ at any point $y > x_{U(m)}$, is then

$$S^P_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda,\mu) = \int_0^\infty \int_0^\infty \left[\frac{1}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i \, (x_{U(m)} - \mu)^{2i} \times \, e^{\lambda (x_{U(m)} - \mu)^2} \frac{\Gamma\left(n-m-i, \lambda (y-\mu)^2\right)}{\lambda^{-i}} \right] \pi(\lambda, \mu | \underset{\sim}{x}) \, d\lambda \, d\mu. \tag{19}$$

Eq. (19) cannot be evaluated analytically. Based on the MCMC samples $\{(\lambda_j, \mu_j); j = 1, ..., M\}$ and under the square error loss function LF_1, the estimate of the predictive survival function for $X_{U(n)}$ can be written as

$$\hat{S}^P_{X_{U(n)}|\underset{\sim}{x}}(y) = \frac{1}{M} \sum_{j=1}^{M} \left[\frac{1}{(n-m-1)!} \sum_{i=0}^{n-m-1} \binom{n-m-1}{i} (-1)^i \, (x_{U(m)} - \mu_j)^{2i} \, e^{\lambda_j (x_{U(m)} - \mu_j)^2} \times \frac{\Gamma\left(n-m-i, \lambda_j (y-\mu_j)^2\right)}{\lambda_j^{-i}} \right].$$

Under the absolute error loss function LF_2, and based on the MCMC samples $\{(\lambda_i, \mu_i); i = 1, ..., M\}$, we find the estimate predictive survival function of $X_{U(n)}$ as follows: Evaluate Eq. (18) for each $\{(\lambda_i, \mu_i), i = 1, ..., M\}$ to get S_1 , S_2 , ... , S_M, where $S_i = S_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda_i, \mu_i)$. Order S_1 , S_2 , ... , S_M to get $S_{(1)} < S_{(2)} < ... < S_{(M)}$, then the estimate predictive survival function of $X_{U(n)}$ is

$$\hat{S}^P_{X_{U(n)}|\underset{\sim}{x}}(y) = Median\left[S_{(1)} \, , \, S_{(2)} \, , \, ... \, , \, S_{(M)}\right]$$

Under the LINEX loss function LF_3, the estimate predictive survival function of $X_{U(n)}$ is obtained as

$$\hat{S}^P_{X_{U(n)}|\underset{\sim}{x}}(y) = -\frac{1}{a^*} \ln \left[\frac{1}{M} \sum_{j=1}^{M} e^{-a^* S_{X_{U(n)}|\underset{\sim}{x}}(y|\lambda_j, \mu_j)} \right].$$

Consider

$$\hat{S}^P_{X_{U(n)}|\underset{\sim}{x}}(L) = 1 - \frac{\gamma}{2}, \tag{20}$$

and

$$\hat{S}^P_{X_{U(n)}|\underset{\sim}{x}}(U) = \frac{\gamma}{2}. \tag{21}$$

Solving the non-linear equations (20) and (21) for L and U under different loss functions

LF_1, LF_2 and LF_3, give the $(1-\gamma)100\%$ prediction interval for $X_{U(n)}$, $n > m$. We need to apply a suitable numerical technique to solve these non-linear equations as they cannot be solved analytically.

5. Simulation and data analysis

Here, we conduct a simulation study to examine the behaviour of the MLEs and BEs as well as BPs that developed in the previous sections under different loss functions based on record data and study a real life examples with the Rayleigh fitting distribution. All computations are performed using Mathematica 11 software.

5.1. Simulation study

In this simulation, the values of the Rayleigh parameters are considered as $\lambda = 2$, $\mu = 1$ to generate record data from $Ra(\lambda, \mu)$. The first n observed records are generated by using the transformation:

$$X_{U(k)} = \left(\frac{\sum_{i=1}^{k} e(i)}{\mu} \right)^{\frac{1}{\lambda}}, k = 1, 2, ..., n,$$

where $\{e(i), i \geq 0\}$ is a sequence of *iid* $Exp(1)$, [see Arnold et al. (1998), p.20]. For the Rayleigh parameters λ and μ, we have computed the BEs, under the three different loss functions; LF_1, LF_2 and LF_3 for some values of a^* $(0.1, 1.0, 5)$. To compute the different BEs we have assumed $\pi_1(\lambda)$, the prior of λ, has gamma density function with the shape and scale parameters c and d, respectively. Regarding the computations of BEs, we consider two types of prior for both λ and μ; Prior 0: the non-informative prior $(i.e.\ a = b = c = d = 0)$ and Prior 1: the informative prior $(i.e.\ a = b = 1, c = 2, d = 1)$. We have computed the mean squared errors (MSEs) for BEs and MLEs based on 1000 replications to compare their performances under different number of observed records. The CIs for the Rayleigh parameters λ and μ are also computed. In the prediction problem and based on observed sequences of record data, we compute the point predictors and 95% PIs for the future n^{th} record $X_{U(n)}$ for Prior 1 under the suggested loss functions; LF_1, LF_2 and LF_3 and for the values of a^* $(0.1, 1.0, 5.0)$. The prediction computations are conducted for the following cases of sample sizes $n = 3, n = 5$, and $n = 7$.

In Table 1, we present the MLEs as well as the BEs of the scale and location parameters of the Rayleigh distribution λ and μ, under the different loss functions used in this paper, when Prior 0 is used. Also, we present the MSEs of MLEs and BEs for the scale and location parameters λ and μ. MCMC samples are used to compute the MSEs based on $M = 1000$ replications. In Table 2, we present the BEs of λ and μ, under the different loss functions, when Prior 1 is used. In Table 3, we show numerical comparisons between the average lengths of the credible intervals of λ and μ when Prior 0 and 1 are used for all of the considered cases.

Table 1: MLEs and BEs (Bayes Estimates) with respect to different loss functions when Prior 0 is used, for $\lambda = 2$ and $\mu = 1$.

Cases		MLE	Sq. err. Bayes 1	Abs. err. Bayes 2	$a^* = 0.1$ Bayes 3	$a^* = 1.0$ Bayes 4	$a^* = 5$ Bayes 5
	λ	3.5966	1.9841	1.9761	1.9784	1.9738	1.9537
$n = 3$		(0.9098)	(0.0206)	(0.0208)	(0.0206)	(0.0207)	(0.0215)
	μ	1.2771	0.9165	0.9069	0.9069	0.8991	0.8668
		(0.0210)	(0.0163)	(0.0170)	(0.0164)	(0.0166)	(0.0191)
	λ	2.1131	1.9695	1.9566	1.9639	1.9594	1.9400
$n = 5$		(0.0971)	(0.0202)	(0.0206)	(0.0203)	(0.0203)	(0.0211)
	μ	1.2794	0.9488	0.9496	0.9389	0.9307	0.8953
		(0.0186)	(0.0169)	(0.0178)	(0.0170)	(0.0173)	(0.0203)
	λ	1.5667	1.9402	1.9160	1.9351	1.9310	1.9140
$n = 7$		(0.0521)	(0.0183)	(0.0190)	(0.0183)	(0.0184)	(0.0190)
	μ	1.2753	0.9629	0.9652	0.9530	0.9448	0.9089
		(0.0195)	(0.0171)	(0.0180)	(0.0172)	(0.0175)	(0.0205)

Note: The first entry represents the average estimate and the second entry is the MSE.

Table 2: BEs with respect to different loss functions when Prior 1 is used, for $\lambda = 2$ and $\mu = 1$.

Cases		Sq. err. Bayes 1	Abs. err. Bayes 2	$a^* = 0.1$ Bayes 3	$a^* = 1.0$ Bayes 4	$a^* = 5$ Bayes 5
	λ	1.9775	1.9687	1.9719	1.9674	1.9479
$n = 3$		(0.0202)	(0.0204)	(0.0202)	(0.0203)	(0.0210)
	μ	0.9521	0.9512	0.9428	0.9352	0.9020
		(0.0161)	(0.0170)	(0.0162)	(0.0164)	(0.0190)
	λ	1.9620	1.9464	1.9567	1.9523	1.9339
$n = 5$		(0.0194)	(0.0198)	(0.0195)	(0.0195)	(0.0202)
	μ	0.9614	0.9695	0.9526	0.9452	0.9123
		(0.0153)	(0.0159)	(0.0154)	(0.0156)	(0.0180)
	λ	1.9352	1.9092	1.9302	1.9262	1.9097
$n = 7$		(0.0179)	(0.0188)	(0.0180)	(0.0180)	(0.019)
	μ	0.9853	0.9879	0.9783	0.9726	0.9468
		(0.0123)	(0.0131)	(0.0124)	(0.0125)	(0.0141)

Table 3: Average CI lengths (AL) and coverage percentage (CP).

Cases		Prior 0		Prior 1	
		AL	CP	AL	CP
$n = 3$	λ	0.7468	0.96	0.7249	0.95
	μ	0.6377	0.96	0.5697	0.96
$n = 5$	λ	0.7265	0.94	0.6789	0.93
	μ	0.6343	0.95	0.5628	0.93
$n = 7$	λ	0.6972	0.95	0.6610	0.95
	μ	0.6273	0.96	0.5585	0.92

From Tables 1 and 2, it is clear that as n increases the performances of MLEs of λ and μ become better in terms of the MSEs. Also, we observe that the Bayes estimates of λ and μ obtained by using Prior 0 and with respect to different loss functions LF_1, LF_2 and LF_3, are quite close to each other and are much better than the MLEs of λ and μ in terms of the MSEs for all considered cases. It can also noticed that the Bayes estimators of λ and μ obtained by using Prior 1 (informative prior) are much better than the Bayes estimators of λ and μ obtained by using Prior 0 (non-informative prior) in terms of the MSEs in most of the considered cases. Moreover, it is worth noting here that the BEs based on the squared error loss function (LF_1) are much better than other BEs that depends on other loss functions based on the reported MSEs values. In Table 3, we observe that the average lengths of the credible intervals for λ and μ, when Prior 1 is used, become smaller as expected, and decrease as n increases. For both Prior 0 and 1, the simulated probabilities for 0.95 are quite close to 0.95.

In Table 4, we present the point predictors and the corresponding 95% PIs for the future n^{th} record $X_{U(n)}, 1 \leq m < n$, based on set of observed records of size m, for all considered cases and for all used loss functions LF_1, LF_2 and LF_3 with many choices of $a^* : 0.1, 1.0, 5.0$. The simulated point predictors and 95% PIs are computed based on MCMC samples $\{(\lambda_i, \mu_i), i = 1, 2, ..., M\}$ and $M = 1000$. In this table the first three future n^{th} records after the last observed record are computed. It is observed from Table 4 that the predicted values for the future records $X_{U(n)}$ (unobserved record) under different loss functions, are quite close to each other and fall in their corresponding 95% PIs. It can be also noticed that the PIs computed based on LINEX loss function (LF_3) when $a^* = 0.5$ are better than other PIs in terms of the length of the PIs reported in the table. Also, as expected, the PIs lengths increase as n increases for given values of m.

5.2. Data analysis

Example 1 (real data):

In this example, we analyze the data of thirty successive March precipitation (in inches) observations. These data are presented in Hinkley (1977), pp. 67-69. The data set is:

```
0.77  1.74  0.81  1.20  1.95  1.20  0.47  1.43  3.37
2.2   3     3.09  1.51  2.1   0.52  1.62  1.31  0.32
0.59  0.81  2.81  1.87  1.18  1.35  4.75  2.48  0.96
1.89  0.90  2.05
```

The Kolmogorov-Smirnov (KS) distance is found to be 0.0770 and the corresponding p-value is 0.9900. Therefore, KS indicates that the Rayleigh distribution can be used to analyze these data. Moreover, graphical tools of empirical and theoretical CDFs and the Q-Q plot given in Figures 1 and 2, give a very good evidence that the Rayleigh distribution fits the data very well. From these data, we have $n = 5$ observed upper record values: 0.77, 1.74, 1.95, 3.37 and 4.75.

Table 4: Point predictors and the corresponding PIs for future records $X_{U(n)}, 1 \leq m < n$ based on some observed records.

Cases	$X_{U(n)}$	Loss function	Predicted	95% PIs
$m=3$	$X_{U(4)}$	LF_1	2.4043	(2.2988, 3.2135)
		LF_2	2.4699	(2.3009, 3.4610)
		$LF_3(a^*=0.1)$	2.5331	(2.2987, 3.0233)
		$LF_3(a^*=1.0)$	2.5241	(2.2987, 2.9080)
		$LF_3(a^*=5.0)$	2.4921	(2.2987, 2.7335)
	$X_{U(5)}$	LF_1	2.6253	(2.3511, 3.6133)
		LF_2	2.6903	(2.3628, 3.5598)
		$LF_3(a^*=0.1)$	2.7460	(2.3509, 3.2943)
		$LF_3(a^*=1.0)$	2.7308	(2.3507, 3.1122)
		$LF_3(a^*=5.0)$	2.6749	(2.3500, 2.8980)
	$X_{U(6)}$	LF_1	2.8236	(2.4329, 3.9355)
		LF_2	2.8870	(2.4643, 3.8435)
		$LF_3(a^*=0.1)$	2.9381	(2.4322, 3.5032)
		$LF_3(a^*=1.0)$	2.9181	(2.4317, 3.2669)
		$LF_3(a^*=5.0)$	2.8434	(2.4291, 3.0304)
$m=5$	$X_{U(6)}$	LF_1	3.1282	(3.0323, 4.0164)
		LF_2	3.3455	(3.0327, 3.7852)
		$LF_3(a^*=0.1)$	3.3799	(3.0321, 3.8923)
		$LF_3(a^*=1.0)$	3.3715	(3.0319, 3.8292)
		$LF_3(a^*=5.0)$	3.3378	(3.0311, 3.6714)
	$X_{U(7)}$	LF_1	3.3592	(3.1533, 4.3148)
		LF_2	3.5844	(3.2271, 4.3505)
		$LF_3(a^*=0.1)$	3.6084	(3.1516, 4.1179)
		$LF_3(a^*=1.0)$	3.5980	(3.1501, 4.0403)
		$LF_3(a^*=5.0)$	3.5558	(3.1432, 3.8750)
	$X_{U(8)}$	LF_1	3.5373	(3.2850, 4.5489)
		LF_2	3.7616	(3.4005, 4.5491)
		$LF_3(a^*=0.1)$	3.7851	(3.2822, 4.2852)
		$LF_3(a^*=1.0)$	3.7736	(3.2797, 4.1954)
		$LF_3(a^*=5.0)$	3.7269	(3.2680, 4.0173)
$m=7$	$X_{U(8)}$	LF_1	3.7850	(3.6146, 4.3022)
		LF_2	3.7377	(3.6166, 4.4392)
		$LF_3(a^*=0.1)$	3.7909	(3.6146, 4.2066)
		$LF_3(a^*=1.0)$	3.7872	(3.6146, 4.1380)
		$LF_3(a^*=5.0)$	3.7728	(3.6146, 4.0217)
	$X_{U(9)}$	LF_1	3.9596	(3.6527, 4.6313)
		LF_2	3.9096	(3.6504, 4.4205)
		$LF_3(a^*=0.1)$	3.9616	(3.6526, 4.4628)
		$LF_3(a^*=1.0)$	3.9547	(3.6526, 4.3459)
		$LF_3(a^*=5.0)$	3.9278	(3.6522, 4.1901)
	$X_{U(10)}$	LF_1	4.1248	(3.7153, 4.9056)
		LF_2	4.0737	(3.7120, 4.6262)
		$LF_3(a^*=0.1)$	4.1233	(3.7150, 4.6696)
		$LF_3(a^*=1.0)$	4.1137	(3.7147, 4.5100)
		$LF_3(a^*=5.0)$	4.0758	(3.7135, 4.3282)

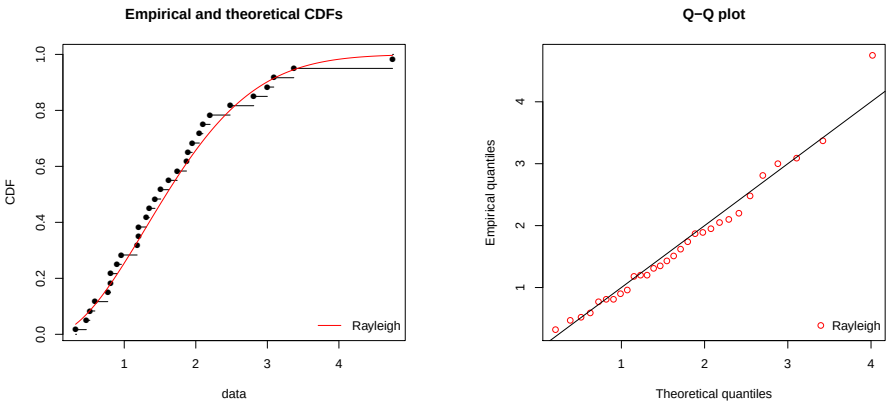


Figure 1: Empirical and fitted distribution functions and Q-Q Plots for data set of example 1.

Example 2 (real data):

In this example, we analyze the survival times in (days) of a group of size $m = 16$ lung cancer patients (Lawless [1982, p. 319]) were considered as follows:

```
6.96   9.30   6.96   7.24   9.30   4.90   8.42   6.05
10.18  6.82   8.58   7.77   11.94  11.25  12.94  12.94
```

From these data, we have $n = 5$ observed upper record values: 6.96, 9.30, 10.18, 11.94 and 12.94. Soliman and Al-Aboud (2008) showed that the Rayleigh distribution fits the observed record values well. Seo and Kim (2018) used this real example to apply an objective Bayesian method under the observed upper record values.

For the above mentioned examples, we compute the BEs based on different loss functions: LF_1, LF_2 and LF_3 and for the values a^* $(0.1, 1.0, 5.0)$. The results are presented in Tables 5 and 6. We can simply see from Tables 5 and 6 that all the estimates are quite close to each other. Furthermore, we obtain the 95% credible intervals for λ and μ, respectively for both examples. For example 1, the 95% CI are given by $(1.5457, 2.4826)$ and $(0.5693, 1.4614)$. For example 2, they are $(1.5473, 2.4828)$ and $(0.6092, 1.4683)$. It can be noticed that the BEs of λ and μ are falling in their credible intervals.

Also, we consider the prediction of the 6^{th}, 7^{th} and 8^{th} future records. The predicted values and the 95% PIs for the 6^{th}, 7^{th} and 8^{th} future records are presented in Tables 7 and 8, for examples 1 and 2, respectively. It is observed that all predicted values, under different loss functions, are all ordered and fall in their corresponding prediction intervals.

Table 5: BEs based on different loss functions for example 1.

	LF_1	LF_2	$LF_3(a^*=0.1)$	$LF_3(a^*=1.0)$	$LF_3(a^*=5.0)$
λ	2.0848	2.1239	2.0636	2.0455	1.9614
μ	1.0215	1.0263	0.9867	0.9568	0.8284

Table 6: BEs based on different loss functions for example 2.

	LF_1	LF_2	$LF_3(a^*=0.1)$	$LF_3(a^*=1.0)$	$LF_3(a^*=5.0)$
λ	2.0841	2.1229	2.0628	2.0446	1.9608
μ	1.0774	1.0899	1.0457	1.0177	0.8874

Table 7: Point predictors and PIs for the 6^{th}, 7^{th} and 8^{th} future records for example 1.

Number of observed records	$X_{U(n)}$	Loss function	Predicted values	95% PIs
$m=5$	$X_{U(6)}$	LF_1	4.8320	(4.7548, 5.5207)
		LF_2	4.8840	(4.7559, 5.4676)
		$LF_3(a^*=0.1)$	4.9453	(4.7548, 5.3315)
		$LF_3(a^*=0.5)$	4.9418	(4.7548, 5.2064)
		$LF_3(a^*=1.0)$	4.9278	(4.7547, 5.0566)
	$X_{U(7)}$	LF_1	4.9840	(4.7928, 5.8867)
		LF_2	5.0593	(4.8058, 5.7628)
		$LF_3(a^*=0.1)$	5.1235	(4.7926, 5.5638)
		$LF_3(a^*=0.5)$	5.1169	(4.7925, 5.3605)
		$LF_3(a^*=1.0)$	5.0908	(4.7918, 5.1797)
	$X_{U(8)}$	LF_1	5.1589	(4.8531, 6.1875)
		LF_2	5.2227	(4.8893, 5.9981)
		$LF_3(a^*=0.1)$	5.2881	(4.8523, 5.7464)
		$LF_3(a^*=0.5)$	5.2788	(4.8518, 5.4798)
		$LF_3(a^*=1.0)$	5.2418	(4.8490, 5.2816)

Table 8: Point predictors and PIs for the 6^{th}, 7^{th} and 8^{th} future records for example 2.

Number of observed records	$X_{U(n)}$	Loss function	Predicted values	95% PIs
$m=5$	$X_{U(6)}$	LF_1	13.4808	(12.9604, 15.6118)
		LF_2	13.4851	(12.9617, 15.6530)
		$LF_3(a^*=0.1)$	13.6791	(12.9604, 15.5010)
		$LF_3(a^*=0.5)$	13.6638	(12.9603, 15.4030)
		$LF_3(a^*=1.0)$	13.6036	(12.9603, 15.0070)
	$X_{U(7)}$	LF_1	14.1850	(13.1308, 16.7771)
		LF_2	14.2062	(13.1448, 16.8060)
		$LF_3(a^*=0.1)$	14.3694	(13.1307, 16.5900)
		$LF_3(a^*=0.5)$	14.3428	(13.1306, 16.4170)
		$LF_3(a^*=1.0)$	14.2361	(13.1301, 16.6830)
	$X_{U(8)}$	LF_1	14.8853	(13.4155, 17.7224)
		LF_2	14.8836	(13.4533, 17.7340)
		$LF_3(a^*=0.1)$	15.0188	(13.4150, 17.4670)
		$LF_3(a^*=0.5)$	14.9836	(13.4146, 17.2250)
		$LF_3(a^*=1.0)$	14.8406	(13.4128, 16.6900)

6. Conclusion

In this work, we have considered the problem of classical and Bayesian estimations of the Rayleigh record model. We have also developed a Bayesian approach to compute the point future records as well as their corresponding prediction intervals. For comparison purposes, we have employed Markov Chain Monte Carlo (MCMC) samples generated from the Rayleigh record model to compute the Bayesian estimators and the point predictors of the future data. Monte Carlo simulations are used to study the behaviour of the obtained methods. Also, two real data examples have been analyzed to illustrate the procedures developed in this article.

Acknowledgements

The authors would like to thank the reviewers for their valuable notes and comments that lead to improve this version of the manuscript.

References

Abu Awwad, R. R., Bdair, O. M. and Abufoudeh, G. K., (2018). One and Two-Sample Prediction for the Progressively Censored Rayleigh Residual Data. *Journal of Statistical Theory and Practice*, 12(4), pp. 669–687.

Ahmadi, J. and Doostparast, M., (2006). Bayesian estimation and prediction for some life distributions based on record values. *Statistical Papers*, 47(3), pp. 373–392.

Arnold, B. C, Balakrishnan, N. and Nagaraja, H. N., (1998). Records. *John Weily, New York.*

Ahsanullah, M., (1980). Linear prediction of record values for the two parameter exponential distributrion. *Annals of the Institute of Statistical Mathematics*, 32, pp. 363–368.

Ahsanullah, M., (2004). Record Values-Theory and Applications. *University Press of America, New York.*

Al-Hussaini, E. K. and Ahmad, A. A., (2003). On Bayesian interval prediction of future records. *Test*, 12, pp. 79–99.

Bdair, O.M. and Raqab, M. Z., (2009). On the mean residual waiting time of records. *Statistics & Decisions International mathematical journal for stochastic methods and models*, 27(3), pp. 249-–260,
DOI: https://doi.org/10.1524/stnd.2009.1050.

Bdair, O. M. and Raqab, M. Z., (2016). One-sequence and two-sequence prediction for future Weibull records. *Journal of Statistical Theory and Applications*, 15(4), pp. 345–366.

Berger, J. O. and Sun., D., (1993). Bayesian analysis for the poly-Weibull distribution. *Journal of the American Statistical Association*, 88(424), pp. 1412-–1418.
DOI:10.1080/01621459.1993.10476426.

David Hinkley, (1977). On quick choice of power transformation. *Journal of the Royal Statistical Society Series C*, 26(1), pp. 67–69.

Gulati, S. and Padgett, W. J., (2003). Parametric and nonparametric inference from record-breaking data. *Springer-Verlag, New York.*

Johnson, N. L, Kotz, S. and Balakrishnan, N., (1994). Continuous univariate distribution. 2-nd edition. *Wiley and Sons, New York.*

Jung In Seo. and Yongku Kim, (2018). Objective Bayesian inference based on upper record values from Rayleigh distribution . *Communications for Statistical Applications and Methods*, 25(4), pp. 411—430.

Kundu, D., (2008). Bayesian inference and life testing plan for the Weibull distribution in presence of progressive censoring. *Technometrics*, 50, pp. 144–154.

Lawless, J.F., (1982). Statistical models and methods for lifetime data. 2nd Edition, *Wiley, New York*.

Madi, M. T. and Raqab, M. Z., (2004). Bayesian prediction of temperature records using the Pareto model. *Environmetrics*, 15, pp. 701–710.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E., (1953). Equations of State Calculations by Fast Computing Machines. *Journal Chemical Physics*, 21, pp. 1087–1091.

Nagaraja, H. N., (1984). Asymptotic linear prediction of extreme order statistics. *Annals of the Institute of Statistical Mathematics*, 36, pp. 289–299.

Nevzorov, V. B., (2000). Records: Mathematical Theory (English Translation). *American Mathematical Society, Providence, Rhode Island.*

Raqab, M. Z., Ahmadi, J. and Doostparast, M. (2007). Statistical inference based on record data from Pareto model. *Statistics*, 42(2), pp. 105–118.

Raqab, M. Z., Bdair, O. M. and Al-Aboud, F. M., (2018). Inference for the two-parameter bathtub-shaped distribution based on record data. *Metrika*, 81(3), pp. 229—253.

Rayleigh, L., (1880). On the resultant of a large number of vibrations of same pitch and of arbitrary phase. *Philosophical Magazine*, 10, pp. 73-–78.

Soliman, A. A. and Al-Aboud, F. M., (2008). Bayesian inference using record values from Rayleigh model with application. *European Journal of Operational Research*, 185(2), pp. 659–672.

Varian, H. R., (1975). A Bayesian approach to real estate assessment. *Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage. North Holland, Amsterdam*, pp. 195–208.

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 81–97, DOI 10.21307/stattrans-2021-028
Received – 18.05.2020; accepted – 09.12.2020

Trade potential under the SAFTA between India and other SAARC countries: the augmented gravity model approach

Vipin Sharma[1], **Vinod Kumar**[2]

ABSTRACT

The study attempts to analyse India's trade potential with other SAARC member states under the SAFTA agreement by means of the augmented gravity model, at annual frequency from 1992 to 2019 in general and from 2004 to 2019 in particular. The findings of this paper prove that the intra-regional trade volumes between SAARC countries can be increased and encouraged. Moreover, the research shows that it is important to introduce structural reforms aiming to boost trade with non-member states. It would be advisable for researchers to take into account the effect locational and infrastructural advantages have on transport costs through the application of a gravity model. Previous research has also demonstrated that the augmented gravity model may prove helpful in explaining some key features of South Asian trade, which traditional gravity models fail to do.

Key words: Cooperation/integration, augmented gravity model, panel data, trade potential, SAARC, SAPTA, SAFTA.

1. Introduction

Globalization has led to many economic activities both at the national and international levels. It has also brought fundamental changes in these economic activities. Economic integration implies close cooperation among member countries and the removal of all types of barriers in intra-regional trade. The SAARC, a regional bloc of 8 countries is a good example of economic cooperation in South Asia.

SAFTA agreement under SAARC, a collective effort of 8 participating countries, aims to enhance their intra-regional trade. There was a perceptible improvement in India's trade performance under SAFTA. It is clear from the rise in trade to gross domestic product (GDP) ratio. From 23% (1991-2003) in the pre-SAFTA, this ratio increased to 48.56% period under SAFTA (2004-2019). India is the largest and more

[1] Assistant Professor (Economics), University School of Business Chandigarh University, Mohali Punjab, India. E-mail: vipinsharmakaushik@gmail.com. ORCID: https://orcid.org/0000-0002-4215-9808.
[2] Panjab University Regional Centre, Sri Muktsar Sahib, India. E-mail: vinod.chd@rediffmail.com. ORCID: https://orcid.org/0000-0002-3234-3855.

developed country among SAARC countries. Its exports as a proportion of SAARC exports rose from 57.90% in 1992 to 67.53% in 2004 and further to 77.93% in 2019. Therefore, it can enhance the volume of total exports to and imports from. In this regard, it would be appropriate to estimate India's trade potential under SAFTA.

2. Review of literature

Velde, Dirk William Te (2011) analyzed how regional integration was important for convergence and growth in developing nations. He used panel data and studied 100 counties over the period 1970-2004. They made use of various analytical techniques both at the micro and macro levels. Among these techniques were "Regional Integration Index, and β and σ-convergence tests". The study found that regional integration did not lead to rapid growth at the macro level. But, it had positive effects on trade and investment in developing countries. The study recommended that regional integration was essential for the growth of member countries as it led to increased trade and investment.

Gul, Najia and Yasin, Hafiz M. (2011) in order to examine the trade potential of Pakistan, made use of the technique of the gravity model of trade. The data for the period 1981-2005 were obtained from the trade statistics of the IMF and World Bank. The trade potential of the country, both worldwide and within a specific region, was estimated using the coefficients obtained in the analysis. The study found that the volume of trade between Pakistan and other SAARC countries and the Economic Cooperation Organization (ECO) was very low. This was in spite of the fact that there existed tremendous trade potential. They cited political and social tensions as the main obstacles among South Asian nations, particularly between India and Pakistan, the two major countries of the SAARC region.

Hassan, M. Kabir (2001) attempted to analyse the viability of economic cooperation arising out of the potential for free trade among the SAARC countries. Data were collected from IMF's DOTS, UNCTAD, and UN COMTRADE for the period 1991-97, and the gravity model was used. The study found that Intra-SAARC trade was low and the countries of the SAARC region traded less with other countries of the world. This study suggested that in order to achieve trade creating benefits the SAARC countries must trade more among themselves and also with the outside world.

Hiranath, S. W. (2004) evaluated the working of the SAARC regional group under SAPTA and also examined the future possibilities of SAFTA. The Panel and Cross-sectional data on bilateral trade flow, GDP, and Per Capita GDP was taken from DOTS (IMF) and the publications of the World Bank for the period 1996-2002. The Gravity Equation was estimated by using the "Generalized Least Squares (GLS) regression technique" and it was corrected for the problems of Heteroscedasticity and Autocorrelation. The study found that there was a significant trade creation effect under SAFTA. However, the study found no evidence of trade diversion effect with the other remaining countries of the world.

Batra, Amita (2006) analysed the world trade flow of 146 countries in her study. She used an "Augmented Gravity Model" equation for her analysis. The main objective of the study was to estimate trade potential for India. The data on population and GNP was taken from WDI (World Bank, CD-ROM, 2003). The study used the OLS estimation technique for the purposes of estimation of the model using cross-sectional data for the year 2000. The findings revealed that the size of India's trade potential was highest in the "Asia-Pacific Region". This was followed by "Western Europe and North America".

Rahman, Shadat, and Das (2006) investigated the effects of the "trade creation and trade diversion on Regional Trade Agreements (RTAs)", with a particular focus on SAFTA. For the purposes of their analysis, they used the gravity model. 'The panel data approach with country-pair specific fixed effect' was used in the study. In addition to that, the regression model included the year-specific fixed effect. They used the bilateral export flow as a dependent variable. The data on bilateral flows of trade and other variables were taken from DOTS, IMF database, World Development Indicator (WDI), IFS CD-ROM for the sample of 61 countries with the time period 1991-2003. All coefficients of the gravity variables were found to carry an expected sign and were also significant statistically. The study concluded that the reduction in "tariff and non-tariff barriers" and the introduction of Rules of Origin (RoO) would increase intra-regional trade in the SAARC countries.

Ekanayake, Mukhergee, and Veeramacheneni (2010) made an attempt to study the effects of trade creation and trade diversion on RTAs in Asian countries and also their effects on intra-regional trade flows. The study made use of an AGM model for the purposes of analysis. The data for the study were taken from various publications of the UN, IMF, and World Bank for the period 1980-2009. A total number of nineteen Asian nations were selected. The results of the model showed that most of the coefficients pertaining to regional dummy variables were positive. They were also found to be significant statistically. This implied that the effect of the multilateral trade agreement on trade was more than BTAs. The study suggested that the fast evolving economic and political environment provided vast possibilities for analysing the success of economic integration in the Asian region.

Akhter and Ghani Ejaz (2010) analysed the trade agreement SAFTA and examined its role in increasing the trade potential of the members of the SAARC region and bringing in trade benefits for the countries. The study for the purposes of analysis uses the gravity model to estimate bilateral trade flows and trade potential among SAARC nations. The data on "GDP and per capita GDP were taken from the publications of World Bank, UNCTAD and WTO" for the time period 2003-2008. The findings of the study showed that the potential for trade creation existed provided there was a regional trade agreement among India, Pakistan, and Sri Lanka. However, as far as the potential for trade creation is concerned, it would be little if the SAARC members signed FTAs with other non-member countries. The basis of these findings was the data pertaining to SAPTA. The study found that SAFTA was more useful in the long run than in the short run. The trade diversion effects under SAFTA would be minimized if trade industrialization, as well as trade liberalization, continued in the region.

Therefore, the study suggested a conducive economic and social setup and also a strong political will for "economic integration and trade liberalization in the region".

Rizwanulhassan & Shafiqurrehman (2015) attempted to evaluate the extent of intra-regional trade among SAARC nations by using the "Extended Gravity Model" for the time period 1991-2010. The data on trade were taken from DOTS (IMF), on population and real GDP from WDI (World Bank), and on distance from the time and date website (timeanddate.com). The study found a significant effect of "GDP, GDP per capita, Exchange Rate Volatility, and Common Border on intra-regional trade".

Abhyaratne, Anoma, and Varma, Sumati (2017) examined the effect of the "India-Sri Lanka Free Trade Agreement (ISLFTA)" on their bilateral trade flows. The study collected panel data for the period 1990-2014 to estimate the "gravity model using the Weighted Least Squares (WLS) Method". It was found that ISLFTA had increased their bilateral trade and produced ample trade creation effects.

2.1. Objective and methodology

The objective of this study is to explore trade possibilities under SAFTA between India and other SAARC member nations by making use of the augmented gravity model. To analyze bilateral trade, the gravity approach is the most widely used empirical technique. Determination of factors affecting bilateral trade and their influence on economic growth is a highly debatable issue among the researchers. Trade potential is generally computed with the help of the gravity model in empirical research. The flow and direction of potential trade in the literature are determined by subtracting the estimated trade flows (predicted by the gravity model) and actual trade flows. Thus, the coefficients obtained from the gravity model are used to forecast trade potential for India. In the case of India, the untapped trade potential is shown by actual trade with any other member country and is less than what is predicted by the gravity model.

In the present study, the augmented gravity model is utilized to find the trade potential of India under SAFTA with other SAARC member countries from 1992 to 2019 at an annual frequency in general and from 2004 to 2019 in particular.

3. Methodological construct

3.1. The model

The gravity equation is an applied model of trade. It analyses bilateral trade among nations/states. It is similar to Newton's physics function, which describes the "force of gravity". "The model explains the flow of trade between a pair of countries as being proportional to their economic ''mass'' (national income) and inversely proportional to the distance between them" (Maryam and Kashim, 2015). Tinbergen (1962) and Poyhonen (1963) used the following equation:

$$T_{ij} = (GDP_i * GDP_j)/Distance_{ij} \quad (1)$$

In this equation, T_{ij} represents the trade between two participating country i and j, GDP_i and GDP_j are country i and j's national incomes respectively. $Distance_{ij}$ measures the bilateral distance between the two countries and is taken to be a constant of proportionality. Taking logarithms of the gravity equation in (1) as given above, the model becomes linear and the estimable equation becomes:

$$\text{Log } T_{ij} = \beta_0 + \beta_1 \log (GDP_i . GDP_j) + \beta_2 \log (Distance_{ij}) + U_{ij} \quad (2)$$

Where β_0, β_1, and β_2 are coefficients to be estimated. Equation (2) is the main equation where trade is found to be a positive function of income and an inverse function of distance.

The basic logic of this model is that trade has a positive relationship with the size of the trading country and a negative relation with the distance between countries. Here, the distance is taken as a proxy for information and transportation costs. As the distance increases trade decreases and vice versa. Tinbergen (1962) used the following modified equation:

$$Tij = \beta_0 \, Yi^{\beta 1} \, Yj^{\beta 2} Dij^{\beta 3} Nij^{\beta 4} \, P_c^{\beta 5} P_b^{\beta 6} \quad (3)$$

Where, Nij is the border dummy for country i and j, Y_i stands for GDP of country i while Y_j depicts the GDP of country j, D_{ij} is the distance between country i and country j, Pc shows the commonwealth preference dummy variable.

3.2. Econometric technique used in this paper

The study makes use of the augmented gravity model. According to this model, the total trade between countries depends on GDP or population, distance (a proxy of transportation costs), and other dummies that may promote or restrict the trade between them. Apart from the basic gravity variables, there are some other factors that affect bilateral trade such as cultural similarities, trade agreements, geographical location, factor endowment, and the role of development. Such factors are used to find out the trade potential of India. Real exchange is another important determinant of the trade potential. However, there were many missing values for this variable for several countries. Therefore, we are unable to include it in our model. The present study makes use of the Tinbergen (1962) following equation:

$$T_{ij} = \beta_0 \, X_{1i}^{\beta 1} \, X_{2j}^{\beta 2} \, X_{3ij}^{\beta 3} A_{ij}^{\beta 4} \, D_{ij}^{\beta 5} \quad (4)$$

Where, T_{ij} represents country i and country j's total trade, X_{1i} and X_{2j} show the GDPs of country i and country j respectively, X_{3ij} shows the distance between the two countries, Aij is the set of other explanatory control variables, D_{ij} shows the set of all

dummy variables for the two countries, and β_0 is the vector of the coefficient of all explanatory variables.

In order to make the model linear, the logarithm of the equation (4) is taken. Thus, the model in the log-linear form becomes:

$$\log T_{ijt} = \beta_0 + \beta_1 \log X_{1it} + \beta_2 \log X_{2jt} + \beta_3 \log X_{3ijt} + \beta_4 \log A_{ijt} + \beta_5 D_{ijt} + U_{ijt} \qquad (5)$$

In the above equation, the variable A_{ijt} represents explanatory variables such as per capita GDP differential (PCGDPD), and total trade to GDP (T/Y_i, T/Y_j) ratio. D_{ijt} is used for dummies like common border, language, etc.

Per Capita GDP differential (PCGDPD) is used to test the Heckscher-Ohlin (H-O) or Linder's hypothesis. The total trade to GDP ratio is taken for showing trade openness. Following the Wang and Winter (1991) study, the present study includes the border and language variables to estimate the effects of cultural factors between the countries.

The study used three gravity models for India's bilateral trade with 7 trading countries from 1992 to 2014: "(i) the gravity model of total trade (export + import), (ii) the gravity model of exports, and (iii) the gravity model of India's imports".

Thus, the gravity model for total trade becomes:

$$\log(T_{ijt}) = \beta_0 + \beta_1 \log(GDP_{it}) + \beta_2 \log(GDP_{jt}) + \beta_3 log(PCGDPD_{ijt}) + \beta_4 \log(Dis_{ijt}) + \beta_5 log(T/Y_{it}) + \beta_6 log(T/Y_{jt}) + \beta_7 (Borderij) + \beta_8 (Lan_{ij}) + Uijt \qquad (6)$$

$$\log(\mathrm{EXP}_{ijt}) = \beta_0 + \beta_1 \log(GDP_{it}) + \beta_2 \log(GDP_{jt}) + \beta_3 log(PCGDPD_{ijt}) + \beta_4 \log(Dis_{ijt}) + \beta_5 log(T/Y_{it}) + \beta_6 log(\mathrm{IMP}/Y_{jt}) + \beta_7 (Borderij) + \beta_8 (Lan_{ij}) + Uijt \qquad (7)$$

$$\log(\mathrm{IMP}_{ijt}) = \beta_0 + \beta_1 \log(GDP_{it}) + \beta_2 \log(GDP_{jt}) + \beta_3 log(PCGDPD_{ijt}) + \beta_4 \log(Dis_{ijt}) + \beta_5 log(T/Y_{it}) + \beta_6 log(\mathrm{EXP}/Y_{jt}) + \beta_7 (Borderij) + \beta_8 (Lan_{ij}) + Uijt \qquad (8)$$

Where, T_{ij} represents country i and country j's total trade, EXP_{ij} represents the total exports between two countries, IMP_{ij} is the total imports between the two countries, GDP_i is the income of country i, GDP_j shows the income of country j, $PCGDPD_{ij}$ is per capita gross domestic product differential between countries, $Dist_{ij}$ is the distance between the two countries, T/Y_i is the total trade to GDP ratio $_j$, EXP/Y_j, IMP/Y_j is the total trade/export/import to GDP ratio of country j respectively.

3.3. Panel data framework

There are several merits of using panel data methods which are not there in the case of time series and cross-sectional data. Panel data control due to individual heterogeneity and produces efficient estimates by dealing with multicollinearity in explanatory variables. The most common and widely used panel data estimation

models are the random effect model (REM) and the fixed effect model (RFM) (Gujrati, 2008). REM assumes that the intercept of each cross-section is a random variable (Gujrati, 2008). This model is more used when random intercepts are uncorrelated with explanatory variables, whereas, in the Fixed Effect Model (FEM), both individual and time effects are accounted for. In the FE model, the slope coefficients do not change. The model is useful when individual intercepts are correlated with the independent variables (Gujrati, 2008).

The present study proposes to calculate the effects of both time-invariant and time-variant variables in the factors affecting bilateral trade and India's trade potential. Therefore, Random Effect Model is preferred to the Fixed Effect Model. "REM is also preferred as numbers of cross-sections are greater than time period" (Gujrati, 2008).

3.4. Trade potentials

The calculation of trade potential is also related to the gravity model. Several studies have used various methods to find out trade potentials. The most widely used one is to use point estimates of coefficients on explanatory variables to estimate the trade potential. The study has calculated the trade potentials using the following equation:

$$\text{Trade Potential} = \text{Predicted trade flows} - \text{Actual trade flows} \tag{9}$$

Predicted values are calculated from gravity models of total trade, total exports, and total imports. The positive value means that there is a chance of increasing trade expansion and the negative value means that trade potential with selected trading countries has been already exhausted (Batra, 2004).

3.5. Results

3.5.1. Total trade determinants

The present study has used the REM to estimate the augmented gravity approach equation. The REM was selected on the basis of the Hausman Test, which resulted in p values greater than 0.05 in all the cases. We used the STATA statistical package to conduct this test. The logged dependent and independent variables mean that the estimated coefficients of the independent variables show the elasticity of variables. They show the marginal effect of the predictor variable while keeping the other variables unchanged. The present study has estimated seven models using REM so as to find the determinants of trade in India. The variables, which are considered to be significant, have been included one by one in the model. The results of the augmented gravity model are presented in Table 3.1.

Model 1 shows the standard gravity variables such as GDP and distance as used by Tinbergen (1962). The results of the study reveal that in model 1, the GDP of India and partner countries, i.e. economic size, and distance are significant with expected signs.

We include significant variables one by one to estimate the effect on other variables. We begin with lnPCGDPDij variables, and model 2 takes into account the $\ln PCGDPD_{ij}$ variable while keeping all other variables used in model 1. There is a slight change in the estimated coefficients of model 2 as compared to model 1. In model 3, $Dist_{ij}$ (distance) is included and is found significant at the level of 5%. Model 4 includes $\ln T/Y_i$ and $\ln T/Y_j$ (log trade GDP ratio of country i and j respectively) and excluded $\ln PCGDPD_{ij}$ and $Dist_{ij}$ (distance). LnT/Yi is significant at the level of 1%. Model 5 includes border and language dummy variables and $\ln PCGDPD_{ij}$ which make $\ln GDP_i$ again significant at the level of 1%. Model 6 includes $Dist_{ij}$ (distance), which makes $\ln PCGDPD_{ij}$ and language dummy variables significant at the level of 1%. $Dist_{ij}$ (distance) itself is significant at the level of 1%. Model 7 is the final model that has the GDP of India and other partner countries, PCGDP differential which validates the H-O theory. This implies that countries which have a different factor of endowments generally, trade more, other variables of the model include trade openness of the two countries, distance, and dummy variables. Most of these variables turn out to be significant and show expected signs. The explanatory power of all the estimated models, as measured by R^2, has been found to be approximately the same.

The final augmented gravity model used for ascertaining the determinants of India's total trade is as follows:

$$\log(T_{ijt}) = \beta_0 + \beta_1 \log(GDP_{it}) + \beta_2 \log(GDP_{jt}) + \beta_3 log(PCGDPD_{ijt}) + \beta_4 \log(Dis_{ijt}) + \beta_5 log(T/Y_{it}) + \beta_6 log(T/Y_{jt}) + \beta_7 (Border_{ij}) + \beta_8 (Lan_{ij}) + Uijt$$

$$\log(TSP_{ijt}) = \beta_0 + \beta_1 \log(GDP_{it}) + \beta_2 \log(GDP_{jt}) + \beta_3 log(PCGDPD_{ijt}) + \beta_4 \log(Dis_{ijt}) + \beta_5 log(T/Y_{it}) + \beta_6 log(T/Y_{jt}) + \beta_7 (Border_{ij}) + \beta_8 (Lan_{ij}) + Uijt$$

$$\log(TSF_{ijt}) = \beta_0 + \beta_1 \log(GDP_{it}) + \beta_2 \log(GDP_{jt}) + \beta_3 log(PCGDPD_{ijt}) + \beta_4 \log(Dis_{ijt}) + \beta_5 log(T/Y_{it}) + \beta_6 log(T/Y_{jt}) + \beta_7 (Border_{ij}) + \beta_8 (Lan_{ij}) + Uijt$$

The results of the estimated results for the gravity model of total trade are given below:

$$\log(T_{ijt}) = -28.48 + 0.286 \log(GDP_{it}) + 0.602 \log(GDP_{jt}) + 0.841 log(PCGDPD_{ijt}) - 0.871 \log(Dis_{ijt}) + 1.362 log(T/Y_{it}) - 0.217 log(T/Y_{jt}) + 0.159 (Border_{ij}) + 1.305 (Lan_{ij})$$

$$\log(TSP_{ijt}) = -51.11 + 0.964 \log(GDP_{it}) + 0.648 \log(GDP_{jt}) + 1.246 log(PCGDPD_{ijt}) - 0.504 \log(Dis_{ijt}) + 1.015 log(T/Y_{it}) - 0.0457 log(T/Y_{jt}) + 0.262 (Border_{ij}) + 1.779 (Lan_{ij})$$

$$\log(TSF_{ijt}) = -28.60 + 0.311 \log(GDP_{it}) + 0.788 \log(GDP_{jt}) + 0.607 log(PCGDPD_{ijt}) - 1.328 \log(Dis_{ijt}) + 0.215 log(T/Y_{it}) + 0.567 log(T/Y_{jt}) + 0.388 (Border_{ij}) + 0.782 (Lan_{ij})$$

The GDP of trading countries has an expected sign, which is a conventional and important factor of AGM. It has an expected coefficient and is statistically significant. The result provides evidence in favour of a positive relationship between the size of countries

and trade. The coefficient of per capita GDP differential is statistically significant and positive. The observed sign of the coefficient reveals that the H-O theory has a dominating impact on Linder's hypothesis. This implies that those nations usually trade more which have different factors of the endowments. Distance is also statistically significant and shows an expected negative sign. Distance has been used as a proxy for transportation costs and other time-related costs. The trade to GDP ratio shows an expected positive sign, which is significant. The trade to GDP ratio for country j shows a negative sign. The results reveal that trade openness leads to increased trade volume among them.

In order to capture the cultural effects on trade flows, the study took into account some dummy variables, namely Border and Language. The dummy variable Border takes the value 1 for Afghanistan, Bangladesh, Bhutan Nepal, and Pakistan as they share common borders with India. However, the results contradict the theoretical reasoning of this variable. Such results occur because due to military and political tensions a large volume of trade takes place between India and Pakistan. Another dummy variable, which has been taken into account to study the effects of culture on trade, is Language. This variable has the expected positive sign and is also found to be statistically significant. It is usually believed that countries sharing a common language have more trade.

3.5.2. Total trade potential of India

An important aspect of the gravity model is to estimate trade potentials. The study has calculated the total trade potentials of India with SAARC and bilateral total trade potential with 6 other SAARC member countries for the Pre-SAFTA period (1992-2003) and Under-SAFTA period (2004-2014).

In Table 3.1 we show the mean of total trade potentials by subtracting predicted trade value (P) from actual trade flows (A), i.e. value of P-A. A

Table 3.1. Total Trade Potential of India (Average)

Indicator Countries	Pre-SAFTA		Under-SAFTA	
	(P-A) 1992-1997	(P-A) 1998-2003	(P-A) 2004-2009	(P-A) 2010-2014
SAARC	-184.01	-370.21	1404.22	1567.26
Afghanistan				
Bangladesh	-238.32	-260.45	1210.03	1424.73
Bhutan	5.23	-1.05	-42.39	-60.09
Maldives	-6.54	-1.92	-5.09	34.56
Nepal	27.24	-126.13	-80.15	-751.60
Pakistan	20.43	62.89	-105.91	388.66
Sri Lanka	7.95	-43.56	427.72	531.01

As shown in Table 3.1, the study has estimated an average of 6 years in the Pre-SAFTA period, while the last average (2010-2014) is of 5 years under SAFTA. The average trade potential of India was highest for Nepal followed by Pakistan, Sri Lanka, and Bhutan in that order under the Pre-SAFTA period during 1992-1997. This shows that India had maximum trade potential with these countries, whereas for those countries which show negative value India exceeded its total trade potential, i.e. SAARC as a whole, Bangladesh and Maldives. For the period 1998 to 2003, under Pre-SAFTA, the average value of total trade potential exceeded for all except for Pakistan. So, during this period India had the highest total trade potential with Pakistan.

During the recent years, i.e. 2010 to 2014, under SAFTA, India had the highest trade potential with Bangladesh and Sri Lanka. This is clear from Table 3.1, which shows the highest average P-A value between these two countries. Also in this period, India exceeded its trade potential with Nepal and Bhutan.

3.6. Conclusion and summary

This study provides a more detailed analysis of trade potential using the augmented gravity model and commodity-wise trade possibilities using the Potential Trade Approach among the SAARC countries. This research work has computed trade possibilities and potential under SAFTA of India with other SAARC participating nations using the gravity model. The gravity model has been widely used in research for estimating the trade potential. In our case, the coefficients obtained from the model are then used to forecast trade potential for India. The gravity model shows that in the case of India actual trade with any country is less than predicted and there is untapped trade potential. The analysis is based on the panel data. Panel data estimation has many advantages and is preferred over cross-sectional and time-series data because it controls individual heterogeneity. Panel data technique enhances the efficiency of the regression estimates by decreasing collinearity among explanatory variables with ample degrees of freedom.

The results presented in this study obtained from the analysis of the gravity model show that the GDP of trading partners (a proxy for economic size) has a positive coefficient and it is also found to be significant statistically. This result lends support to the positive relationship between the economic size of trading nations and the trade flows. The estimated coefficient of per capita GDP differential (country i and country j) shows a positive sign and is also found to be statistically significant at a 1% level of significance. The positive sign of the estimated coefficient verifies the H-O hypothesis and Linder's hypothesis. This implies that the countries that generally trade more have different factor endowments. Distance is another important variable. This variable shows the expected negative sign and is also found to be statistically significant at a 1% level of significance. The variable is used as a proxy for transportation cost and other

time-related costs. Theoretically, a negative relationship exists between trade flows and the distance between trading partners. This implies that as the distance between the partners' increases, the trade flows decrease. Thus, there is a theoretical justification for including the distance variable in the hypothesis of the gravity model. Another variable, i.e. Trade to GDP ratio of country, which is the proxy for trade openness, shows an expected positive sign and is also statistically significant at a 1 % level of significance. And Trade to GDP ratio for country j shows a negative sign. The results support the theoretical reasoning of the variable, which states that the more the open economy of trading partners, the more will be the trade between them.

Apart from the above variables, the present study has also made use of some dummy variables in the analysis using the gravity model. The dummy variables were included to assess the impact of cultural effects on intra-regional trade flows. For example, the Border is a dummy, which takes the value 1 for countries like Afghanistan, Bangladesh, Bhutan, Nepal, and Pakistan as they share a common border with India. And the value zero when they do not share a common border with India. However, it is assumed that the countries sharing common borders generally share common traditions, customs, and consumption patterns also. Consequently, a positive association between a common border and trade flows is expected. But the results are contradictory and are not in harmony with the theoretical reasoning of this variable. For example, in the case of India and Pakistan, this may be because the low volume of trade between the two countries is due to political factors and strained relations. Another dummy variable used in the model for assessing cultural effects is Language. This variable shows the desired positive sign and is statistically significant. Generally, it is assumed that countries sharing a common language have more trade.

The findings of this study show that intra-regional trade volumes between SAARC nations can be increased and encouraged. It is important to undertake structural reforms so that the trade with non-member countries can also be boosted. The researchers should try to take into account the effect of locational and infrastructural advantages on transportation costs using gravity. Previous research has also argued that an augmented gravity model may help in explaining some key features of South Asian trade, which may not be explained by traditional gravity models.

3.7. Future area of research

The researchers should try to take into account the effect of locational and infrastructural advantages on transportation costs using gravity. Services ought to likewise be included and accentuation ought to incorporate whatever number of services as could be expected under the circumstances.

References

Balassa, Bela, (1962). The Theory of Economic Integration, George Allen & Unwin, London.

Bandara, J., W. Yu, (2001). How Desirable is the South Asian Free Trade Area? A Quantitive Assessment, Paper presented to the Fourth Annual Conference on Global Economic Ananlysis, Purdue University, Indiana, USA, June 27-29.

Batra, Amita, (2007). South Asia's Free Trade Agreement: Strategies and Options, Economic & Political weekly, Vol.42, No.38, pp. 3878–3885.

Bhagwati, Jagdish, (1995). U.S. Trade Policy: The Infatuation with Free Trade Agreements, Discussion Paper Series No. 726.

Bhattacharyya, B., V. Katti, (2002). Regional Trade Enhancement: "SAPTA" and Beyond, http://www.iift.edu/publications/paper3.pdf.

Chowdhury, M. B., (2005). Trade Reforms and Economic Integration in South Asia. Applied Econometrics and International Development, Vol. 5(4), pp. 23–40.

Chowdhury, M., (1998). Resources Boom and Macroeconomic Adjustment: Theory and Evidence from Papua New Guinea, PhD Dissertation, Research School of Pacific and Asian Studies, Australian National University, Canberra, Australia.

Chowdhury, M., (2004). Resources Boom and Macroeconomic Adjustments in Developing Countries, Ashgate Publishing Limited, Gower House, England.

Department of Commerce under the Ministry of Finance, Government of India.

Din, M. U., Nasir S., (2004). Regional Economic Integration in South Asia: The Way Forward. The Pakistan Development Review, Vol. 43(4), pp. 959–974.

Ekanayake, E. M., Mukherjee, A. and Veeramacheneni B., (2010). Trade Blocks and Gravity Model: A Study of Economic Integration among Asian Developing Countries. Journal of Economic Integration, Vol. 25(4), pp. 627–643.

Hoekman, B., et al. (eds.), (2002). Development, Trade, and the WTO: A Handbook, The World Bank: Washington, D.C.

IMF, (2001). Balance of Payments Statistics Yearbook 2001, Vol. 2. IMF, Direction of Trade Statistics Yearbooks, various issues.

Kohli, E., (2004). Human Resource Development in SAARC Countries. Unpublished dissertation submitted to Guru Nanak Dev University, Amritsar for the degree of MSc in Economics at the Punjab School of Economics.

Pannu, N. K., (2005). Development Pattern of SAARC and Central Asian Countries. Unpublished dissertation submitted to Guru Nanak Dev University, Amritsar for the degree of MSc in Economics at the Punjab School of Economics.

Pitigala, N., (2005). What Does Regional Trade in South Asia Reveal about Future Trade Integration? Some Empirical Evidence, World Bank Policy Research Working Paper 3497, February 2005.

Rahman, M., Shadat, W. B. and Das, N. C., (2006). Trade Potential in "SAFTA": An Application of Augmented Gravity Model. Occasional Paper, the Centre for Policy Dialogue, Dhaka, Bangladesh, Series 61.

S Perera, M. S., (2009). The South Asian Free Trade Area: an Analysis of Policy Options for Sri Lanka. Journal of Economic Integration, Vol. 24(3), pp. 530–562.

S Perera, M. S., (2009): "The South Asian Free Trade Area: an analysis of Policy Options for Sri Lanka", Journal of Economic Integration, Vol. 24, No. 3, pp. 530–562.

SAARC Website: data collected from http://saarc-sec.org/areaofcooperation/detail.php?activity_id=5

SAARC's Official website.

Sejuti, Jha, (2013). Utility of Regional Trade Agreements? Experience From India's Trade Regionalism, Foreign Trade Review, Vol. 48, No. 2, pp. 233–245.

Sharma, N., Kumar, P., (2012). Growth and Pattern of India's Intra-industry Trade with SAARC Nations. The Indian Economic Journal, Vol. 60(2), pp. 114–125.

Sharma, R., (2002). India's Trade Relations with SAARC Countries in the context of Globalisation and Liberalisation. Unpublished PhD Thesis submitted at South Asian Studies Centre, University of Rajasthan, Jaipur.

Sharma, Vipin, (2019). India and Bangladesh Trade Potential under "SAFTA", International Journal of Trade & Global Perspectives; Vol. 8(1), pp. 4109–4134.

Taneja, N., Prakash, S. and Kalita, P., (2013). India's Role in Facilitating Trade under "SAFTA". Indian Council for research on International Economic Relations (ICRIER), New Delhi, Working Paper, No. 26, pp. 31–19.

The Kathmandu Post, (2002). Nepal Gets up to 30 per cent Customs Tariff Concession, Kathmandu, November 27.

Udagedera, S., (2001). "SAPTA" Negotiations: Constraints and Challenges, in Impediments to Regional Economic Cooperation in South Asia, edited by Saman Kelegama, Institutie of Policy Studies of Sri Lanka and Coalition for Action on

South Asian Cooperation in association with Friederich-Ebert-Stiftung, Colombo Office.

Velde, D. W. T., (2011). Regional Integration, growth and Convergence, Journal of Economic Integration, Vol. 26(1), pp. 1–28.

Vipin and Vinod, (2016). Economic Integration among SAARC Countries: From "SAPTA" to "SAFTA", Inter-State Conflicts and Contentious Issues in South Asia: Challenges and Prospects For SAARC, Kalpaz publications.

Vipin and Vinod, (2016). Trade Reforms and Economic Integration in SAARC Countries: "SAPTA" to "SAFTA", International Journal of Entrepreneurship & Business Environment perspectives, Vol. 5, No. 4, pp. 2984–2996.

Weerakoon, Dushni, (2001): "Does "SAFTA" Have a Future", Economic & Political Weekly, Vol. 36, No. 34, pp. 3214–3216.

Weerakoon, Dushni and Tayanthi Thennakoon, (2006). SAFTA" Myth of Free Trade", Economic & Political Weekly, Vol. 41, No. 37, pp. 3920–3923.

Wooster, R. B., Banda, T. M. & Dube, S., (2008). The Contribution of Intra-regional and Extra-regional Trade to Growth: Evidence from the European Union. Journal of economic Integration, Vol. 23(1), pp. 161–182.

Zaheer, Dr. R., (2013). The Economic Performance of SAARC Member Countries. Research on Humanities and Social Sciences, Vol. 3(5), pp. 201–214.

APPENDIX

Table A.1. Panel Data Results of Gravity Trade Model (Random Effects Model)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Lntrade	Lntrade	Lntrade	lntrade	Lntrade	lntrade	lntrade
Lngdpi	1.098***	1.051***	1.005***	0.369	1.011***	0.991***	0.286
	(0.146)	(0.152)	(0.120)	(0.240)	(0.126)	(0.0565)	(0.203)
Lngdpj	0.598***	0.635***	0.697***	0.578***	0.667***	0.649***	0.602***
	(0.146)	(0.146)	(0.116)	(0.157)	(0.118)	(0.0303)	(0.0523)
Lnpcgdpdij		0.176			0.321	0.854***	0.841***
		(0.221)			(0.228)	(0.112)	(0.107)
Lndistij			-1.172**			-0.886***	-0.871***
			(0.506)			(0.0991)	(0.0951)
Lntyi				1.348***			1.362***
				(0.327)			(0.345)
Lntyj				-0.0587			-0.217
				(0.192)			(0.201)
Border					0.566	0.286	0.159
					(1.048)	(0.196)	(0.220)
Language					0.929	1.330***	1.305***
					(1.005)	(0.149)	(0.143)
Constant	-37.85***	-37.35***	-36.98***	-21.98***	-44.02***	-45.15***	-28.48***
	(1.439)	(1.558)	(1.436)	(4.106)	(7.813)	(1.823)	(4.465)
Observations	138	138	138	138	138	138	138
Number of countrypair1	6	6	6	6	6	6	6

Note: Standard errors are in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.2. Gravity Model of Total Trade under Pre-SAFTA

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	model1	model2	model3	model4	model5	model6	model7
Lngdpi	1.621***	1.649***	1.524***	1.311*	1.678***	1.628***	0.964
	(0.269)	(0.270)	(0.258)	(0.780)	(0.227)	(0.215)	(0.799)
Lngdpj	0.595***	0.586***	0.689***	0.539**	0.595***	0.641***	0.648***
	(0.169)	(0.170)	(0.152)	(0.211)	(0.0705)	(0.0409)	(0.110)
Lnpcgdpdij		0.392			1.274***	1.233***	1.246***
		(0.350)			(0.267)	(0.178)	(0.185)
Lndistij			-0.982			-0.512***	-0.504***
			(0.643)			(0.150)	(0.153)
Lntyi				0.569			1.015
				(1.114)			(1.159)
Lntyj				-0.165			0.0457
				(0.464)			(0.526)
Border					0.0641	0.226	0.262
					(0.586)	(0.267)	(0.584)
Language					1.585***	1.760***	1.779***
					(0.538)	(0.207)	(0.301)
Constant	-51.78***	-52.15***	-50.79***	-43.37***	-64.29***	-65.12***	-51.11***
	(5.513)	(5.509)	(5.538)	(16.54)	(7.049)	(5.821)	(17.29)
Observations	72	72	72	72	72	72	72
Number of country1	6	6	6	6	6	6	6

Note: Standard errors are in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A.3 Gravity Model of Total Trade under SAFTA

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Lngdpi	0.689***	0.640***	0.500***	0.591***	0.670***	0.468***	0.311**
	(0.158)	(0.177)	(0.106)	(0.185)	(0.255)	(0.111)	(0.121)
Lngdpj	0.494***	0.542***	0.675***	0.513***	0.513**	0.708***	0.788***
	(0.142)	(0.162)	(0.0864)	(0.158)	(0.240)	(0.0916)	(0.0241)
Lnpcgdpdij		0.168			0.131	0.307*	0.607***
		(0.205)			(0.287)	(0.172)	(0.0839)
Lndistij			-1.330***			-1.403***	-1.328***
			(0.322)			(0.322)	(0.0565)
Lntyi				0.277			0.215
				(0.306)			(0.367)
Lntyj				0.142			0.567***
				(0.162)			(0.132)
Language					0.0957	0.911	0.782***
					(2.152)	(0.714)	(0.120)
Border					0.131	0.626	0.388***
					(2.207)	(0.756)	(0.148)
Constant	-24.03***	-23.77***	-22.27***	-23.38***	-24.71	-29.05***	-28.60***
	(1.779)	(1.829)	(1.689)	(2.110)	(16.53)	(5.626)	(2.457)
Observations	77	77	77	77	77	77	77
Number of country1	7	7	7	7	7	7	7

Note: Standard errors are in parentheses
*** p<0.01, ** p<0.05, * p<0.1

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 99–121, DOI 10.21307/stattrans-2021-029
Received – 24.01.2020; accepted – 01.06.2021

A new reciprocal Rayleigh extension: properties, copulas, different methods of estimation and a modified right-censored test for validation

Haitham M. Yousof[1], M. Masoom Ali[2], Hafida Goual[3], Mohamed Ibrahim[4]

ABSTRACT

In this article, a new reciprocal Rayleigh extension called the Xgamma reciprocal Rayleigh model is defined and studied. The relevant statistical properties are derived, and the useful results related to the convexity and concavity are addressed. We discussed the estimation of the parameters using different estimation methods such as the maximum likelihood estimation method, the ordinary least squares estimation method, the weighted least squares estimation method, the Cramer-Von-Mises estimation method, and the bootstrapping method. A simulation study was conducted to assess the performances of the proposed estimation methods are investigated through a simulation study. Many bivariate and multivariate type model have also been derived based on Farlie-Gumbel-Morgenstern copula, the Clayton copula, Renyi's entropy copula and the Ali-Mikhail-Haq copula. A modified Nikulin-Rao-Robson test for right-censored validation is applied to a censored real data set.

Key words: Xgamma model, reciprocal Rayleigh model, simulations, bootstrapping, Farlie Gumbel Morgenstern copula, least squares, Cramer-Von-Mises, bootstrapping, Ali-Mikhail-Haq copula, convexity, concavity.

1. Introduction

The probability density function (PDF) and cumulative distribution function (CDF) of the reciprocal Rayleigh (RR) distribution are given, respectively, by

$$g_{\theta_2}(y) = 2\theta_2^2 y^{-3} e^{-\left(\frac{\theta_2}{y}\right)^2}\Big|_{y\in\mathbb{R}^+},$$

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

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

[3] Laboratory of Probability and Statistics, University of Badji Mokhtar, Annaba, Algeria. E-mail: goual.hafida@gmail.com. ORCID: https://orcid.org/0000-0003-4932-4332.

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

and

$$G_{\theta_2}(y) = e^{-\left(\frac{\theta_2}{y}\right)^2}|_{y\in\mathbb{R}^+},$$

where $\theta_2 > 0$ refers to the scale parameter. The RR model is a special case from the well-known inverse Weibull distribution. The RR model was originally proposed by Fréchet (1927). It has many applications in accelerated life testing, earthquakes, floods, wind speed, horse racing, rainfall, queues in supermarkets, and sea waves. Gusmao et al. (2011) defined and studied the generalized reciprocal Rayleigh (GRR) distribution. Krishna et al. (2013) proposed some applications of the Marshall-Olkin reciprocal Rayleigh (MORR) distribution. Mahmoud and Mandouh (2013) proposed and studied the transmuted reciprocal Rayleigh (TRR) distribution. Haq et al. (2017) presented a new four-prarameter reciprocal Rayleigh version for modeling extreme values. Korkmaz et al. (2017) studied some theoretical and computational aspects of the odd Lindley reciprocal Rayleigh (OLRR) distribution. Yousof et al. (2018d) defined a new family called the odd reciprocal Rayleigh G (ORR-G) family of distributions. Yousof et al. (2019) defined a new compound version of the reciprocal Rayleigh (OBRR) distribution. Salah et al. (2020) defined and studied a new version of RR model called the odd Burr RR model with different copula, different estimation methods, applications and validation testing. Recently, Cordeiro (2020) proposed and studied the Xgamma-G (Xg-G) family of distribution with CDF and PDF (for $\theta_1 > 0$) given by

$$F_{\theta_1,\underline{\xi}}(y) = 1 - \frac{1+\theta_1-\theta_1\log\left[1-G_{\underline{\xi}}(y)\right]+\frac{1}{2}\theta_1^2\left\{\log\left[1-G_{\underline{\xi}}(y)\right]\right\}^2}{1+\theta}\left[1-G_{\underline{\xi}}(y)\right]^{\theta_1}|_{y\in\mathbb{R}}, \tag{1}$$

and

$$f_{\theta_1,\underline{\xi}}(y) = \frac{\theta_1}{1+\theta_1}g_{\underline{\xi}}(y)\left[1-G_{\underline{\xi}}(y)\right]^{\theta_1-1}\left(\theta+\frac{1}{2}\theta_1^2\left\{\log\left[1-G_{\underline{\xi}}(y)\right]\right\}^2\right)|_{y\in\mathbb{R}}, \tag{2}$$

respectively, where $g_{\underline{\xi}}(y)$ and $G_{\underline{\xi}}(y)$ are the baseline PDF and CDF respectively with a parameter vector $\underline{\boldsymbol{\xi}}$. To this end, we define the CDF of the Xgamma reciprocal Rayleigh (XgRR) model. Using (1), the CDF of the XgRR can be written as

$$F_{\theta_1,\theta_2}(y) = 1 - \frac{1}{1+\theta_1}\left[1-e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^{\theta_1}\left(\begin{array}{c}1+\theta_1-\theta_1\log\left[1-e^{-\left(\frac{\theta_2}{y}\right)^2}\right]\\+\frac{1}{2}\theta_1^2\left\{\log\left[1-e^{-\left(\frac{\theta_2}{y}\right)^2}\right]\right\}^2\end{array}\right)|_{y\in\mathbb{R}^+}. \tag{3}$$

The PDF corresponding to (3) reduces to

$$f_{\theta_1,\theta_2}(y) = 2\theta_2^2\frac{\theta_1 y^{-3}e^{-\left(\frac{\theta_2}{y}\right)^2}}{1+\theta_1}\left[1-e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^{\theta_1-1}\left(\theta_1+\frac{1}{2}\theta_1^2\left\{\log\left[1-e^{-\left(\frac{\theta_2}{y}\right)^2}\right]\right\}^2\right)|_{y\in\mathbb{R}^+}. \tag{4}$$

The XgRR family density in (4) can be expressed as

$$f_{\theta_1,\theta_2}(y) = \frac{2\theta_1^2\theta_2^2}{1+\theta_1} y^{-3} \overbrace{e^{-\left(\frac{\theta_2}{y}\right)^2}}^{g_{\theta_2}(y)} \left[1 - e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^{\theta_1 - 1} + \left(\frac{\theta_1^3}{2(1+\theta_1)} \overbrace{2\theta_2^2 y^{-3} e^{-\left(\frac{\theta_2}{y}\right)^2}}^{g_{\theta_2}(y)} \left[1 - e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^{\theta_1 - 1} \underbrace{\left\{\log\left[1 - e^{-\left(\frac{\theta_2}{y}\right)^2}\right]\right\}^2}_{A(y;\theta_2)} \right) |_{y \in \mathbb{R}^+}. \tag{5}$$

Consider

$$\log\left(1 - \frac{a_1}{a_2}\right) = -\sum_{i=0}^{\infty} \frac{1}{i+1} \left(\frac{a_1}{a_2}\right)^{i+1} |_{\left|\frac{a_1}{a_2}\right| < 1}, \tag{6}$$

and the power series raised to a positive integer n (see Gradshteyn and Ryzhik (2002))

$$\left(\sum_{j=0}^{\infty} a_j\, u^j\right)^n = \sum_{j=0}^{\infty} c_{(n,j)}\, u^j, \tag{7}$$

where the coefficients $c_{(n,j)}$ (for $j = 1,2, \dots$) can be easily determined from the recurrence equation

$$c_{(n,j)} = (j a_0)^{-1} \sum_{m=1}^{j} [m(n+1) - j]\, a_m c_{(n,j-m)} \text{ and } c_{(n,0)} = a_0^n.$$

The coefficient $c_{(n,j)}$ can be calculated from $c_{(n,0)}, \dots, c_{(n,j-1)}$ and hence from the quantities $a_0, \dots, a_j$. For $\left|\frac{a_1}{a_2}\right| < 1$ and $a_3 > 0$, the power series holds

$$\left(1 - \frac{a_1}{a_2}\right)^{a_3} = \sum_{j=0}^{\infty} \frac{\Gamma(1+a_3)}{j!\,\Gamma(a_3 - j + 1)} \left(-\frac{a_1}{a_2}\right)^j. \tag{8}$$

Applying (6) to the quantity $A(y;\theta_2)$ in the PDF in (5), the PDF can be expressed as

$$f_{\theta_1,\theta_2}(y) = \frac{\theta_1^2}{1+\theta_1} \overbrace{2\theta_2^2 y^{-3} e^{-\left(\frac{\theta_2}{y}\right)^2}}^{g_{\theta_2}(y)} \left[1 - e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^{\theta_1 - 1} + \left(\frac{\theta_1^3}{2(1+\theta_1)} \overbrace{2\theta_2^2 y^{-3} e^{-\left(\frac{\theta_2}{y}\right)^2}}^{g_{\theta_2}(y)} \left[1 - e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^{\theta_1 - 1} \left[e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^2 \underbrace{\left\{\sum_{i=0}^{\infty} \frac{1}{i+1} e^{-i\left(\frac{\theta_2}{y}\right)^2}\right\}^2}_{B(y;\theta_2)} \right).$$

Expanding the quantity $B(y;\theta_2)$ using (7), the $f_{\theta_1,\theta_2}(y)$ can be written as

$$f_{\theta_1,\theta_2}(y)=\frac{\theta_1^2}{1+\theta_1}\overbrace{2\theta_2^2y^{-3}e^{-\left(\frac{\theta_2}{y}\right)^2}}^{g_{\theta_2}(y)}\underbrace{\left[1-e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^{\theta_1-1}}_{C(y;\theta_1,\theta_2)}$$

$$+\left\{\frac{\theta_1^3}{2(1+\theta_1)}\overbrace{2\theta_2^2y^{-3}e^{-\left(\frac{\theta_2}{y}\right)^2}}^{g_{\theta_2}(y)}\sum_{i=0}^{\infty}c_{2,i}\,e^{-(2+i)\left(\frac{\theta_2}{y}\right)^2}\underbrace{\left[1-e^{-\left(\frac{\theta_2}{y}\right)^2}\right]^{\theta_1-1}}_{C(y;\theta_1,\theta_2)}\right\},$$

where $c_{2,i}=1/(i+1)$. Applying the power series (8) to the quantity $C(y;\theta_1,\theta_2)$, we obtain

$$f(y)=\sum_{j=0}^{\infty}\left[\nabla_j\,\pi_{1+j}(y)+\sum_{i=0}^{\infty}\nabla_{i,j}\;\pi_{3+i+j}(y)\right], \tag{9}$$

where

$$\nabla_j=\frac{(-1)^j\theta_1^2\Gamma(\theta_1)}{(1+j)(1+\theta_1)\Gamma(\theta_1-j)},\ \nabla_{i,j}=\frac{(-1)^j\theta_1^3\Gamma(\theta_1)\;c_{2,i}}{2(1+\theta_1)(3+i+j)j!\,\Gamma(\theta_1-j)},$$

and $\pi_\varsigma(y)$ is the RR density with scale parameter $\theta_2\varsigma^{\frac{1}{2}}$ and shape parameter 2. So, the density of Y is a linear combination of RR densities. The CDF of Y follows by integrating (8) as

$$F_{\theta_1,\theta_2}(y)=\sum_{j=0}^{\infty}\left[\nabla_j\,H_{1+j}(y)+\sum_{i=0}^{\infty}\nabla_{i,j}\;H_{3+i+j}(y)\right], \tag{10}$$

where $H_\varsigma(y)$ is the RR density with scale parameter $\theta_2\varsigma^{\frac{1}{2}}$ and shape parameter 2. Equations (9) and (10) are the main results of this section. We provide some plots of the PDF and hazard rate function (HRF) of the XgRR model to show its flexibility. Figure 1 displays some plots of the XgRR density for selected parameter values. These plots reveal that the new density can be right skewed with different flexible shapes. The HRF plots of the XgRR distribution can be upside down or increasing. Many other useful real data sets can be found in Aryal and Yousof (2017), Merovci et al. (2017 and 2020), Korkmaz et al. (2017), Hamedani et al. (2017), Brito et al. (2017), Alizadeh et al. (2018), Korkmaz et al. (2018), Yousof et al. (2018a-d), Hamedani et al. (2018), Cordeiro et al. (2018), Hamedani et al. (2019), Ibrahim (2019), Nascimento et al. (2017, 2018 and 2019), Ibrahim et al. (2019), Goual and Yousof (2019), Korkmaz et al. (2019), Alizadeh et al. (2019) and Goual et al. (2020), Ibrahim (2020 a and b) and Yadav et al. (2020).

After studying the mathematical properties of the XgRR model, we discussed the estimation of the parameters using different estimation methods such as maximum likelihood estimation method, ordinary least squares estimation method, weighted least-squares estimation method, Cramer-Von-Mises estimation method and

bootstrapping method. Then, the Nikulin-Rao-Robson (N.R.R) statistic and modified N.R.R are discussed. In particular, the modified chi-squared test for composite hypothesis for complete samples was first considered by Nikulin (1973a, 1973b and 1973c), Rao and Robson (1974), among others. On the other hand, several goodness-of-fit tests have been suggested by the statisticians for censored data.

2. Properties

2.1. Moments

Let Y_{ς} be a rv having density $\pi_{\varsigma}(y)$. The r^{th} ordinary moment of Y, say $\mu'_{r,Y}$, follows from (9) as

$$\mu'_{r,Y} = E(Y^r) = \sum_{j=0}^{\infty}\left[\nabla_j\, E\left(Y_{1+j}^r\right) + \sum_{i=0}^{\infty}\nabla_{i,j}\; E\left(Y_{3+i+j}^r\right)\right].$$

Therefore,

$$\mu'_{r,Y} = \theta_2^r \Gamma\left(1-\frac{r}{2}\right)\sum_{j=0}^{\infty}\left[\nabla_j^{(r,1+j)} + \sum_{i=0}^{\infty}\nabla_{i,j}^{(r,3+i+j)}\right]\; |_{(2>r)}, \tag{11}$$

where

$$\nabla_j^{(r,1+j)} = b_j\,(1+j)^{\frac{r}{2}} \text{ and } \nabla_{i,j}^{(r,3+i+j)} = b_{i,j}\,(3+i+j)^{\frac{r}{2}},$$

and

$$\Gamma(1+\tau)|_{(\tau\in R^+)} = \tau! = \prod_{w=0}^{\tau-1}(\tau-w) = \int_0^{\infty} y^{\tau}\, eyp(-y)\, dy.$$

Setting $r = 1$ in (11) gives the mean of Y

$$E(Y) = \theta_2 \Gamma\left(1-\frac{1}{2}\right)\sum_{j=0}^{\infty}\left[\nabla_j\,(1+j)^{\frac{1}{2}} + \sum_{i=0}^{\infty}\nabla_{i,j}\;(3+i+j)^{\frac{1}{2}}\right].$$

2.2. Incomplete moments

The r^{th} incomplete moment of Y is defined by

$$m_{r,Y}(t) = \int_{-\infty}^{t} y^r f(y) dy.$$

We can write from (9)

$$m_{r,Y}(t) = \sum_{j=0}^{\infty}\left[\nabla_j\, m_{r,Y,1+j}(y) + \sum_{i=0}^{\infty}\nabla_{i,j}\; m_{r,Y,3+i+j}(y)\right].$$

Therefore,

$$m_{r,Y}(t) = \theta_2^r \sum_{j=0}^{\infty}\left[\begin{array}{c}\nabla_j^{(r,1+j)}\gamma\left(1-\frac{r}{2},(1+j)\left(\frac{\theta_2}{t}\right)^2\right) \\ +\sum_{i=0}^{\infty}\nabla_{i,j}^{(r,3+i+j)}\gamma\left(1-\frac{r}{2},(3+i+j)\left(\frac{\theta_2}{t}\right)^2\right)\end{array}\right]|_{(2>r)}, \tag{12}$$

where

$$\gamma(\tau,q)=\int_0^q t^{\tau-1}\,e^{-t}dt=\frac{q^{\tau}}{\tau}\{F_{1:1}[\tau;\tau\theta_2+1;-q]\}=\sum_{j=0}^{\infty}\frac{(-1)^j q^{\tau+j}}{j!\,(\tau+j)}=\Gamma(\tau)-\Gamma(\tau,q),$$

$$\Gamma(\tau,q)|_{(y>0)}=\int_q^{\infty} t^{\tau-1}\,eyp(-t)\,dt,$$

and $F_{1:1}[\cdot,\cdot,\cdot]$ is a confluent hypergeometric function (see Johnson et al. (2005)). Setting $r=1$ in (12) gives the first incomplete moment

$$m_{1,Y}(t)=\theta_2\sum_{j=0}^{\infty}\left[\begin{array}{c}\nabla_j^{(1,1+j)}\gamma\left(\frac{1}{2},(1+j)\left(\frac{\theta_2}{t}\right)^2\right)\\+\sum_{i=0}^{\infty}\nabla_{i,j}^{(1,3+i+j)}\gamma\left(\frac{1}{2},(3+i+j)\left(\frac{\theta_2}{t}\right)^2\right)\end{array}\right]$$

Two important applications of the $m_{1,Y}(t)$ are related to the mean deviation about the mean (S_1) and median and to the Bonferroni and Lorenz curves. The mean deviation about the mean

$$S_{1,Y}=E(|Y-E(Y)|)=2\mu'_{1,Y}F(E(Y))-2m_{1,Y}(E(Y))$$

and about the median

$$S_{2,Y}=E(|Y-M|)=E(Y)-2m_{1,Y}(M)$$

where $E(Y)$, $M=Q(u)=F^{-1}(u)$ is the median of Y, $F(\mu'_{1,Y})$ is easily calculated.

2.3. Moment generating function

The moment generating function (MGF) of Y, say $M_Y(t)=E(e^{tY})$, is obtained from (9) as

$$M_Y(t)=\sum_{j=0}^{\infty}\left[\nabla_j\,M_{Y,1+j}(t)+\sum_{i=0}^{\infty}\nabla_{i,j}\;M_{Y,3+i+j}(t)\right].$$

Therefore,

$$M_Y(t)=\sum_{j,r=0}^{\infty}\left[\begin{array}{c}\nabla_j[(t\theta_2)^r/r!](1+j)^{\frac{r}{2}}\Gamma\left(1-\frac{r}{2}\right)\\+\sum_{i=0}^{\infty}\nabla_{i,j}\;[(t\theta_2)^r/r!](3+i+j)^{\frac{r}{2}}\Gamma\left(1-\frac{r}{2}\right)\end{array}\right]\;|_{(2>r)}.$$

2.4. Convexity and concavity

Convex densities play an important role in several areas of mathematics. They are important in studying the "problems of optimization" where they are distinguished by several convenient characteristics. In mathematical analysis, a certain density defined on a certain n-dimensional interval is called "convex density" if the line between any two points on the graph of the density lies above the graph between the two points. The PDF in (5) is said to be "concave density" if for any $Y_1 \sim$ XgRR (θ_1,θ_2) and $Y_2 \sim$ XgRR (c_1,c_2) the PDF satisfies

$$f\left(\mathbf{b}y_1+\;\bar{\mathbf{b}}y_2\right)\geq\mathbf{b}f_{\theta_1,\theta_2}(y_1)+\;\bar{\mathbf{b}}f_{c_1,c_2}(y_2)|_{0\leq\mathbf{b}\leq1\text{ and }\;\bar{\mathbf{b}}=1-\mathbf{b}}.$$

If the function $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is twice differentiable, then if $f^{//}(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2) < 0, \forall\, y \in \mathbb{R}^+$, $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "strictly convex density". If $f^{//}(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2) \leq 0, \forall\, y \in \mathbb{R}^+$, then $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "convex". The density in (5) is said to be "convex density" if for any $Y_1 \sim$ XgRR (θ_1, θ_2) and $Y_2 \sim$ XgRR (c_1, c_2) the density satisfies

$$f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2) \leq \mathbf{b}f_{\theta_1,\theta_2}(y_1) + \bar{\mathbf{b}}f_{c_1,c_2}(y_2)|_{0\leq\mathbf{b}\leq 1 \text{ and } \bar{\mathbf{b}}=1-\mathbf{b}}.$$

If the function $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is twice differentiable, then if $f^{//}(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2) > 0, \forall\, y \in \mathbb{R}^+$, $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "strictly convex density". If $f^{//}(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2) \geq 0, \forall\, y \in \mathbb{R}^+$, then $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "convex". If $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "convex" and c is a constant, then the function $cf(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "convex". If $f(\mathrm{b}y_1 + \bar{\mathbf{b}}y_2)$ is "convex density", then $[cf(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)]$ is convex for every $c > 0$. If $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ and $g(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ are "convex density", then $[f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2) + g(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)]$ is also "convex density". If $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ and $g(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ are "convex density", then $[f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2).g(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)]$ is also "convex density". If the function $-f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "convex density", then the function $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "convex density". If $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "concave density", then $\frac{1}{f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)}$ is "convex density" if $f(y) > 0$. If $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "concave density", $\frac{1}{f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)}$ is "convex density" if $f(y) < 0$. If $f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)$ is "concave density", $\frac{1}{f(\mathbf{b}y_1 + \bar{\mathbf{b}}y_2)}$ is "convex density".

3. Copulas

For modelling the bivariate real data sets, we can consider the bivariate XgRR type generated via the FGM copula, modified FGM copula, Clayton copula and Renyi's entropy copula. Many other types of copula could be considered in separate works. In this Section, we derive some new bivariate type XgRR (BXgRR) model using the theorems of FGM copula, modified FGM copula, Clayton copula and Renyi's entropy. The Multivariate XgRR (MvXgRR) type is also presented. However, future works may be allocated to study these new models. First, we consider the joint CDF (JCDF) of the FGM family, where $C_\rho(m, w) = mw(1 + \rho m^* w^*)|_{m^*=1-m, w^*=1-w}$, where the marginal function $m = F_1 = F_{\theta_1,\theta_2}(y_1)$, $w = F_2 = F_{a_1,a_2}(y_2)$, $\rho \in (-1,1)$ is a dependence parameter and for every $m, w \in (0,1)$, $C(m, 0) = C(0, w) = 0$, which is "grounded minimum", and $C(m, 1) = m$ and $C(1, w) = w$ which is "grounded maximum", $C(m_1, w_1) + C(m_2, w_2) - C(m_1, w_2) - C(m_2, w_1) \geq 0$.

3.1. Via FGM copula

A copula is continuous in m and w where

$$|C(m_2, w_2) - C(m_1, w_1)| \leq |m_2 - m_1| + |w_2 - w_1|,$$

is the stronger Lipschitz condition. For $0 \leq m_1 \leq m_2 \leq 1$ and $0 \leq w_1 \leq w_2 \leq 1$, we have

$$Pr(m_1 \leq M \leq m_2, w_1 \leq W \leq w_2) = C(m_1, w_1) + C(m_2, w_2) - C(m_1, w_2) - C(m_2, w_1) \geq 0.$$

Then, setting

$$m^* = S_{\theta_1,\theta_2}(y_1) = 1 - F_{\theta_1,\theta_2}(y_1)|_{[m^*=(1-m)\in(0,1)]}$$

and

$$w^* = S_{a_1,a_2}(y_2) = 1 - F_{a_1,a_2}(y_2)|_{[w^*=(1-w)\in(0,1)]},$$

we can easily obtain the JCDF of the FGM family from

$$C_\rho(y_1, y_2) = F_{\theta_1,\theta_2}(y_1)F_{a_1,a_2}(y_2)\{1 + \rho[1 - F_{\theta_1,\theta_2}(y_1)][1 - F_{a_1,a_2}(y_2)]\},$$

where

$$F_{\theta_1,\theta_2}(y_1) = 1 - \frac{\hbar_{\theta_2}(y_1)^{\theta_1}}{1+\theta_1}\left\{\begin{matrix}1 + \theta_1 - \theta_1 \log \hbar_{\theta_2}(y_1) \\ + \frac{1}{2}\theta_1^2[\log \hbar_{\theta_2}(y_1)]^2\end{matrix}\right\},$$

$$\hbar_{\theta_2}(y_1) = 1 - e^{-\left(\frac{\theta_2}{y_1}\right)^2},$$

$$F_{a_1,a_2}(y_2) = 1 - \frac{\hbar_{a_2}(y_2)^{a_1}}{1+a_1}\left\{\begin{matrix}1 + a_1 - a_1 \log \hbar_{a_2}(y_2) \\ + \frac{1}{2}\theta_1^2[\log \hbar_{a_2}(y_2)]^2\end{matrix}\right\},$$

The JPDF can then be derived from

$$c_\rho(m, w) = 1 + \rho(1 - 2m)(1 - 2w)$$

or from

$$f(y_1, y_2) = C(F_1, F_2)f_1 f_2.$$

3.2. Via modified FGM copula

The modified FGM copula is defined as

$$C_\rho(m, w) = mw[1 + \rho V(m)A(w)]|_{\rho\in(-1,1)}$$

or

$$C_\rho(m, w) = mw + \rho\dot{V}(m)\dot{A}(w)|_{\rho\in(-1,1)},$$

where $\dot{V}(m) = mV(m)$, and $\dot{A}(w) = wA(w)$, where $V(m)$ and $A(w)$ are being two continuous functions on $(0,1)$ where $V(0) = V(1) = A(0) = A(1) = 0$. Let

$$c_1 = inf\left\{\frac{\partial}{\partial m}\dot{V}(m)|\boldsymbol{q}_1(m)\right\} < 0, c_2 = sup\left\{\frac{\partial}{\partial m}\dot{V}(m)|\boldsymbol{q}_1(m)\right\} < 0,$$

$$d_1 = inf\left\{\frac{\partial}{\partial w}\dot{A}(w)|\boldsymbol{q}_2(w)\right\} > 0$$

and

$$d_2 = sup\left\{\frac{\partial}{\partial w}\dot{A}(w)|\boldsymbol{q}_2(w)\right\} > 0.$$

Then, $1 \leq min(c_1 c_2, d_1 d_2) \leq \infty$, where

$$\frac{\partial}{\partial m} V(m) = V(m) + m \frac{\partial}{\partial m} V(m),$$

$$\boldsymbol{q}_1(u) = \left\{ m: m \in (0,1) | \frac{\partial \dot{V}(m)}{\partial m} \text{ exists} \right\}$$

and

$$\boldsymbol{q}_2(w) = \left\{ w: w \in (0,1) | \frac{\partial \dot{A}(w)}{\partial w} \text{ exists} \right\}.$$

Type-I

Recall the following functional forms for both $V(m)$ and $A(w)$. Then, the BXgRR-FGM (Type-**I**) can be derived from

$$C_\rho(y_1, y_2) = F_{\theta_1,\theta_2}(y_1) F_{a_1,a_2}(y_2) + \rho \dot{V}(m) \dot{A}(y_2)|_{\rho \in (-1,1)}$$

where

$$\dot{V}(y_1) = F_{\theta_1,\theta_2}(y_1) S_{\theta_1,\theta_2}(y_1)$$

and

$$\dot{A}(y_2) = F_{a_1,a_2}(y_2) S_{a_1,a_2}(y_2).$$

Type-II

Let $V(y_1)^*$ and $A(y_2)^*$ be two functional forms satisfying all the conditions stated earlier where

$$V(y_1)^*|_{(\rho_1 > 0)} = S_{\theta_1,\theta_2}(y_1)^{1-\rho_1} F_{\theta_1,\theta_2}(y_1)^{\rho_1}$$

and

$$A(y_2)^*|_{(\rho_2 > 0)} = S_{a_1,a_2}(y_2)^{1-\rho_2} F_{a_1,a_2}(y_2)^{\rho_2}.$$

Then, the corresponding BXgRR-FGM (Type-**II**) can be derived from

$$C_{\rho,\rho_1,\rho_2}(y_1, y_2) = F_{\theta_1,\theta_2}(y_1) F_{a_1,a_2}(y_2) [1 + \rho V(y_1)^* A(y_2)^*].$$

Type-III

Let $\ddot{V}(y_1)$ and $\ddot{A}(y_2)$ be two functional forms for satisfying all the conditions stated earlier where

$$\ddot{V}(y_1) = F_{\theta_1,\theta_2}(y_1) \log\left[1 + S_{\theta_1,\theta_2}(y_1)\right],$$

and

$$\ddot{A}(y_2) = F_{a_1,a_2}(y_2) \log\left[1 + S_{a_1,a_2}(y_2)\right].$$

In this case, one can also derive a closed form expression for the associated CDF of the BXgRR-FGM (Type-**III**) from

$$C_\rho(y_1, y_2) = F_{\theta_1,\theta_2}(y_1) F_{a_1,a_2}(y_2) \left[1 + \rho \ddot{V}(m)\, \ddot{A}(m)\right].$$

3.3. Via Clayton copula

The Clayton copula can be considered as

$$C(w_1, w_2) = [(1/w_1)^\rho + (1/w_2)^\rho - 1]^{-\rho^{-1}}|_{\rho \in (0,\infty)}.$$

Setting $w_1 = F_{\theta_1,\theta_2}(y_1) \in (0,1)$ and $w_2 = F_{a_1,a_2}(y_2) \in (0,1)$, the BXgRR type can be derived from $C(y_1, y_2) = C(F_{\theta_1,\theta_2}(y_1), F_{a_1,a_2}(y_2))$. Similarly, the MvXgRR (D-dimensional extension) from the above can be derived from

$$C(w_h) = \left(\sum_{h=1}^{D} w_h^{-\rho} + 1 - D\right)^{-\rho^{-1}}.$$

3.4. Via Renyi's entropy

Let $m \in (0,1) = F_{\theta_1,\theta_2}(y_1)$ and $w \in (0,1) = F_{a_1,a_2}(y_2)$. Then, the Renyi's entropy copula can be expressed as

$$C(y_1, y_2) = y_2 F_{\theta_1,\theta_2}(y_1) + y_1 F_{a_1,a_2}(y_2) - y_1 y_2.$$

Then, the associated BXgRR can be directly derived from

$$C(y_1, y_2) = C(F_{\theta_1,\theta_2}(y_1), F_{a_1,a_2}(y_2)).$$

3.5. Via Ali-Mikhail-Haq copula

Under the stronger Lipschitz condition, the Archimedean Ali-Mikhail-Haq copula can be expressed as

$$C_\rho(\mathrm{u}, \varpi) = \frac{\mathrm{u}\varpi}{1 - \rho\overline{\mathrm{u}}\,\overline{\varpi}}\Big|_{\rho\in(-1,1)}.$$

Then, for $\overline{\mathrm{u}} = 1 - F_{\theta_1,\theta_2}(y_1)$, $\overline{\varpi} = 1 - F_{a_1,a_2}(y_2)$ we have the following Bv XgRR type

$$C_\rho(y_1, y_2) = \frac{F_{\theta_1,\theta_2}(y_1) F_{a_1,a_2}(y_2)}{1 - \rho S_{\theta_1,\theta_2}(y_1) S_{a_1,a_2}(y_2)}\Big|_{\rho\in(-1,1)}.$$

4. Estimation

4.1. Maximum likelihood estimation (MLE)

Here, we consider the estimation of the unknown parameters of the new family from complete samples by maximum likelihood. Let $y_1, \cdots, y_n$ be a random sample from the XgRR model with a (2×1) parameter vector. The log-likelihood function for θ_1, θ_2 is given by

$$\ell_n(\theta_1, \theta_2) = n\log\theta_1 - n\log(1+\theta_1) + n\,log\,2 + 2n\log\theta_2 - 3\sum_{i=1}^{n}\log(y_i)$$
$$-\sum_{i=1}^{n}\left(\frac{\theta_2}{y_i}\right)^2 + (\theta_1 - 1)\sum_{i=1}^{n}\log[\hbar_{\theta_2}(y_i)] + \sum_{i=1}^{n}\log\left(\theta_1 + \frac{1}{2}\theta_1^2\{\log[\hbar_{\theta_2}(y_i)]\}^2\right).$$

where $\hbar_{\theta_2}(y_i) = 1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2}$. The log-likelihood function in $\ell_n(\theta_1, \theta_2)$ can be maximized numerically by using R (optim), SAS (PROC NLMIXED) or Ox program

(sub-routine MaxBFGS), among others. For interval estimation of the parameters, the elements of the 2×2 observed information matrix $\mathbf{J}(\theta_1, \theta_2)$ can be evaluated numerically.

4.2. Ordinary and weighted least-squares estimators

The theory of least square estimation and weighted least square estimation was proposed by Swain et al. (1988) to estimate the parameters of the Beta distribution. It is based on the minimization of the sum of the square of differences of theoretical cumulative distribution function and empirical distribution function. Suppose $F_{\theta_1,\theta_2}(y_{i:n})$ denotes the distribution function of the XgRR distribution and $y_1 < y_2 < \cdots < y_n$ be the n ordered random sample. The ordinary least square estimates (OLSEs) are obtained by minimizing

$$OLS(\theta_1, \theta_2) = \sum_{i=1}^{n} \left[F_{\theta_1,\theta_2}(y_{i:n}) - \frac{i}{n+1} \right]^2.$$

Now, using (1) we have

$$OLS(\theta_1, \theta_2) = \sum_{i=1}^{n} \left\{ \left[1 - \frac{\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right]^{\theta_1}}{1+\theta_1} \left(\begin{array}{l} 1 + \theta_1 - \theta_1 \log\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right] \\ \\ + \frac{1}{2}\theta_1^2 \left\{ \log\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right] \right\}^2 \end{array} \right) \right] - \frac{i}{n+1} \right\}^2.$$

Then, least square estimators (LSE) of the parameters are obtained by simultaneously solving the following non-linear equations:

$$\sum_{i=1}^{n} \left\{ \left[1 - \frac{\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right]^{\theta_1}}{1+\theta_1} \left(\begin{array}{l} 1 + \theta_1 - \theta_1 \log\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right] \\ \\ + \frac{1}{2}\theta_1^2 \left\{ \log\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right] \right\}^2 \end{array} \right) \right] - \frac{i}{n+1} \right\} \xi_{\theta_1|\theta_1,\theta_2}(y_i) = 0,$$

and

$$\sum_{i=1}^{n} \left\{ \left[1 - \frac{\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right]^{\theta_1}}{1+\theta_1} \left(\begin{array}{l} 1 + \theta_1 - \theta_1 \log\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right] \\ \\ + \frac{1}{2}\theta_1^2 \left\{ \log\left[1 - e^{-\left(\frac{\theta_2}{y_i}\right)^2} \right] \right\}^2 \end{array} \right) \right] - \frac{i}{n+1} \right\} \xi_{\theta_2|\theta_1,\theta_2}(y_i) = 0,$$

where $\xi_{\theta_1|\theta_1,\theta_2}(y_i)$ and $\xi_{\theta_2|\theta_1,\theta_2}(y_i)$ are the values of the first derivatives with respect to parameters of XgRR distribution. The weighted least squares estimates (WLSE) are obtained by minimizing the given form of equation with respect to the parameters

$$WLS(\theta_1, \theta_2) = \sum_{i=1}^{n} \frac{(n+1)^2(n+2)}{i(n-i+1)} \left[F_{\theta_1,\theta_2}(y_i) - \frac{i}{n+1} \right]^2.$$

The WLSE of the parameters are obtained by solving the following non-linear equations;

$$\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}\left\{\left[1-\frac{\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]^{\theta_1}}{1+\theta_1}\left(\begin{array}{c}1+\theta_1-\theta_1\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\\ +\frac{1}{2}\theta_1^2\left\{\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\right\}^2\end{array}\right)\right]-\frac{i}{n+1}\right\}\xi_{\theta_1|\theta_1,\theta_2}(y_i)=0,$$

and

$$\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}\left\{\left[1-\frac{\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]^{\theta_1}}{1+\theta_1}\left(\begin{array}{c}1+\theta_1-\theta_1\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\\ +\frac{1}{2}\theta_1^2\left\{\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\right\}^2\end{array}\right)\right]-\frac{i}{n+1}\right\}\xi_{\theta_2|\theta_1,\theta_2}(y_i)=0,$$

where $\xi_{\theta_1|\theta_1,\theta_2}(y_i)$ and $\xi_{\theta_2|\theta_1,\theta_2}(y_i)$ are the values of first derivatives of the CDF of XgRR distribution.

4.3. Method of Cramer-Von-Mises estimation

The Cramer-Von-Mises estimation (CVME) method of the parameters is based on the theory of minimum distance estimation. It was proposed by MacDonald (1971) and justified that the bias of the estimator is smaller than the other minimum distance estimators. Thus, The Crammer-Von-Mises estimates of the parameter θ_1 and θ_2 are obtained by minimizing the following expression with respect to the parameters θ_1 and θ_2 respectively.

$$CVM(\theta_1,\theta_2)=\frac{1}{12n}+\sum_{i=1}^{n}\left[F_{\theta_1,\theta_2}(y_i)-\frac{2i-1}{2n}\right]^2,$$

and

$$CVM(\theta_1,\theta_2)=\sum_{i=1}^{n}\left[1-\frac{\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]^{\theta_1}}{1+\theta_1}\left(\begin{array}{c}1+\theta_1-\theta_1\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\\ +\frac{1}{2}\theta_1^2\left\{\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\right\}^2\end{array}\right)-\frac{2i-1}{2n}\right]^2.$$

The CVME of the parameters is obtained by solving the following non-linear equations

$$\sum_{i=1}^{n}\left[1-\frac{\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]^{\theta_1}}{1+\theta_1}\left(\begin{array}{l}1+\theta_1-\theta_1\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\\ +\frac{1}{2}\theta_1^2\left\{\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\right\}^2\end{array}\right)-\frac{2i-1}{2n}\right]\xi_{\theta_1|\theta_1,\theta_2}(y_i)=0,$$

and

$$\sum_{i=1}^{n}\left[1-\frac{\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]^{\theta_1}}{1+\theta_1}\left(\begin{array}{l}1+\theta_1-\theta_1\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\\ +\frac{1}{2}\theta_1^2\left\{\log\left[1-e^{-\left(\frac{\theta_2}{y_i}\right)^2}\right]\right\}^2\end{array}\right)-\frac{2i-1}{2n}\right]\xi_{\theta_2|\theta_1,\theta_2}(y_i)=0,$$

where $\xi_{\theta_1|\theta_1,\theta_2}(y_i)$ and $\xi_{\theta_2|\theta_1,\theta_2}(y_i)$ are the values of the first derivatives of the CDF of XgRR distribution with respect to θ_1 and θ_2 respectively.

4.4. Bootstrapping method

Bootstrapping method is a powerful statistical technique. It is especially useful when the sample size that we are working with is small. Under the usual circumstances, sample sizes of less than 40 cannot be dealt with by assuming a normal or a t distributions. Bootstrap techniques work quite well with samples that have less than 40 elements. The reason for this is that bootstrapping involves resampling. These kinds of techniques assume nothing about the distribution of our data. Bootstrapping has become more popular as computing resources have become more readily available.

5. Numerical results for comparing estimation methods

In this Section, a Monte Carlo simulation study is conducted for comparing the performance of the different estimators of the unknown parameters of the XgRR distribution. The performance of the different estimators proposed in the previous Section is evaluated in terms of their mean squared errors (MSEs). All the computations in this section are done by Mathcad program Version 15.0. We generate 1000 samples of the XgRR distribution, where n = (20,50,100,200,500) and θ_1 and θ_2 are chosen as follows:

	I	II	III
θ_2	2.0	0.9	1.2
θ_1	1.5	0.3	0.6

The average values (AVs) of estimates and the corresponding MSEs of MLEs, LSEs, WLSEs, CVM, MPSD and Bootstrap method are obtained and reported in Tables 1, 2, 3, 4 and 5. We observe that all the estimates show the property of consistency, i.e. the MSEs decrease as the sample size increases.

Table 1. AVs and the corresponding MSEs (in parentheses) for n=20

Parameters	MLE	LS	WLS	CVM	Bootstrap
θ_2=2.0	2.06686	2.07031	2.09047	2.08087	3.14584
	(0.21967)	(0.20541)	(0.24017)	(0.23470)	(4.25306)
θ_1=1.5	1.52659	1.52061	1.51782	1.51839	1.19740
	(0.02907)	(0.03600)	(0.03636)	(0.04756)	(1.19181)
θ_2=0.9	0.99944	0.93704	0.93355	0.92649	0.77474
	(0.35047)	(0.04215)	(0.03902)	(0.03634)	(0.03767)
θ_1=0.3	0.31287	0.28547	0.23809	0.23838	0.43615
	(0.00451)	(1.15021)	(5.42295)	(5.30798)	(0.03477)
θ_2=1.2	1.24013	1.25164	1.24611	1.23804	1.03133
	(0.07944)	(0.08231)	(0.07525)	(0.07098)	(0.08032)
θ_1=0.6	0.62025	0.60920	0.61635	0.61905	0.79832
	(0.00956)	(0.04099)	(0.01483)	(0.01549)	(0.07196)

Table 2. AVs and the corresponding MSEs (in parentheses) for n=50.

Parameters	MLE	LS	WLS	CVM	Bootstrap
θ_2=2.0	2.03571	2.03548	2.04436	2.04309	1.89752
	(0.07429)	(0.08252)	(0.07604)	(0.07831)	(0.07750)
θ_1=1.5	1.50659	1.50738	1.50208	1.50268	1.56873
	(0.01035)	(0.01499)	(0.01267)	(0.01405)	(0.02128)
θ_2=0.9	0.97998	0.91650	0.91229	0.90955	1.01084
	(0.11707)	(0.01234)	(0.01270)	(0.01250)	(0.02619)
θ_1=0.3	0.29945	0.30253	0.30513	0.30620	0.27442
	(0.00206)	(0.00193)	(0.00195)	(0.00223)	(0.00152)
θ_2=1.2	1.22066	1.21680	1.20984	1.20756	1.19130
	(0.02568)	(0.02424)	(0.02090)	(0.02104)	(0.01494)
θ_1=0.6	0.60589	0.60694	0.60833	0.60960	0.61487
	(0.00319)	(0.00528)	(0.00421)	(0.00478)	(0.00325)

Table 3. AVs and the corresponding MSEs (in parentheses) for n=100

Parameters	MLE	LS	WLS	CVM	Bootstrap
θ_2=2.0	2.01577	2.01094	2.01555	2.01478	1.87189
	(0.03717)	(0.03506)	(0.03362)	(0.03489)	(0.03867)
θ_1=1.5	1.50394	1.50564	1.50324	1.50379	1.56382
	(0.00515)	(0.00693)	(0.00605)	(0.00673)	(0.00922)
θ_2=0.9	0.96579	0.90754	0.90982	0.90921	0.93491
	(0.05381)	(0.00584)	(0.00524)	(0.00541)	(0.01081)
θ_1=0.3	0.29496	0.30160	0.30052	0.30070	0.29517
	(0.00134)	(0.00096)	(0.00075)	(0.00092)	(0.00114)
θ_2=1.2	1.20965	1.21049	1.20931	1.20922	1.14055
	(0.01303)	(0.01178)	(0.01082)	(0.01108)	(0.01160)
θ_1=0.6	0.60315	0.60230	0.60269	0.60268	0.57792
	(0.00155)	(0.00245)	(0.00207)	(0.00238)	(0.00253)

Table 4. AVs and the corresponding MSEs (in parentheses) for n=200

Parameters	MLE	LS	WLS	CVM	Bootstrap
θ_2=2.0	2.00332	2.01138	2.00587	2.00564	1.84919
	(0.01841)	(0.01710)	(0.01740)	(0.01811)	(0.03575)
θ_1=1.5	1.50390	1.50005	1.50278	1.50295	1.57880
	(0.00263)	(0.00329)	(0.00313)	(0.00350)	(0.00910)
θ_2=0.9	0.95153	0.90173	0.90503	0.90449	0.98419
	(0.02540)	(0.00267)	(0.00264)	(0.00270)	(0.01048)
θ_1=0.3	0.29346	0.30154	0.30022	0.30040	0.27433
	(0.00087)	(0.00046)	(0.00038)	(0.00046)	(0.00091)
θ_2=1.2	1.20178	1.19981	1.20636	1.20563	1.25721
	(0.00634)	(0.00491)	(0.00513)	(0.00527)	(0.00807)
θ_1=0.6	0.60278	0.60331	0.60044	0.60074	0.57895
	(0.00079)	(0.00111)	(0.00099)	(0.00115)	(0.00119)

Table 5. AVs and the corresponding MSEs (in parentheses) for n=500.

Parameters	MLE	LS	WLS	CVM	Bootstrap
θ_2=2.0	1.99647	1.99838	1.99702	1.99672	1.99960
	(0.00668)	(0.00654)	(0.00629)	(0.00666)	(0.00620)
θ_1=1.5	1.50344	1.50277	1.50334	1.50356	1.50239
	(0.00100)	(0.00133)	(0.00121)	(0.00137)	(0.00136)
θ_2=0.9	0.94897	0.89934	0.89870	0.89860	0.90144
	(0.01044)	(0.00104)	(0.00629)	(0.00106)	(0.00105)
θ_1=0.3	0.28999	0.30117	0.30139	0.30151	0.30039
	(0.00049)	(0.00019)	(0.00016)	(0.00019)	(0.00018)
θ_2=1.2	1.19801	1.20037	1.20245	1.20247	1.22109
	(0.00229)	(0.00198)	(0.00196)	(0.00202)	(0.00292)
θ_1=0.6	0.60204	0.60112	0.60018	0.60015	0.59211
	(0.00030)	(0.00044)	(0.00038)	(0.00044)	(0.00049)

6. Modified Right-Censored Test for Validation

6.1. The N.R.R statistic test

Many goodness-of-fit tests are used to indicate whether or not it is reasonable to assume that a random sample comes from a specific distribution. For this purpose, researchers proposed many different goodness-of-fit tests. For the complete data, Nikulin (1973a, 1973b and 1973c) and Rao and Robson (1974) separately proposed a statistic known today as the N.R.R statistic. This statistical test is a natural modification of the Pearson statistic. To test the hypothesis H_0 we have

$$H_0: P\{T_i \leq t\} = F(t,\zeta)|_{(t\in R,\ \zeta=(\zeta_1,\zeta_2,\cdots,\zeta_s)^T)},$$

where ζ represents the vector of unknown parameters. Nikulin (1973a, 1973b and 1973c) and Rao and Robson (1974) proposed the N.R.R statistic defined as follows: Observations $T_1, T_2, \cdots, T_n$ are grouped in r subintervals and $\nu_j = (\nu_1, \nu_2, \cdots, \nu_r)^T$ is the vector of frequencies, where ν_j is frequency of ith group and $\sum_{j=1}^{r} \nu_j = n$. The tests are based on the following Pearson's statistic

$$Y^2\left(\underset{\sim}{\hat{\zeta}}_n\right) = \chi_n^2\left(\underset{\sim}{\hat{\zeta}}_n\right) + n^{-1}\ell^T\left(\underset{\sim}{\hat{\zeta}}_n\right)\left(\mathbf{I}\left(\underset{\sim}{\hat{\zeta}}_n\right) - \mathbf{J}\left(\underset{\sim}{\hat{\zeta}}_n\right)\right)^{-1}\ell\left(\underset{\sim}{\hat{\zeta}}_n\right),$$

where

$$\chi_n^2(\zeta) = \left(\frac{\nu_1 - np_1(\zeta)}{\sqrt{np_1(\zeta)}}, \frac{\nu_2 - np_2(\zeta)}{\sqrt{np_2(\zeta)}}, \cdots, \frac{\nu_r - np_r(\zeta)}{\sqrt{np_r(\zeta)}}\right)^T,$$

and $p(\zeta)$ is the vector of probabilities and ζ is the vector of parameters which can be known (simple hypothesis) or unknown (composite hypothesis). The Y^2 statistic follows a chi-square distribution with $(r-1)$ degrees of freedom (for more details see Nikulin (1973a, 1973b and 1973c)).

6.2. Application to right-censored real data

To test the null hypothesis H_0, we use the N.R.R statistic. We compute the maximum likelihood estimators

$$\widehat{\theta_1} = 0.95473 \text{ and } \widehat{\theta_2} = 1.24885.$$

We then deduce the value of $Y^2 = 11.05847$. The critical value is

$$\chi^2_{0.05}(6-1) = 11.0705.$$

Then, the N.R.R Y^2 statistic value is less than the critical value, we say that taxes revenue data can be fitted by the XgRR model. The modified chi-squared test for composite hypothesis for complete samples was first considered by Nikulin (1973a, 1973b and 1973c), Rao and Robson (1974). Several goodness-of-fit tests have been suggested by the statisticians for censored data. Bagdonavicius and Nikulin $(2011a, b)$ proposed a modification of the N.R.R statistic that takes into account random right censorship and based on the maximum likelihood estimators on the initial data, also follows a limiting Chi-square distribution. In this Section we develop the approach proposed by Bagdonavicius and Nikulin $(2011a, b)$ to confirm the adequacy of XgRR model when the parameters are unknown and data are censored. Let us consider the composite hypothesis

$$H_0 \ : \ F(t) \in F_0 = F_0(t, \zeta)|_{(t\in R^1,\ \zeta\in\Psi\subset R^s)},$$

where ζ is an unknown m-dimensional parameter and F_0 is a differentiable and completely specified cdf with the support $(0, \infty)$. Let us consider a finite time interval, say, $[0, \tau]$, where τ is the maximum time of the study, and divide it into $k > s$ smaller intervals $I_j = (a_{j-1}, a_j]$, where

$$0 < a_0 < a_1 \ldots < a_{k-1} < a_k < +\infty.$$

In this case the estimated $\hat{a}_{\boldsymbol{k}}$ is given by

$$\hat{a}_{\boldsymbol{k}} = \Lambda^{-1}\left\{\frac{1}{n-i+1}\left[E_j - \sum_{l=1}^{i-1}\Lambda\left(\mathrm{T}_{(l)}, \underline{\hat{\zeta}}\right)\right], \underline{\hat{\zeta}}\right\}, \hat{a}_{\boldsymbol{k}} = \mathrm{t}_{(n)} \,|\, \boldsymbol{j} = 1,2,\dots,k$$

where $\underline{\hat{\zeta}}$ is the maximum likelihood estimator of the parameter ζ, Λ^{-1} is the inverse of cumulative hazard function Λ, $T_{(i)}$ is the i^{th} element in the ordered statistics $(T_{(1)}, \dots, T_{(n)})$ and

$$E_{\boldsymbol{j}} = (n+1-i)\Lambda\left(\hat{a}_{(j)}, \underline{\hat{\zeta}}\right) + \sum_{\boldsymbol{l}=1}^{i-1}\Lambda\left(\mathrm{T}_{(l)}, \underline{\hat{\zeta}}\right),$$

and a_j are random data functions such as the k intervals have equal expected numbers of failures e_j. Usually in real application we fix k. The test statistic for H_0 is given in Goual et al. (2020) and Goual and Yousof (2019). The survival times in days are for the $n = 51$ patients. The data are: 7, 34, 42, 63, 64, 74*, 83, 84, 91, 108, 112, 129, 133, 133, 139, 140, 140, 146, 149, 154, 157, 160, 160, 165, 173, 176, 185*, 218, 225, 241, 248, 273, 277, 279*, 297, 319*, 405, 417, 420, 440, 523*, 523, 583, 594, 1101, 1116*, 1146, 1226*, 1349*, 1412*, 1417. (* censored). We suppose that these data are distributed according to the XgRR distribution, we transform the survival times in months (1 month = 30.438 days), so the maximum likelihood estimates of the parameter vector ζ are

$$\zeta = (5.00248, 1.378452)^T$$

We choose $r = 7$ as the number of classes. The elements of the test statistic Y_n^2 is presented as follows, we find $Y_n^2 = 14.000154$ and the critical value $\chi^2_{0.05}(7) = 14.00924$. Comparing the critical value and the statistic test Y_n^2, we can say that Arm-A head and neck cancer data can be adjusted by the XgRR model.

7. Concluding remarks

In this article, a new reciprocal Rayleigh extension called the Xgamma reciprocal Rayleigh model is defined and studied. Relevant statistical properties such as raw moments, incomplete moments and moment generating function are derived. After a quick study for their properties, different non-Bayesian estimation methods under uncensored schemes are considered and described such as the maximum likelihood estimation method, ordinary least square estimation method, weighted least square estimation method, Cramér–von-Mises estimation method and Bootstrapping method. The performances of the proposed estimation methods are investigated through a simulation study. Many bivariate and multivariate type models have been also derived based on Farlie Gumbel Morgenstern copula, Clayton copula, Renyi's entropy copula and Ali–Mikhail–Haq copula. A modified right-censored test for validation is applied to a right-censored real data set.

As a potential future work, we can use and apply the well-known Bagdonavičius-Nikulin goodness-of-fit test and the modified version of the Bagdonavičius-Nikulin goodness-of-fit test to our new XgRR model and many other useful lifetime models (see Goual et al. (2019), Ibrahim et al. (2020), Yadav et al. (2020) and Mansour et al. (2020a-f) for more details). The reciprocal Rayleigh distribution can be extended using some new G families such as presented by Alizadeh et al. (2020) and El-Morshedy et al. (2021). Some useful real-life data sets can be cited from Elgohari and Yousof (2020a and 2020b).

References

Alizadeh, M., Ghosh, I., Yousof, H. M., Rasekhi, M. and Hamedani G. G., (2017). The generalized odd generalized exponential family of distributions: properties, characterizations and applications, J. Data Sci., Vol. 15, pp. 443–466.

Alizadeh, M., Jamal, F., Yousof, H. M., Khanahmadi, M. and Hamedani, G. G., (2020). Flexible Weibull generated family of distributions: characterizations, mathematical properties and applications. University Politehnica of Bucharest Scientific Bulletin-Series A-Applied Mathematics and Physics, Vol. 82, pp. 145–150.

Alizadeh, M., Rasekhi, M., Yousof, H. M. and Hamedani G. G., (2018). The transmuted Weibull G family of distributions. Hacettepe Journal of Mathematics and Statistics, Vol. 47, pp. 1–20.

Alizadeh, M., Yousof, H. M. Rasekhi, M. and Altun, E., (2019). The odd log-logistic Poisson-G Family of distributions, Journal of Mathematical Extensions, Vol. 12, pp. 81–104.

Aryal, G. R., Yousof, H. M., (2017). The exponentiated generalized-G Poisson family of distributions. Economic Quality Control, 32, pp. 1–17.

Bagdonavicius, V., Nikulin, M., (2011a). Chi-squared Goodness-of-fit Test for Right Censored Data. International Journal of Applied Mathematics and Statistics, Vol. 24, pp. 30–50.

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

Brito, E., Cordeiro, G. M., Yousof, H. M., Alizadeh, M. and Silva, G. O., (2017). Topp-Leone Odd Log-Logistic Family of Distributions, Journal of Statistical Computation and Simulation, Vol. 87, pp. 3040–3058.

Cordeiro, G. M., Altun, E., Korkmaz, M. C., Pescim, R. R., Afify, A. Z. and Yousof, H. M., (2020). The xgamma Family: Censored Regression Modelling and Applications. Revista Colombiana de Estadística, Vol. 18, pp. 593–612.

Cordeiro, G. M., Yousof, H. M., Ramires, T. G. and Ortega, E. M. M., (2018). The Burr XII system of densities: properties, regression model and applications. Journal of Statistical Computation and Simulation, Vol. 88, pp. 432–456.

Chakraborty, S., Handique, L., Altun, E. and Yousof, H. M., (2018). A new statistical model for extreme values: mathematical properties and applications. International Journal of Open Problems in Computer Science and Mathematics, Vol. 12, pp. 1–18.

Elgohari, H., Yousof, H. M., (2020a). A Generalization of Lomax Distribution with Properties, Copula and Real Data Applications. *Pakistan Journal of Statistics and Operation Research*, Vol. 16, pp. 697–711.

Elgohari, H., Yousof, H. M., (2020b). New Extension of Weibull Distribution: Copula, Mathematical Properties and Data Modeling. *Statistics, Optimization & Information Computing*, Vol. 8, pp. 972–993.

El-Morshedy, M., Alshammari, F. S., Hamed, Y. S., Eliwa, M. S., Yousof, H. M., (2021). A new family of continuous probability distributions. Entropy, Vol. 23, 194. https://doi.org/10.3390/e23020194

Fréchet, M., (1927). Sur la loi de probabilité de lécart maximum. Ann. Soc. Pol. Math., Vol. 6, pp. 93–116.

Gradshteyn, I. S., Ryzhik, I. M., (2002). Table of integrals, series, and products. San Diego, CA: Academic Press.

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

Goual, H., Yousof, H. M., (2019). Validation of Burr XII inverse Rayleigh model via a modified chi-squared goodness-of-fit test. Journal of Applied Statistics, Vol. 47, pp. 1–32.

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

Gusmao, F. R. S., Ortega, E. M. M. and Cordeiro, G. M., (2011). The generalized inverse Weibull distribution. Statistical Papers, Vol. 52, pp. 591–619.

Haq, M. A., Yousof, H. M. and Hashmi, S., (2017). A New Five-Parameter Fréchet Model for Extreme Values. Pakistan Journal of Statistics and Operation Research, Vol. 13(3), pp. 617–632.

Hamedani G. G., Altun, E, Korkmaz, M. C., Yousof, H. M. and Butt, N. S., (2018). A new extended G family of continuous distributions with mathematical properties, characterizations and regression modeling. Pakistan Journal of Statistics and Operation Research, Vol. 14, pp. 737–758.

Hamedani G. G., Rasekhi, M., Najibi, S. M., Yousof, H. M. and Alizadeh, M., (2019). Type II general exponential class of distributions. Pakistan Journal of Statistics and Operation Research , Vol. 15, pp. 503–523.

Hamedani G. G., Yousof, H. M., Rasekhi, M., Alizadeh, M. and Najibi, S. M., (2017). Type I general exponential class of distributions. Pakistan Journal of Statistics and Operation Research, Vol. 14, pp. 39–55.

Ibrahim, M., (2019). A new extended Fréchet distribution: properties and estimation. Pakistan Journal of Statistics and Operation Research, Vol. 15, pp. 773–796.

Ibrahim, M., (2020a). The compound Poisson Rayleigh Burr XII distribution: properties and applications. Journal of Applied Probability and Statistics, Vol. 15, pp. 73–97.

Ibrahim, M., (2020b). The generalized odd Log-logistic Nadarajah Haghighi distribution: statistical properties and different methods of estimation. Journal of Applied Probability and Statistics, Vol. 15, pp. 61–84.

Ibrahim, M., Altun, E. and Yousof, H. M., (2020). A new distribution for modeling lifetime data with different methods of estimation and censored regression modeling. Statistics, Optimization & Information Computing, Vol. 8, pp. 610–630.

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

Jahanshahi, S. M. A., Yousof, H. M. and Sharma, V. K., (2019). The Burr X Fréchet Model for Extreme Values: Mathematical Properties, Classical Inference and Bayesian Analysis. Pakistan Journal of Statistics and Operation Research, Vol. 15, pp. 797–818.

Johnson, N. L., Kemp, A. W. and Kotz, S., (2005). Univariate discrete distributions, 3rd edn. Wiley, Hoboken.

Korkmaz, M. C., Altun, E., Yousof, H. M. and HAMEDANI G. G., (2019). The odd power Lindley generator of probability distributions: properties, characterizations and regression modeling, International Journal of Statistics and Probability, Vol. 8, pp. 70–89.

Korkmaz, M. C., Yousof, H. M. and Ali, M. M., (2017). Some theoretical and computational aspects of the odd Lindley Fréchet distribution, Journal of Statisticians: Statistics and Actuarial Sciences, Vol. 2, pp. 129–140.

Korkmaz, M. C., Yousof, H. M. and Hamedani G. G., (2018). The exponential Lindley odd log-logistic G family: properties, characterizations and applications. Journal of Statistical Theory and Applications, Vol. 17, pp. 554–571.

Krishna, E., Jose, K. K., Alice, T. and Risti, M. M., (2013). The Marshall-Olkin Fréchet distribution. Communications in Statistics-Theory and Methods, Vol. 42, pp. 4091–4107

Mahmoud, M. R., Mandouh, R. M., (2013). On the transmuted Fréchet distribution. Journal of Applied Sciences Research, Vol. 9, pp. 5553–5561.

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

Mansour, M. M., Butt, N. S., Ansari, S. I., Yousof, H. M., Ali, M. M. and Ibrahim, M., (2020b). A new exponentiated Weibull distribution's extension: copula, mathematical properties and applications. Contributions to Mathematics, Vol. 1 (2020), pp. 57–66, doi: 10.47443/cm.2020.0018

Mansour, M., Korkmaz, M. Ç., Ali, M. M., Yousof, H. M., Ansari, S. I. and Ibrahim, M., (2020c). A generalization of the exponentiated Weibull model with properties, Copula and application. Eurasian Bulletin of Mathematics, Vol. 3, pp. 84–102.

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

Mansour, M., Yousof, H. M., Shehata, W. A. and Ibrahim, M., (2020e). A new two parameter Burr XII distribution: properties, copula, different estimation methods

and modeling acute bone cancer data. Journal of Nonlinear Science and Applications, Vol.13, pp. 223–238.

Mansour, M. M., Butt, N. S., Yousof, H. M., Ansari, S. I. and Ibrahim, M., (2020f). A Generalization of Reciprocal Exponential Model: Clayton Copula, Statistical Properties and Modeling Skewed and Symmetric Real Data Sets. Pakistan Journal of Statistics and Operation Research, Vol. 16, pp. 373–386.

Merovci, F., Alizadeh, M., Yousof, H. M. and Hamedani G. G., (2017). The exponentiated transmuted-G family of distributions: theory and applications, Communications in Statistics-Theory and Methods, Vol. 46, pp. 10800–10822.

Merovci, F., Yousof, H. M. and Hamedani, G. G., (2020). The Poisson Topp Leone Generator of Distributions for Lifetime Data: Theory, Characterizations and Applications. Pakistan Journal of Statistics and Operation Research, Vol. 16, pp. 343–355.

Nikulin, M. S., (1973a). Chi-square Test for Continuous Distributions with Shift and Scale Parameters. teor. Veroyatn. Primen, Vol. 18**,** pp. 559–568.

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

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

Salah, M., El-Morshedy, M., Eliwa, M. S. and Yousof, H. M., (2020). Expanded Fréchet Model: Mathematical Properties, Copula, Different Estimation Methods, Applications and Validation Testing. Mathematics, Vol. 8, pp.19–49.

Swain, J. Venkatraman, S. and Wilson, J., (1988). Least squares estimation of distribution function in Johnsons translation system, J. Stat. Comput. Simul., Vol. 29, pp. 271–297.

Nascimento, A. D. C., Silva, K. F., Cordeiro, G. M., Alizadeh, M. and Yousof, H. M., (2019). The odd Nadarajah-Haghighi family of distributions: properties and applications. Studia Scientiarum Mathematicarum Hungarica, Vol. 56, pp. 1–26.

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

Yadav, A. S., Goual, H., Alotaibi, R. M. Rezk, H., Ali, M. M. and Yousof, H. M., (2020). Validation of the Topp-Leone-Lomax model via a modified Nikulin-Rao-Robson goodness-of-fit test with different methods of estimation. Symmetry, Vol. 12, pp. 1–26. DOI: 10.3390/sym12010057

Yousof, H. M., Afify, A. Z., Alizadeh, M., Butt, N. S., Hamedani, G. G. and Ali, M. M., (2015). The transmuted exponentiated generalized-G family of distributions. Pakistan Journal of Statistics and Operation Research, Vol. 11, pp. 441–464.

Yousof, H. M., Afify, A. Z., Ebraheim, A. N., Hamedani, G. G. and Butt, N. S., (2016). On six-parameter Fréchet distribution: properties and applications, Pakistan Journal of Statistics and Operation Research, Vol. 12, pp. 281–299.

Yousof, H. M., Alizadeh, M., Jahanshahiand, S. M. A., Ramires, T. G., Ghosh, I. and Hamedani G. G., (2017). The transmuted Topp-Leone G family of distributions: theory, characterizations and applications, Journal of Data Science. Vol. 15, pp. 723–740.

Yousof, H. M., Afify, A. Z., Alizadeh, M., Nadarajah, S., Aryal, G. R. and Hamedani, G. G., (2018a). The Marshall-Olkin generalized-G family of distributions with Applications, STATISTICA, Vol. 78, pp. 273–295.

Yousof, H. M., Altun, E. and Hamedani, G. G., (2018b). A new extension of Frechet distribution with regression models, residual analysis and characterizations. Journal of Data Science, Vol. 16, pp. 743–770.

Yousof, H. M., Afify, A. Z., Alizadeh, M., Hamedani G. G., Jahanshahi, S. M. A. and Ghosh, I., (2018c). The generalized transmuted Poisson-G family of Distributions. Pakistan Journal of Statistics and Operation Research, Vol. 14, pp. 759–779.

Yousof, H. M., Altun, E., Ramires, T. G., Alizadeh, M. and Rasekhi, M., (2018c). A new family of distributions with properties, regression models and applications, Journal of Statistics and Management Systems, Vol. 21, pp. 163–188.

Yousof, H. M., Butt, N. S., Alotaibi, R., Rezk, H. R., Alomani, G. A. and Ibrahim, M., (2019). A new compound Fréchet distribution for modeling breaking stress and strengths data. Pakistan Journal of Statistics and Operation Research, Vol. 15, pp. 1017–1035.

Yousof, H. M., Hamedani, G. G. and Ibrahim, M., (2020). The Two-parameter Xgamma Fréchet Distribution: Characterizations, Copulas, Mathematical Properties and Different Classical Estimation Methods. *Contributions to Mathematics*, Vol. 2 (2020), pp. 32–41.

Yousof, H. M., Rasekhi, M., Altun, E. and Alizadeh, M., (2018d). The extended odd Frechet family of distributions: properties, applications and regression modeling. International Journal of Applied Mathematics and Statistics, Vol. 30, pp. 1–30.

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 123–140, DOI 10.21307/stattrans-2021-030
Received – 18.09.2020; accepted – 09.04.2021

Robust Bayesian insurance premium in a collective risk model with distorted priors under the generalised Bregman loss

Agata Boratyńska[1]

ABSTRACT

The article presents a collective risk model for the insurance claims. The objective is to estimate a premium, which is defined as a functional specified up to unknown parameters. For this purpose, the Bayesian methodology, which combines the prior knowledge about certain unknown parameters with the knowledge in the form of a random sample, has been adopted. The generalised Bregman loss function is considered. In effect, the results can be applied to numerous loss functions, including the square-error, LINEX, weighted square-error, Brown, entropy loss. Some uncertainty about a prior is assumed by a distorted band class of priors. The range of collective and Bayes premiums is calculated and posterior regret Γ-minimax premium as a robust procedure has been implemented. Two examples are provided to illustrate the issues considered - the first one with an unknown parameter of the Poisson distribution, and the second one with unknown parameters of distributions of the number and severity of claims.

Key words: classes of priors, posterior regret, distortion function, Bregman loss, insurance premium

1. Introduction

We consider a Bayesian collective risk model. Our objective is to estimate a premium, which is defined as a functional H that assigns to any risk S a real number $H(S)$, the premium for taking the risk S. In practical situations the premium $H(S)$ can be calculated if the distribution of the risk S is known. We shall consider the case in which the distribution of S or the premium $H(S)$ is specified up to an unknown parameter θ, thus the risk premium will be denoted by $H(\theta)$. The premium $H(\theta)$ can be calculated according to different principles, from the simplest net premium to more sophisticated ones (see Kaas et al. (2009), Furman and Zitikis (2008)). Next we ought to estimate $H(\theta)$. We will use the Bayesian methodology, which combines the prior knowledge about a parameter θ (defined by a prior distribution π) with the knowledge

[1] Warsaw School of Economics SGH, Collegium of Economic Analysis, Institute of Econometrics, Al. Niepodległości 162, 02-554 Warszawa, Poland. E-mail: aborata@sgh.waw.pl. ORCID: 0000-0001-7363-1960.

in the form of a random sample $X = (X_1, X_2, \dots, X_n)$, where the distribution of this random variable depends on θ. The quality of an estimator is measured by the expected value of a loss function. There are a lot of different loss functions considered in the literature (see Heilmann (1989), Gómez-Déniz (2008), Boratyńska (2008) and Karimnezhad and Parsian (2018) for more references). Its choice depends on the severity of the error related to overestimation or underestimation. The most popular square-error loss equally penalizes over- and under-estimation of the same magnitude, the LINEX loss with $c < 0$ gives a greater error for underestimation than for overestimation, under the generalized entropy loss an error depends on the ratio between the estimated function and a considered action (for definitions of losses see Table 1). Again under- and over-estimation are not penalized equally. We will use the generalized Bregman loss (GB loss) function introduced by Karimnezhad and Parsian (2018) (for definition see Section 2). The class of GB loss functions contains different losses (weighted, symmetric, asymmetric, precautionary). All the loss functions mentioned above belong to that class. Thus, a practitioner has the great family of loss functions and he can choose one that expresses the severity of the estimation error very well.

Now, having some prior information about a parameter $\theta \in \Theta$, described by a prior distribution π (we will use the same notation for a probability distribution and its density (p.d.f.) with respect to the chosen measure on a probability space Θ), and a loss function $L(H(\theta), a)$ (measuring an error between the estimated parameter $H(\theta)$ and our estimate a) we can calculate the collective premium $\widehat{H}_\pi^C$, which minimizes the expected loss

$$E_\pi L(H(\theta), a) = \int_\Theta L(H(\theta), a)\pi(d\theta)$$

in a class of actions $a \in R$.

If, additionally, we have a random sample $X = (X_1, X_2, \dots, X_n)$ and X has a p.d.f. depended on a parameter θ, then for every value x of a random variable X we can calculate a Bayes premium $\widehat{H}_\pi^B(x)$, which minimizes the posterior risk equal the expected value of the loss function, if θ has the posterior distribution, thus

$$R_x(\pi, a) = E_\pi(L(H(\theta), a)|x) = \int_\Theta L(H(\theta), a)\pi(d\theta|x),$$

where $\pi(\cdot\,|x)$ denotes the posterior p.d.f. and a denotes a chosen action. Two premiums (defined above) express two situations. For example, the first premium is a premium in a class of risk. The prior expresses the population behaviour of an unknown parameter θ. The second premium combines knowledge about the population and about one considered risk (a policy).

The collective and Bayes premiums depend on a choice of a prior. The elicitation of a prior is difficult and can be uncertain. To model uncertainty of the prior

information the robust Bayesian inference uses a class Γ of priors. In literature there are a lot of different classes Γ of priors: parametric classes of priors, ε-contamination classes, density and distribution band classes, quantile classes. For general references see Berger (1994), Ríos Insua and Ruggeri (2000). In insurance the robust Bayesian analysis was considered in many papers, for example: Young (2000), Chan et al. (2008), Gómez-Déniz (2009), Karimnezhad and Parsian (2014), Boratyńska (2017). Most of them present parametric or ε-contamination classes. We will use a class of priors based on distortion functions defined by Arias-Nicolás et al. (2016) (for definition see Section 3). The class is easily elicited and interpretable. It is connected with the stochastic and likelihood ratio orders. It quantifies a prior uncertainty in terms of distortion of a cumulative distribution function (c.d.f.). A parametric class of priors very often has a fixed shape of a c.d.f. During elicitation of a prior a practitioner has only approximate knowledge about a prior and narrowing down to a certain parametric family may be unjustified. The family considered in the paper can be an alternative. In insurance this class was considered by Sánchez-Sánchez et al. (2019). The concept of distortion functions has been used in actuarial science to model risk measure (see, for example, Balbas et al. (2009)).

Having a class Γ of priors we choose a measure of robustness of a statistical procedure and some concept of optimality. As a measure of robustness the range of posterior quantity, like the Bayes estimator, can be considered. If the range is small, then one may used the Bayes estimator as the robust procedure with respect to misspecifications of the prior (see Berger (1994), Ríos Insua and Ruggeri (2000) and Arias-Nicolás et al. (2016), among others). On the other hand, if conclusions differ widely, we should aim at eliciting additional information about the prior. However, the expert may not be willing to provide more information, and the practitioner is interested in choosing a single action from the set of actions provided by a global procedure. In this moment we can choose several concepts of an optimal procedure: the stable procedure, conditional Γ-minimax procedure or posterior regret Γ-minimax (PRGM) procedure (see Sivaganesan and Berger (1989), Ríos Insua et al. (1995), Boratyńska (1997, 2002), Ríos Insua and Ruggeri (2000), among others). We will use the last concept. Given the imprecision in elicitation of a prior, we try to make a decision, and this decision cannot be a Bayes action for every prior in the class Γ. Thus, we choose an action (in our problem an estimator of a premium), which minimizes the maximum loss of optimality in the class Γ and the largest possible increase in risk, resulting from making the wrong choice of a prior distribution, is kept as small as possible. The PRGM estimator depends on bands of the Bayes estimator when a prior runs over the class Γ. Thus, computing a PRGM estimator is simple provided that we have procedures to compute the range of Bayes estimators.

The article is organized as follows. Section 2 presents a guide for collective, Bayes and PRGM premiums under the GB loss. Section 3 reviews the structure of the class of priors based on distortion functions. Considering the GB loss function we find the bands of the Bayes estimator for a distorted band class of priors, thus we can compute the PRGM estimators. We note that for every value of a random sample X the optimal PRGM premium is the Bayes premium with respect to one prior from the considered class of priors. Section 4 contains PRGM estimators of a premium in some actuarial models with the GB loss function. We present some generalization for the case in which an unknown parameter is bidimensional (it is a case where a parameter of a probability distribution of a number of claims and a parameter of a probability distribution of a severity of claims are unknown and some prior information about them is known). Section 5 contains some concluding remarks.

2. Collective, Bayes and PRGM premiums under the GB loss function

Generally, let X be an observed random variable with a p.d.f. $f(\cdot|\theta)$ indexed by a real unknown parameter θ. Suppose θ has a prior distribution π. Let $L(H(\theta), a)$ be the generalized Bregman loss function (GB loss), measuring the penalty of incorrect estimation of a premium $H(\theta)$ by a real decision action a, defined as follows:

$$L(H(\theta), a) = w(H(\theta))\left[\phi(g(a)) - \phi\left(g(H(\theta))\right) - \left(g(a) - g(H(\theta))\right)\phi'\left(g(H(\theta))\right)\right],$$

where real functions w, g and ϕ are fixed and $w(H(\theta)) > 0$ for every value $H(\theta)$, $g(\cdot)$ is a monotone function and $\phi(\cdot)$ is a convex, differentiable function and $\phi'(g(\theta)) = \frac{d}{dz}\phi(z)|_{z=g(\theta)}$. The shape of the GB loss depends of the choice of functions w, g and ϕ, for example, taking $w(z) = e^{-cz}$, $g(z) = z$ and $\phi(z) = e^{cz}$ $(c \neq 0)$ we obtain the LINEX loss function introduced by Varian (1974), taking $w(z) = 1$, $g(z) = z$ and $\phi(z) = z^2$ we have the square-error loss. Table 1 presents some examples of the GB loss. The following theorem is the corollary of Theorem 3.1. in Karimnezhad and Parsian (2018).

Theorem 1. *Let $X = x$. Then, under the GB loss function and a prior π, the collective $\widehat{H}_\pi^C$ and Bayes $\widehat{H}_\pi^B(x)$ premiums satisfy the following equations:*

$$\phi'(g(\widehat{H}_\pi^C)) = \frac{E_\pi(w(H(\theta))\phi'(g(H(\theta))))}{E_\pi(w(H(\theta)))},$$

$$\phi'(g(\widehat{H}_\pi^B(x))) = \frac{E_\pi(w(H(\theta))\phi'(g(H(\theta)))|x)}{E_\pi(w(H(\theta))|x)}.$$

Now, suppose that our knowledge about a prior is described by a family Γ of priors. Let $r^C(\Gamma)$ and $r^B(\Gamma, x)$ denote the range of a collective and a Bayes premium when priors run the class Γ, respectively, thus if $X = x$, then

$$r^C(\Gamma) = \sup_{\pi\in\Gamma} \widehat{H}_\pi^C - \inf_{\pi\in\Gamma} \widehat{H}_\pi^C \qquad \text{and} \qquad r^B(\Gamma, x) = \sup_{\pi\in\Gamma} \widehat{H}_\pi^B(x) - \inf_{\pi\in\Gamma} \widehat{H}_\pi^B(x).$$

Consider the posterior regret of an action a given by

$$r_x(\pi, a) = R_x(\pi, a) - R_x(\pi, \widehat{H}_\pi^B(x)).$$

In a sense, for $X = x$, it measures the loss of optimality due to choosing a instead of the optimal Bayes estimate. The estimator $\widehat{H}_\Gamma^{PR}$ is the posterior regret Γ-minimax premium (PRGM premium) if for every value x of X

$$\inf_{a\in R} \sup_{\pi\in\Gamma} r_x(\pi, a) = \sup_{\pi\in\Gamma} r_x(\pi, \widehat{H}^{PR}(x)).$$

We will use the following theorem to calculate the PRGM premium.

Theorem 2. (Karimnezhad and Parsian (2018)) *In estimating $H(\theta)$ under the GB loss function, let $X = x$, Γ be a class of prior distributions and let $\underline{H} = \underline{H}(x) = \inf_{\pi\in\Gamma} \widehat{H}_\pi^B(x)$, $\overline{H} = \overline{H}(x) = \sup_{\pi\in\Gamma} \widehat{H}_\pi^B(x)$ and $\underline{H} < \overline{H}$.*
If $w(H) = const$, then

$$g(\widehat{H}^{PR}(x)) = \frac{\phi(g(\overline{H})) - \phi(g(\underline{H})) - \left(g(\overline{H})\phi'(g(\overline{H})) - g(\underline{H})\phi'(g(\underline{H}))\right)}{\phi'(g(\underline{H})) - \phi'(g(\overline{H}))}.$$

If there exists a constant k such that $E_\pi(w(H(\theta))|x) = \frac{k}{\phi'(g(\widehat{H}_\pi^B(x)))}$, then

$$\frac{\phi(g(\widehat{H}^{PR}(x))) - \phi(g(\overline{H})) - \phi'(g(\overline{H}))(g(\widehat{H}^{PR}(x)) - g(\overline{H}))}{\phi(g(\widehat{H}^{PR}(x))) - \phi(g(\underline{H})) - \phi'(g(\underline{H}))(g(\widehat{H}^{PR}(x)) - g(\underline{H}))} = \frac{\phi'(g(\overline{H}))}{\phi'(g(\underline{H}))}.$$

Directly from the proof of Theorem 2 we have the following corollaries.

Corollary 1. *Under the assumptions of Theorem 2 for every x of X*

$$\underline{H}(x) \le \widehat{H}^{PR}(x) \le \overline{H}(x).$$

Corollary 2. *Under the assumptions of Theorem 2, if for every value x of X the set $\{\widehat{H}_\pi^B(x): \pi \in \Gamma\}$ is a connected set, then for every x there exists $\pi \in \Gamma$ such that $\widehat{H}^{PR}(x) = \widehat{H}_\pi^B(x)$.*

Table 1 presents collective, Bayes and PRGM premiums for different loss functions belonging to the class of GB loss functions.

Table 1. Examples of GB loss functions and collective, Bayes and PRGM premiums (for more examples and details see Karimnezhad and Parsian (2018))

$L(H,a)$	$\hat{H}_\pi^C$	$\hat{H}_\pi^B$	$\hat{H}^{PR}$
square-error loss $(H-a)^2$	$E_\pi H$	$E_\pi(H\|x)$	$0.5(\underline{H}+\overline{H})$
LINEX loss $e^{c(a-H)}-c(a-H)-1$	$\frac{-1}{c}\ln E_\pi e^{-cH}$	$\frac{-1}{c}\ln E_\pi(e^{-cH}\|x)$	$\underline{H}+\frac{1}{c}\ln\left(\frac{c(\underline{H}-\overline{H})}{\exp(c(\underline{H}-\overline{H}))-1}\right)$
weighted squared loss (1) $\frac{1}{H}(a-H)^2$	$\left(E_\pi\frac{1}{H}\right)^{-1}$	$\left(E_\pi\frac{1}{H}\|x\right)^{-1}$	$\sqrt{\underline{H}\overline{H}}$
weighted squared loss (2) $\frac{1}{H^2}(a-H)^2$	$\frac{E_\pi\frac{1}{H}}{E_\pi\frac{1}{H^2}}$	$\frac{E_\pi(\frac{1}{H}\|x)}{E_\pi(\frac{1}{H^2}\|x)}$	Th.2 is not applicable
Brown loss $(\ln a-\ln H)^2$	$e^{E_\pi\ln H}$	$e^{E_\pi(\ln H\|x)}$	$\sqrt{\underline{H}\overline{H}}$
precautionary loss $\frac{H}{a}+\frac{a}{H}-2$	$\sqrt{\frac{E_\pi H}{E_\pi\frac{1}{H}}}$	$\sqrt{\frac{E_\pi(H\|x)}{E_\pi(\frac{1}{H}\|x)}}$	Th.2 is not applicable
generalized entropy loss $\left(\frac{a}{H}\right)^q-q\ln\frac{a}{H}-1$	$\left(E_\pi\left(\frac{1}{H^q}\right)\right)^{\frac{-1}{q}}$	$\left(E_\pi\left(\frac{1}{H^q}\|x\right)\right)^{\frac{-1}{q}}$	$\left(\frac{\ln\underline{H}^q-\ln\overline{H}^q}{\overline{H}^{-q}-\underline{H}^{-q}}\right)^{\frac{1}{q}}$

3. Distorted band class of priors

We start with recalling the definition of the stochastic and likelihood ratio orders and a distortion function.

Let π_1 and π_2 be two probability distributions on the space Θ and F_{π_1} and F_{π_2} their cumulative distribution functions. We say that π_1 is smaller than π_2 in the stochastic order (denoted by $\pi_1 \preceq \pi_2$) if and only if for every $t \in R$ we have $F_{\pi_1}(t) \geq F_{\pi_2}(t)$. We say that π_1 is smaller than π_2 in the likelihood ratio order (denoted by $\pi_1 \preceq_{lr} \pi_2$) if and only if the ratio of their densities $\frac{\pi_2(\theta)}{\pi_1(\theta)}$ increases over the union of the supports of π_1 and π_2 (here $a/0$ is taken to be equal to $+\infty$ whenever $a > 0$ and a support of a p.d.f. π is a closure of a set $\{\theta \in \Theta: \pi(\theta) > 0\}$).

Let V and W be two random variables such that $V \sim \pi_1$ and $W \sim \pi_2$. It is well known that

$$\pi_1 \preceq_{lr} \pi_2 \quad \Longrightarrow \quad \pi_1 \preceq \pi_2$$

and

$$\pi_1 \preceq \pi_2 \quad \Leftrightarrow \quad E\psi(V) \leq E\psi(W), \qquad (*)$$

for all increasing functions ψ for which the expectations exist. For more details about stochastic orders see Shaked and Shanthikumar (2007), for the stochastic ordering of posterior distributions, marginal distributions of data and predictive distributions see Męczarski (2015).

Let h: $[0,1] \longrightarrow [0,1]$ be a nondecreasing, continuous function such that $h(0) = 0$ and $h(1) = 1$. Then h is called a distortion function. Let π be a probability distribution on Θ, then a probability distribution π_h, with a c.d.f. of the form $F_{\pi_h} = h(F_\pi)$, is called the distorted distribution with the distortion function h.

Suppose that the prior distribution is not exactly specified and consider the following class of priors.

Definition (Arias-Nicolás et al. (2016)). *Let $\bar{\pi}$ be a specific prior belief. The distorted band class $\Gamma_{\bar{\pi},h_1,h_2}$ associated with $\bar{\pi}$, based on h_1 and h_2, a concave and convex distortion functions, respectively, is defined as*

$$\Gamma_{\bar{\pi},h_1,h_2} = \{\pi\colon\ \bar{\pi}_{h_1} \preceq_{lr} \pi \preceq_{lr} \bar{\pi}_{h_2}\}.$$

The following properties are very useful (for details see Arias-Nicolás et al. (2016)):

- easy elicitation and structure,
- $\Gamma_{\bar{\pi},h_1,h_2} \subseteq \{\pi\colon\ \bar{\pi}_{h_1} \preceq \pi \preceq \bar{\pi}_{h_2}\}$,
- if $\pi_1, \pi_2 \in \Gamma_{\bar{\pi},h_1,h_2}$, then for every $\varepsilon \in [0,1]$ and $\pi_\varepsilon = (1-\varepsilon)\pi_1 + \varepsilon\pi_2$ we have $\pi_\varepsilon \in \Gamma_{\bar{\pi},h_1,h_2}$,
- for every $\pi \in \Gamma_{\bar{\pi},h_1,h_2}$ and every x the posterior distribution satisfies

$$\bar{\pi}_{h_1}(\cdot\,|x) \preceq_{lr} \pi(\cdot\,|x) \preceq_{lr} \bar{\pi}_{h_2}(\cdot\,|x).$$

Example 1. Let $\bar{\pi}$ be a fixed prior on the space Θ. Consider a class

$$\Gamma_1 = \{\pi\colon\ \bar{\pi}_{h_{1,c_1}} \preceq_{lr} \pi \preceq_{lr} \bar{\pi}_{h_{2,c_2}}\},$$

where h_{1,c_1}, h_{2,c_2} are two distortion functions such that

$$h_{1,c_1}(z) = 1-(1-z)^{c_1}, \quad h_{2,c_2}(z) = z^{c_2},$$

$$\bar{\pi}_{h_{1,c_1}}(\theta) = \frac{d}{d\theta}(1-(1-F_{\bar{\pi}}(\theta))^{c_1}), \ \ \bar{\pi}_{h_{2,c_2}}(\theta) = \frac{d}{d\theta}((F_{\bar{\pi}}(\theta))^{c_2}),$$

and $c_1 > 1$, $c_2 > 1$ are fixed numbers. Thus, if c_1 and c_2 are integers, then the bounds distributions are the distributions of the first and the last order statistics. The following properties describe the dependence on parameters c_1 and c_2.

- If $c'_1 > c_1$, then $\bar{\pi}_{h_{1,c'_1}} \preceq_{lr} \bar{\pi}_{h_{1,c_1}}$.
- If $c'_2 > c_2$, then $\bar{\pi}_{h_{2,c_2}} \preceq_{lr} \bar{\pi}_{h_{2,c'_2}}$.
- Similar order is for posterior distributions.
- The Kolmogorov distance (see Arias-Nicolas et al. (2016))

$$dK\left(\bar{\pi}, \bar{\pi}_{h_{1,c_1}}\right) = (c_1-1)c_1^{\frac{-c_1}{c_1-1}}, \quad dK(\bar{\pi}, \bar{\pi}_{h_{2,c_2}}) = (c_2-1)c_2^{\frac{-c_2}{c_2-1}}.$$

We will use that class for elicitation priors in Section 4.

Now, considering the GB loss we would like to find bounds of a set of Bayes estimators of the premium. The following lemma presents the preservation of order of the collective and Bayes premiums computed under the GB loss function when prior distributions are in the likelihood ratio order.

Lemma 1. *Let H be an increasing function of θ. Let π_1 and π_2 be two priors such that $\pi_1 \preceq_{lr} \pi_2$ and $\widehat{H}^C_{\pi_i}$, $\widehat{H}^B_{\pi_i}$ be the collective and Bayes premium under the GB loss and the prior π_i, for $i = 1, 2$. Then for every x of X*

$$\widehat{H}^C_{\pi_1} \le \widehat{H}^C_{\pi_2} \quad and \quad \widehat{H}^B_{\pi_1}(x) \le \widehat{H}^B_{\pi_2}(x).$$

If H is decreasing, then in the above inequalities there is the sign change.

Proof. Assume H is increasing (if H is decreasing, then the proof is similar, only we have opposite inequalities in $(**)$).

Having a probability distribution π and a positive integrable function w, define the probability distribution π^w with the p.d.f. equal $\pi^w(\theta) = \frac{w(H(\theta))\pi(\theta)}{\int_\Theta w(H(\theta))\pi(d\theta)}$. If $\pi_1 \preceq_{lr} \pi_2$, then $\frac{\pi_2^w(\theta)}{\pi_1^w(\theta)} = \frac{\int_\Theta w(H(\theta))\pi_1(d\theta)}{\int_\Theta w(H(\theta))\pi_2(d\theta)} \cdot \frac{\pi_2(\theta)}{\pi_1(\theta)}$ is an increasing function of θ, hence $\pi_1^w \preceq_{lr} \pi_2^w$ and $\pi_1^w \preceq \pi_2^w$.

Note that $\phi'(g(\widehat{H}^C_{\pi_i}))$ (see the formula in Theorem 1) is the expected value of the function $\phi'(g(H(\theta)))$ if θ has the probability distribution π_i^w, $i = 1,2$. Now, applying the property $(*)$ of the stochastic order, if g is increasing, we have

$$\phi'\left(g\left(\widehat{H}^C_{\pi_1}\right)\right) \le \phi'\left(g\left(\widehat{H}^C_{\pi_2}\right)\right) \qquad (**)$$

(if g is decreasing we have opposite inequalities) and obtain the assertion for the collective premium. The proof for the Bayes premium is similar, we only put a posterior distribution $\pi(\cdot\,|x)$ in the place of π.

The following theorem presents the bounds of a set of Bayes estimators and it is a conclusion from Lemma 1.

Theorem 3. *Under the GB loss function and the distorted band class $\Gamma_{\overline{\pi},h_1,h_2}$ of priors, if H is an increasing function of θ and for every $\pi \in \Gamma_{\overline{\pi},h_1,h_2}$ and every x of X there exist $\widehat{H}^C_\pi$ and $\widehat{H}^B_\pi(x)$, then*

$$\inf_{\pi\in\Gamma_{\overline{\pi},h_1,h_2}} \widehat{H}^C_\pi = \widehat{H}^C_{\overline{\pi}_{h_1}}, \qquad \sup_{\pi\in\Gamma_{\overline{\pi},h_1,h_2}} \widehat{H}^C_\pi = \widehat{H}^C_{\overline{\pi}_{h_2}},$$
$$\inf_{\pi\in\Gamma_{\overline{\pi},h_1,h_2}} \widehat{H}^B_\pi = \widehat{H}^B_{\overline{\pi}_{h_1}}, \qquad \sup_{\pi\in\Gamma_{\overline{\pi},h_1,h_2}} \widehat{H}^B_\pi = \widehat{H}^B_{\overline{\pi}_{h_2}}.$$

If H is decreasing, then inf and sup change places.

Having the upper and lower bounds for the set of Bayes premiums and applying Theorem 2, we can calculate the PRGM premium if the class of priors is equal $\Gamma_{\overline{\pi},h_1,h_2}$.

Remarks

1. Arias-Nicolás et al. (2016) define the class of submodular loss functions and obtain the bounds of the set of Bayes actions under priors belonging to $\Gamma_{\overline{\pi},h_1,h_2}$, if a loss function is convex in a and submodular. If $w(\theta) = const$ then the GB loss is

submodular $\left(\frac{\partial^2 L(\theta,a)}{\partial\theta\partial a} = -g'(a)g'(\theta)\varphi''(g(\theta)) \leq 0\right)$, but if $w(\theta) \neq const$, then a GB loss may not have the submodularity property. As an example consider $L(\theta,a) = \frac{1}{\theta^2}(a-\theta)^2$.

2. Applying Remark 8 in Sánchez-Sánchez et al. (2019) and Corollaries 1 and 2 we obtain that for every x there exists $\pi_0 \in \Gamma_{\overline{\pi},h_1,h_2}$ such that $\widehat{H}^{PR}(x)$ is equal to the Bayes estimator with respect to the prior π_0.

Example 2. In that example we present the exact formula for the PRGM estimator for some GB losses and a certain class $\Gamma_{\overline{\pi},h_1,h_2}$ of priors.

Let X be an observed random variable with the negative binomial distribution, $bin^-(r,\theta)$, where $\theta \in (0,1)$ is unknown and $r > 0$ is known, with the p.d.f. given by $f(x|\theta) = \frac{\Gamma(r+x)}{\Gamma(r)x!}\theta^r(1-\theta)^x$, if $x = 0,1,2,\ldots$.

Let $\overline{\pi}$ be a prior of θ with the p.d.f. equal $\overline{\pi}(\theta) = 2\theta$ if $\theta \in (0,1)$. We are interested in estimating a function $H(\theta) = \frac{1-\theta}{\theta}$. Note that $E(X|\theta) = rH(\theta)$. Hence, if X describes the number of claims, then we are interested in estimating the expected value of the number of claims. Consider $h_1(z) = z^{0.75}$ and $h_2(z) = z^2$ and a class $\Gamma_{\overline{\pi},h_1,h_2}$ of priors. Then $F_{\overline{\pi}}(\theta) = \theta^2$, $F_{\overline{\pi}_{h_1}}(\theta) = \theta^{1.5}$, $F_{\overline{\pi}_{h_2}}(\theta) = \theta^4$ for $\theta \in (0,1)$. If $X = x$, then posterior distributions for priors $\overline{\pi}, \overline{\pi}_{h_1}$ and $\overline{\pi}_{h_2}$ are beta distributions $Beta(r+2, x+1)$, $Beta(r+0.5, x+1)$ and $Beta(r+4, x+1)$, where a beta distribution with parameters $\alpha > 0$ and $\beta > 0$, $Beta(\alpha,\beta)$, has the p.d.f. given by $\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, if $\theta \in (0,1)$.

Table 2. Bayes and PRGM estimators and the oscillation $r^B(\Gamma_{\overline{\pi},h_1,h_2}, x)$ under some losses, notation: $A = \frac{\Gamma(r+1.5)}{\Gamma(r+1.5+q)}$ and $B = \frac{\Gamma(r+4)}{\Gamma(r+4+q)}$.

Loss function	$(H-a)^2$	$\frac{1}{H}(H-a)^2$	$\left(\frac{a}{H}\right)^q - q\ln\frac{a}{H} - 1$
$\widehat{H}^B_{\overline{\pi}}(x)$	$\frac{x+1}{r+1}$	$\frac{x}{r+2}$	$\left(\frac{\Gamma(r+2)x!}{\Gamma(r+2+q)\Gamma(x-q+1)}\right)^{\frac{1}{q}}$
$\widehat{H}^B_{\overline{\pi}_{h_1}}(x)$	$\frac{x+1}{r+0.5}$	$\frac{x}{r+1.5}$	$\left(\frac{\Gamma(r+1.5)x!}{\Gamma(r+1.5+q)\Gamma(x-q+1)}\right)^{\frac{1}{q}}$
$\widehat{H}^B_{\overline{\pi}_{h_2}}(x)$	$\frac{x+1}{r+3}$	$\frac{x}{r+4}$	$\left(\frac{\Gamma(r+4)x!}{\Gamma(r+4+q)\Gamma(x-q+1)}\right)^{\frac{1}{q}}$
$r^B(\Gamma_{\overline{\pi},h_1,h_2}, x)$	$\frac{2.5(x+1)}{(r+0.5)(r+3)}$	$\frac{2.5x}{(r+1.5)(r+4)}$	$\left(\frac{x!}{\Gamma(x-q+1)}\right)^{\frac{1}{q}}\left(A^{\frac{1}{q}} - B^{\frac{1}{q}}\right)$
$\widehat{H}^{PR}(x)$	$\frac{(x+1)}{2}\left(\frac{1}{r+0.5} + \frac{1}{r+3}\right)$	$\frac{x}{\sqrt{(r+4)(r+1.5)}}$	$\left(\frac{x!}{\Gamma(x-q+1)} \cdot \frac{\ln(B/A)}{1/A-1/B}\right)^{\frac{1}{q}}$

Now, applying formulas from Table 1 we can calculate Bayes and PRGM estimators under selected loss functions. Table 2 presents results.

In the above example the interesting prior and posterior distributions are easy to compute. In practice, it is not easy to compute the exact distributions and interesting

posterior quantities (for example the expected value). In the next section we apply the acceptance-rejection method. The algorithm applying this method for simulation a random sample from prior and posterior distributions π_h and $\pi_h(\cdot\,|x)$, knowing distributions π and $\pi(\cdot\,|x)$, is presented in Arias-Nicolás et al. (2016).

4. The collective risk models and premium calculations, examples

Let $N, Y_1, Y_2, \ldots$ be independent random variables, where N describes the number of claims and $Y_1, Y_2, \ldots$ are identically distributed random variables describing severity of claims. We consider two models.

4.1. Unknown parameter θ in the Poisson model

Assume that N has the Poisson distribution with an unknown parameter $\theta > 0$ and a distribution of Y_1 is known. The parameter θ can represent a driver's propensity to make a claim and the prior indicates how that propensity is distributed throughout the population of insured drivers (see Lemaire (1979), Gómes-Déniz (2009)). Consider the premium $H(\theta)$ which is a linaer function of θ, thus $H(\theta) = t\theta$ (the net premium, the variance principle premium, the Esscher premium, the exponential premium are examples, see Boratyńska (2008)). Now, let $X_1, X_2, \ldots, X_n$ be observed i.i.d. random variables with the Poisson distribution $Poiss(\theta)$ and consider following GB loss functions (for shape see Figure 1):

- the square-error loss $L_s(a,\theta) = (\theta - a)^2$,
- the LINEX loss $L_l(a,\theta) = e^{-0{,}5(a-\theta)} + 0{,}5(a-\theta) - 1$,
- the Brown loss $L_B(a,\theta) = (\ln\theta - \ln a)^2$,
- the generalized entropy losses with q equal 2, 1 and -1:

$$L_2(a,\theta) = \left(\frac{a}{\theta}\right)^2 - 2\ln\frac{a}{\theta} - 1,\ L_1(a,\theta) = \frac{a}{\theta} - \ln\frac{a}{\theta} - 1, L_{(-1)}(a,\theta) = \frac{\theta}{a} + \ln\frac{a}{\theta} - 1.$$

For all these loss functions it is enough to find the collective, Bayes and PRGM estimators of θ, because if $H(\theta) = t\theta$, then

$$\widehat{H}_\pi^C = t\hat{\theta}_\pi^C, \quad \widehat{H}_\pi^B = t\hat{\theta}_\pi^B, \quad \widehat{H}^{PR} = t\hat{\theta}^{PR},$$

for the square-error, Brown and generalized entropy losses. For the LINEX loss $L_l(a,\theta) = e^{c(a-\theta)} + c(a-\theta) - 1$, with a constant c, we have

$$\widehat{H}_\pi^C = t\hat{\theta}_{\pi,tc}^C\ , \quad \widehat{H}_\pi^B = t\hat{\theta}_{\pi,tc}^B\ , \quad \widehat{H}^{PR} = t\hat{\theta}_{tc}^{PR},$$

where $\hat{\theta}_{\pi,tc}^C$, $\hat{\theta}_{\pi,tc}^B$, $\hat{\theta}_{tc}^{PR}$ are estimators for the LINEX loss with a constant tc (see Boratyńska (2008)).

Note that the collective and the Bayes estimator of θ for loss functions L_s and $L_{(-1)}$ are equal, but PRGM estimators are different (see Table 1).

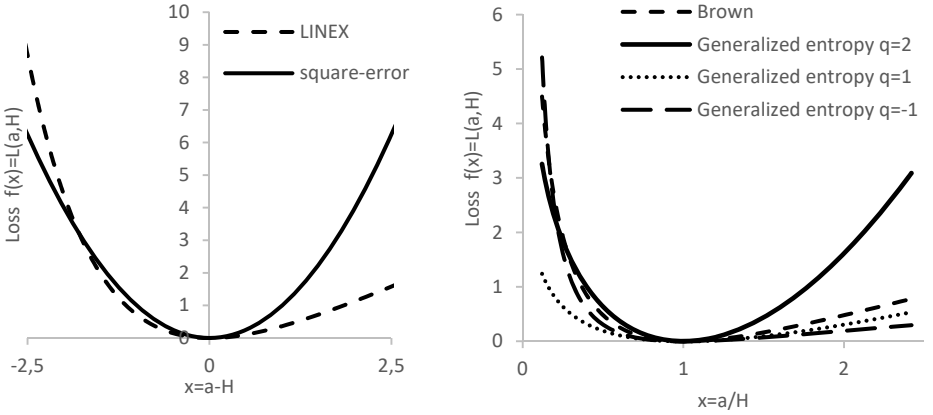


Figure 1. The graphs of loss functions

We assume that the actuary is unable to specify a simple prior distribution of the expected number of claims. Thus, let $\bar{\pi} = Gamma(3, 15)$ be the fixed prior distribution of θ with a p.d.f. $\bar{\pi}(\theta) = \frac{15^3}{2}\theta^2 \exp(-15\theta)$ for $\theta > 0$, and

$$\Gamma = \{\pi\colon\ \bar{\pi}_{h_1} \preceq_{lr} \pi \preceq_{lr} \bar{\pi}_{h_2}\}$$

be the family of priors, where

$$\bar{\pi}_{h_1}(\theta) = \frac{d}{d\theta}(1 - (1 - F_{\bar{\pi}}(\theta))^{c_1}), \qquad \bar{\pi}_{h_2}(\theta) = \frac{d}{d\theta}((F_{\bar{\pi}}(\theta))^{c_2})$$

and $c_1 = c_2 = 1.5$. Then $dK(\bar{\pi}, \bar{\pi}_{h_1}) = dK(\bar{\pi}, \bar{\pi}_{h_2}) = 0.148$. The class Γ expresses the inaccuracy in determining the cumulative distribution function of $\bar{\pi}$. The parameters c_1, c_2 provides the degree of distortion and can be elicited by fixing a reasonable distance in terms of Kolmogorov metric.

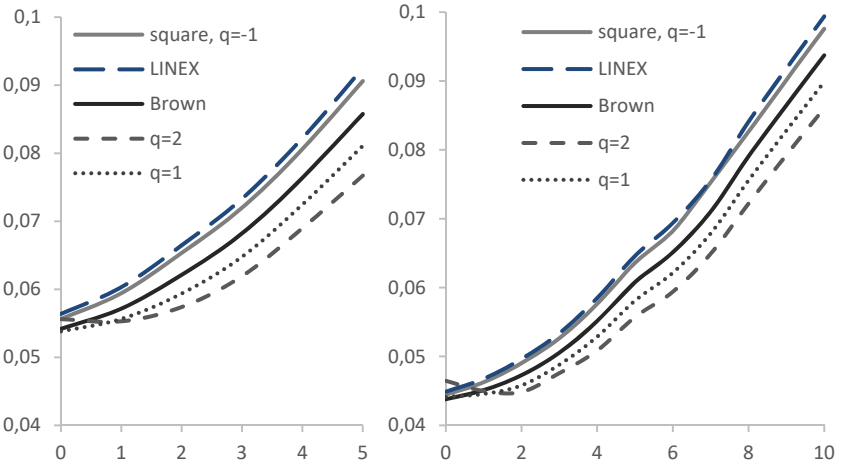


Figure 2. Oscillation $r^B(\Gamma, x)$ of Bayes estimators for different loss functions, $n = 5$ (left) and $n = 10$ (right)

Figure 2 presents the oscillation of the Bayes estimators for $n = 5$ and $n = 10$ and different values of $x = \sum_{i=1}^{n} X_i$. Table 3 shows the oscillation of the collective estimator for different losses. We see that the oscillation for Bayes estimators is an increasing function of x (except the generalized entropy loss with $q = 2$) and it is smaller than the oscillation for the collective estimators for $\frac{x}{n} < 0.5$. The greatest oscillation is for the LINEX loss.

Table 1. Oscillation of the collective estimator of θ

Loss	square	LINEX	Brown	Generalized entropy loss		
				$q = 2$	$q = 1$	$q = -1$
$r^C(\Gamma)$	0.076	0.078	0.073	0.071	0.071	0.076

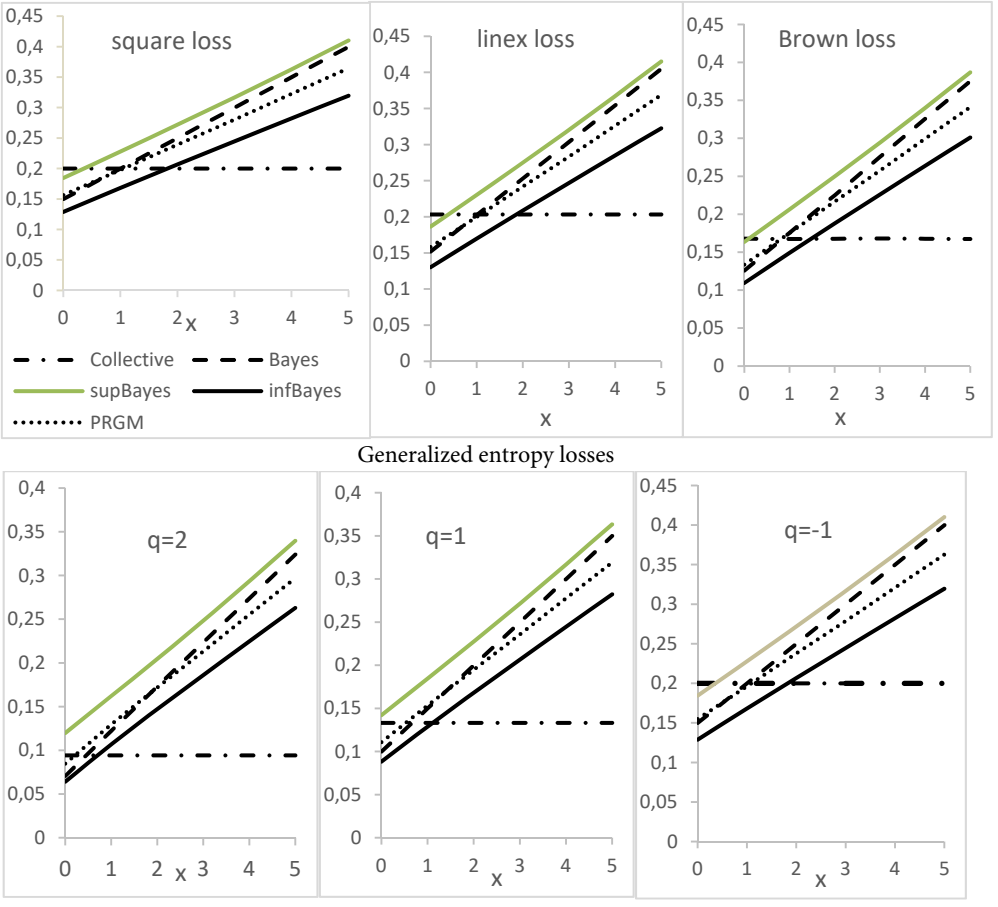


Figure 3. Values of collective, Bayes and PRGM estimators and minimum and maximum of Bayes estimators for different loss functions and $n = 5$

Figures 3 and 4 show values of minimum and maximum of Bayes estimators, collective estimators and Bayes estimators for the prior $\bar{\pi} = Gamma(3,15)$ and PRGM estimators for two values of n and different $x = \sum_{i=1}^{n} X_i$. The oscillation of Bayes estimators is the smallest if $\frac{x}{n}$ is closed to the expected value $E_{\bar{\pi}}\theta = 0.2$.

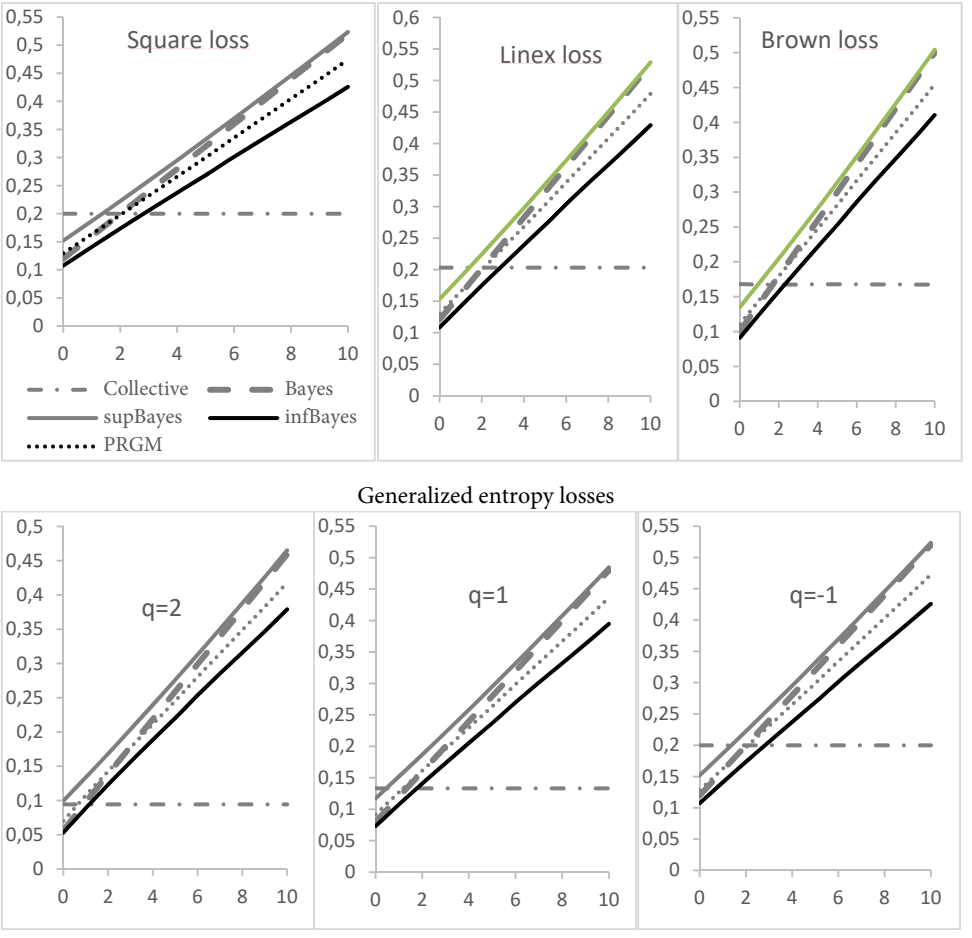


Figure 4. Values of collective, Bayes and PRGM estimators and minimum and maximum of Bayes estimators for different loss functions and $n = 10$

We use n and x small, because n is interpreted as the number of periods (years) we observe, for example, a driver, and x is the number of claims during the n periods. The prior represents the population behaviour of the parameter θ. Our results (Bayesian and PRGM premiums) have similar interpretation as the rules in the credibility theory. They combine knowledge about a single driver with knowledge about the entire population. Similar models with a parametric class of priors or an ε-contamination class of priors and the square-error loss or LINEX loss were considered in Boratyńska (2008) and Gómez-Déniz (2009).

4.2. Unknown parameters θ and λ of distributions of the number and severity of claims

Assume that random variable N has a distribution $f_1(\cdot\,|\theta)$ depending on an unknown parameter $\theta \in \Theta$, and a random variable Y_1 has a distribution $f_2(\cdot\,|\lambda)$ depending on an unknown parameter $\lambda \in \Lambda$. Consider the premium of the form

$$H(\theta, \lambda) = H_1(\theta)H_2(\lambda),$$

where H_1 and H_2 are increasing and continuous functions of θ and λ, respectively.

Let $X_1, X_2, \dots, X_n$ be observed i.i.d. random variables with a p.d.f. $f_1(\cdot\,|\theta)$ and $Z_1, Z_2, \dots, Z_m$ be observed i.i.d. random variables with a p.d.f. $f_2(\cdot\,|\lambda)$, all variables are conditionally independent, knowing parameters θ and λ. Assume that θ and λ are independent, and $\theta \sim \mu$ and $\lambda \sim \upsilon$. Denote $X = (X_1, X_2, \dots, X_n)$ and $Z = (Z_1, Z_2, \dots, Z_m)$. Let x and z be observed values of random variables X and Z. It can be seen directly that the posterior distributions $\mu(\cdot\,|x)$ and $\upsilon(\cdot\,|z)$ are independent. Consider the following GB loss functions: square-error loss, Brown loss, generalized entropy loss (see Table 1). Then, the collective and Bayes premiums are equal

$$\widehat{H}^C_{\mu,\upsilon} = \widehat{H}^C_{1,\mu}\widehat{H}^C_{2,\upsilon}, \quad \widehat{H}^B_{\mu,\upsilon}(x,z) = \widehat{H}^B_{1,\mu}(x)\widehat{H}^B_{2,\upsilon}(z).$$

Let Γ^* be a family of priors on the space $\Theta \times \Lambda$ with a p.d.f. given by

$$\pi(\lambda, \theta) = \mu(\theta)\upsilon(\lambda),$$

where

$$\mu \in \{\mu\colon\ \bar{\mu}_{h_1} \preceq_{lr} \mu \preceq_{lr} \bar{\mu}_{h_2}\},\ \upsilon \in \{\upsilon\colon\ \bar{\upsilon}_{h_3} \preceq_{lr} \upsilon \preceq_{lr} \bar{\upsilon}_{h_4}\},$$

$\bar{\mu}$ and $\bar{\upsilon}$ are fixed priors on the spaces Θ and Λ, respectively, and h_1, h_2, h_3, h_4 are fixed distortion functions (h_1, h_3 are concave and h_2, h_4 are convex). Assume that for every $\pi \in \Gamma^*$ and every x and z the Bayes premium exists. Then (applying Theorem 4) the minimum and maximum of Bayes estimators of the premium H are given by

$$\inf_{\pi\in\Gamma^*}\widehat{H}^B_\pi(x,z) = \widehat{H}^B_{1,\bar{\mu}_{h_1}}(x)\widehat{H}^B_{2,\bar{\upsilon}_{h_3}}(z), \qquad \sup_{\pi\in\Gamma^*}\widehat{H}^B_\pi(x,z) = \widehat{H}^B_{1,\bar{\mu}_{h_2}}(x)\widehat{H}^B_{2,\bar{\upsilon}_{h_4}}(z),$$

and using Theorem 2 we have the PRGM estimator of H.

Example 3. Assume that $N \sim Poiss(\theta)$ and Y_1 has an exponential distribution with a density given by $f_2(y|\lambda) = \frac{1}{\lambda}\exp(-\frac{y}{\lambda})$ for $y > 0$, depended on an unknown parameter $\lambda > 0$. Consider the net premium

$$H(\theta, \lambda) = H_1(\theta)H_2(\lambda) = \theta\lambda.$$

Assume that θ has the prior distribution $\bar{\mu} = Gamma(\alpha, \beta)$ and λ has the prior distribution $\bar{\upsilon} = IGamma(a, b)$ with a density function

$$\bar{\upsilon}(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{-a-1}\exp\left(-\frac{b}{\lambda}\right) \text{ for } \lambda > 0\,,$$

where α, β, a,b are fixed positive parameters and $\alpha > 2$ and $a > 1$.

If $X = x = (x_1, x_2, \dots, x_n)$ and $Z = z = (z_1, z_2, \dots, z_m)$, then the posterior distributions are $\bar{\mu}(\cdot\,|x) = Gamma(\alpha + \sum_{i=1}^{n} x_i, \beta + n)$ and $\bar{v}(\cdot\,|z) = IGamma(a + m, b + \sum_{i=1}^{m} z_i)$. We obtain the following collective and Bayes premiums:

- under the square-error loss and the generalized entropy loss for $q = -1$

$$\widehat{H}^{C}_{\bar{\mu},\bar{v}} = \frac{\alpha b}{\beta(a-1)}, \qquad \widehat{H}^{B}_{\bar{\mu},\bar{v}}(x,z) = \frac{(\alpha + \sum_{i=1}^{n} x_i)(b + \sum_{i=1}^{m} z_i)}{(\beta + n)(a + m - 1)},$$

- under the Brown loss

$$\widehat{H}^{C}_{\bar{\mu},\bar{v}} = \exp(\psi(\alpha,\beta) - \psi(a,b)),$$

$$\widehat{H}^{B}_{\bar{\mu},\bar{v}}(x,z) = \exp\left(\psi(\alpha + n, \beta + \sum_{i=1}^{n} x_i) - \psi(a + m, b + \sum_{i=1}^{m} z_i)\right),$$

where $\psi(s,t) = \int_0^{+\infty} \ln y \frac{t^s}{\Gamma(s)} y^{s-1} e^{-ty} dy$,

- under the generalized entropy loss for $q = 1$

$$\widehat{H}^{C}_{\bar{\mu},\bar{v}} = \frac{(\alpha - 1)b}{\beta a}, \qquad \widehat{H}^{B}_{\bar{\mu},\bar{v}}(x,z) = \frac{(\alpha + \sum_{i=1}^{n} x_i - 1)(b + \sum_{i=1}^{m} z_i)}{(\beta + n)(a + m)},$$

- under the generalized entropy loss for $q = 2$

$$\widehat{H}^{C}_{\bar{\mu},\bar{v}} = \frac{b\sqrt{(\alpha-2)(\alpha-1)}}{\beta\sqrt{a(a+1)}}, \quad \widehat{H}^{B}_{\bar{\mu},\bar{v}}(x,z) = \frac{\sqrt{(\alpha+\sum_{i=1}^{n} x_i - 2)(\alpha+\sum_{i=1}^{n} x_i - 1)}(b+\sum_{i=1}^{m} z_i)}{(\beta+n)\sqrt{(a+m+1)(a+m)}}.$$

For $i = 1,3$ and $j = 2,4$ define fixed numbers $c_i > 1$, $c_j > 1$ and the distortion functions

$$h_i(z) = 1 - (1 - z)^{c_i}, \qquad h_j(z) = z^{c_j}.$$

Now, using the class Γ^* of priors and simmulation methods for calculation of posterior expected values (similarly as in Section 4.1), we obtain the minimum and maximum of collective and Bayes premiums and the PRGM premium.

5. Conclusions

The analysis proposed in this article was used to provide the optimal estimators of the risk premium in Bayesian models with the distorted band class of priors expressing some uncertainty in elicitation of a prior. It is an alternative to the parametric classes of priors used by practitioners. It expresses the uncertainty in determining a prior c.d.f., and that uncertainty is more realistic. The range of Bayes estimators and optimal PRGM estimators is obtained under the large family of GB loss functions, thus the practitioner can find the loss function expressing the severity of under- and over-estimation.

The numerical example presents the difference among the estimators of frequency of claims in the collective risk model under different loss functions. The last example presents the situation where we can apply results for one-dimensional parameter to the bidimensional parameter, thus we can estimate the net premium with unknown frequency and expected severity of claims. Ruggeri et al. (2021) consider the generalization of the distorted band class to the multivariate case. Applying their models we can try to describe a dependence structure between random variables θ and λ connected with frequency and severity of claims. The author believes that this topic could be expanded in the future.

References

Arias-Nicolás, J. P., Ruggeri, F., Suárez-Llorens, A., (2016). New Classes of Priors Based on Stochastic Orders and Distortion Functions, *Bayesian Analysis,* Vol. 11, pp. 1107–1136.

Balbas A., Garrido J., Mayoral S., (2009). Properties of distortion risk measures, *Methodology and Computing in Applied Probability* 11, p. 385, https://doi.org/10.1007/s11009-008-9089-z.

Berger, J. O., (1994). An overview of robust Bayesian analysis, *Test,* Vol. 3, pp. 5–124 (with discussion).

Boratyńska, A., (1997). Stability of Bayesian inference in exponential families, *Statistics & Probability Letters,* Vol. 36, pp. 173–178.

Boratyńska, A., (2002). Robust Bayesian estimation with asymmetric loss function, *Applicationes Mathematicae,* Vol. 29, pp. 297–306.

Boratyńska, A., (2008). Posterior regret Γ-minimax estimation of insurance premium in collective risk model, *ASTIN Bull.,* Vol. 38, pp. 277–291.

Boratyńska, A., (2017). Robust Bayesian estimation and prediction of reserves in exponential model with quadratic variance function, *Insurance Math. Econom.,* Vol. 76, pp. 135–140.

Chan, J., Choy, B., Makov, U., (2008). Robust Bayesian analysis of loss reserves data using the generalized t-distribution, *ASTIN Bull.,* Vol. 38, pp. 207–230.

Furman, E., Zitikis, R., (2008). Weighted premium calculation principles, *Insurance Math. Econom.*, Vol. 42, pp. 459–465.

Gómez-Déniz, E., (2008). A generalization of the credibility theory obtained by using the weighted balanced loss function, *Insurance Math. Econom.,* Vol. 42, pp. 850–854.

Gómez-Déniz, E., (2009). Some Bayesian Credibility Premiums Obtained by Using Posterior Regret Γ-Minimax Methodology, *Bayesian Analysis,* Vol. 4, pp. 223–242.

Heilmann, W., (1989). Decision theoretic foundations of credibility theory, *Insurance Math. Econom.,* Vol. 8, pp. 77–95.

Kaas, R., Goovaerts, M., Dhaene, J., Denuit, M., (2009). *Modern actuarial risk theory*, Springer-Verlag, Berlin Heidelberg.

Karimnezhad, A., Parsian, A., (2014). Robust Bayesian methodology with applications in credibility premium derivation and future claim size prediction, *AstA Advances in Statistical Analysis,* Vol. 98, pp. 287–303.

Karimnezhad, A., Parsian, A., (2018). Bayesian and robust Bayesian analysis in a general setting, *Commun in Statist. Theory and Methods,* Vol. 48, pp. 3899–3920.

Lemaire, J., (1979). How to define a bonus-malus system with an exponential utility function, *ASTIN Bull.,* Vol. 10, pp. 274–282.

Męczarski, M., (2015). Stochastic orders in the Bayesian framework, *Annals of the Collegium of Economic Analysis,* Vol. 37, pp. 339–360.

Ríos Insua, D., Ruggeri, F. (eds.), (2000). Robust Bayesian analysis, *Lecture Notes in Statistics,* Vol. 152, Springer-Verlag, New York.

Ríos Insua, D., Ruggeri, F., Vidakovic, B., (1995). Some results on posterior regret Γ-minimax estimation, *Statistics and Decisions,* Vol. 13, pp. 315–331.

Ruggeri, F., Sánchez-Sánchez, M., Sordo, M.A., Suárez-Llorens, A., (2021). On a New Class of Multivariate Prior Distributions: Theory and Application in Reliability, *Bayesian Analysis,* Vol. 16, pp. 31–60.

Sánchez-Sánchez, M., Sordo, M. A., Suárez-Llorens, A., Gómez-Déniz, E., (2019). Deriving robust Bayesian premiums under bands of prior distributions with applications, *ASTIN Bull.*, Vol. 49, pp. 147–168.

Shaked, M., Shanthikumar, J. G., (2007). *Stochastic Orders*, Springer, New York, USA.

Sivaganesan, S., Berger, J. O., (1989). Ranges of posterior measures for priors with unimodal contaminations, *Annals of Statist.,* Vol. 17, pp. 868–889.

Varian, H. R., (1974). A Bayesian approach to real estate assessment, *Studies in Bayesian Econometrics and Statistics,* North Holland, pp. 195–208.

Young, V., (2000). Credibility using semiparametric models and a loss function with a constancy penalty, *Insurance Math. Econom.,* Vol. 26, pp. 151–156.

STATISTICS IN TRANSITION new series, September 2021
Vol. 23, No. 3, pp. 141–156, DOI 10.21307/stattrans-2021-031
Received – 21.11.2019; accepted – 03.03.2021

An application of persistent homology and the graph theory to linguistics: The case of Tifinagh and Phoenician scripts

Hajar Bouazzaoui[1], **Mohamed Abdou Elomary**[2], **MyIsmail Mamouni**[3]

ABSTRACT

As the origin of the Tifinagh script remains uncertain, this work aims at exploring its probable relatedness with the Phoenician script. Using tools from within topological data analysis and graph theory, the similarity between the two scripts is studied. The clustering of their letter shapes is performed based on the pairwise distances between their topological signatures. The ideas presented in this work can be extended to study the similarity between any two writing systems and as such can serve as the first step for linguists to determine the possibly related scripts before conducting further analysis.

Key words: topological data analysis, persistent homology, graph theory, writing systems, Abjad scripts, Alphabet scripts, Tifinagh script, Phoenician script.

1. Introduction

Living beings - humans and animals alike, have a need for systems of communication to ensure their survival. Humans, by their ingenuity, have developed writing systems as a conventional visual mode to represent their oral communication. While writing and talking are both tools for transmitting messages, writing has the advantage of being a reliable form of data storage that obeys the usual coding and decoding rules, which imply a shared understanding by the author and the reader of the sets of characters composing the used writing system.

Tifinagh, which is the writing system of interest in this paper, is the script adopted for Tamazight or Berber languages more broadly. Berber has been originally spoken in territories ranging from the Atlantic coast to Egypt before the arabisation of North Africa. Millions of Tifinagh inscriptions of various styles and eras tattoo the rocks of North Africa and the Sahara. A long process of cultural and identity changes begun with the emergence of Islam in the seventh century, concurrently, the linguistic map of Tifinagh (see Figure 1) retracted over the centuries until its present form, broken into islands distant from each other.

[1]Hassan I University, Department of Mathematics and Computer Science FST de Settat, IMII Laboratory. Address: Km 3, B.P.: 577 Route de Casablanca, Morocco. E-mail: h.bouazzaoui@uhp.ac.ma, ORCID: https://orcid.org/0000-0003-1860-9757

[2]Hassan I University, Department of Mathematics and Computer Science FST de Settat, IMII Laboratory. Address: Km 3, B.P.: 577 Route de Casablanca, Morocco. E-mail: elomaryabdou@gmail.com

[3]Department of Mathematics, Research Team of Mathematics, Didactic and its Applications (M@DA), CRMEF RABAT, Avenue Allal Al Fassi, Madinat Al Irfane, 10000, Rabat, Morocco. E-mail: mamouni.myismail@gmail.com

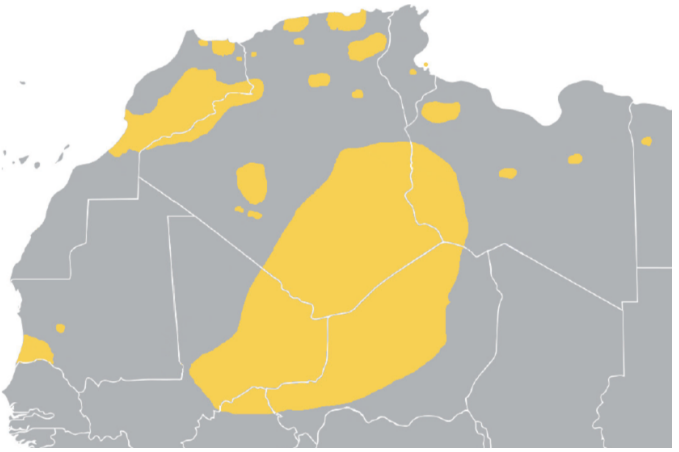


Figure 1: Current Tifinagh speaking map in Africa.

So far, there is no conclusive theory about the origin of the Tifinagh script. The majority of scholars support one these three theories (Blanco 2014):

- South-Semitic origin (Arabian and Latin scripts);
- North-Semitic origin (Phoenician and/or Punic);
- Independent invention with Phoenician influence.

Our aim in the present work is to verify whether the Tifinagh and the Phoenician scripts are indeed related.

From a linguistic point of view, the study of script evolution is not independent from historical, geographic and cultural factors. One cannot then demonstrate the relationship between scripts based solely on the study of individual graphemes (Briquel-Chatonnet 1997). However, analyzing and comparing letter shapes remains an important constituent of that study.

In order to demonstrate linguistic relatedness and to reconstruct a hypothetical common ancestral system of languages, linguists rely, among others, on the comparative method as a technique to study language development and perform comparisons on these languages (McMahon, A. and McMahon, R. 2011). However, the languages to compare are not chosen at random, and an initial stage of deciding whether some languages are related is required.

The present work, which studies the relatedness of the Phoenician and Tifinagh scripts, rely on methods that could be extended to study the relatedness of any two other scripts, and as such, serve as a first step to the comparative method, at least to the extent where only letter shapes are considered.

We believe that this is the first work that investigates the visual relationship between scripts using topological data analysis (TDA). A previous work (Sadouk et al. 2020) established a possible relationship between Phoenician and Tifinagh scripts using deep learning. The authors trained a classifier on a dataset of Phoenician letters and used a transfer learning system based on these shapes to improve the performance of Tifinagh handwritten character

recognition thereby inferring a possible relationship between the two scripts. Still, as with all deep learning systems, large samples of data were required. TDA, on the other hand, can provide robust results with only small samples of data.

To verify the relatedness of the two scripts, we adopt a topological data analysis approach based on persistent homology and graph theory. We represent each letter of the writing systems we are studying as a graph. Our aim is to study the similarity between the graphs corresponding to Phoenician letters and those corresponding to Tifinagh letters. In the literature, many graph similarity measures were studied among which we cite maximum common subgraph (Fernández and Valiente 2001), the number of mismatching edges (Zhu et al. 2012) and graph edit distance (GED) (Gouda and Hassan 2016). GED has been the most adopted one. It is the least expensive sequence of edit operations that can transform a graph G_1 to a graph G_2. In practice, however, finding the minimal edit distance is an NP-hard problem and has the drawback of having an exponential computational complexity in terms of the number of graph edit vertices.

In this work, topological information of interest in each of these graphs is summarized in persitence diagrams. Computing the *Bottleneck distance* between these topological signatures will serve as a mean to verify similarity between letter graphs and thus between Tifinagh and Phoenician scripts.

The paper is organized as follows: in Section 2, we give a brief introduction of mathematical concepts we will be using throughout this paper; we put special emphasis on persistent homology. In Section 3, we describe the method we used to perform our analysis before closing with a discussion of results and future research directions.

2. Materials and background

Homology formalizes the way topological spaces are distinguished by examining their holes. One of the most common approaches to homology is simplicial homology. It is based on associating abelian groups or modules to simplicial complexes built on top of topological spaces. One of its major advantages is that it lends itsef to relatively easy computations.

We first define what simplices are. A simplex or a p-simplex is the generalization of a triangle in p-dimension.

Definition 1 (p-simplex)
Let $e_0, e_1, ..., e_p$ be affinely independent points in $\mathbb{R}^n$. The associated convex hull, denoted $\sigma^p = [e_0, e_1, ..., e_p]$, is called a p-simplex. That is the polyhedron:

$$\sigma^p = \left\{ \sum_{i=0}^{p} t_i e_i,\ t_i \geq 0, \sum_{i=0}^{p} t_i = 1 \right\}$$

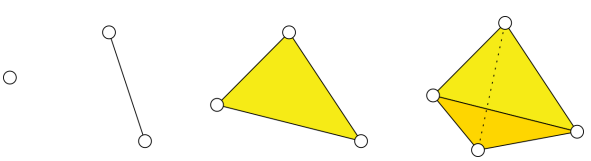


Figure 2: A 0-simplex is a point, a 1-simplex is a line segment, a 2-simplex is a triangle and a 3-simplex a tetrahedron (Zhu 2013)

When σ and α are two simplices such that $\alpha \subset \sigma$, we call α a face of σ and σ a co-face of α.

Definition 2 (Simplicial complex)
A simplicial complex K is a finite collection of simplices satisfying the following conditions:

1. *For any $\sigma \in K$ with a face α, we have $\alpha \in K$;*
2. *If $\sigma_1, \sigma_2 \in K$ then $\sigma_1 \cap \sigma_2 =$ or $\sigma_1 \cap \sigma_2 \in K$.*

The dimension of K is the maximal dimension of its simplices.

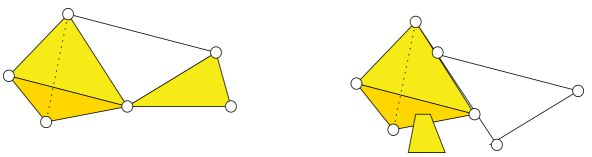


Figure 3: Left: a simplicial complex. Right: not a simplicial complex (Zhu 2013)

Definition 3 (p-chain)
A p-chain is a formal finite sum $\sum_i n_i \sigma_i^p$, where σ_i^p are oriented p-simplices of a simplicial-complex K and $n_i \in \mathbb{Z}$.

The set $C_p(K)$ of all p-chains of K is a $\mathbb{Z}$-module. The following $\mathbb{Z}$-linear map :

$$\partial_p : C_p(K) \to C_{p-1}(K) \tag{1}$$

is called a *boundary map*, it is defined at the level of the generators as follows:

$$\partial_p(\sigma) := \sum_{i=0}^{p} (-1)^i [e_0, e_1, ..., \hat{e_i}, ..., e_p] \tag{2}$$

where $\sigma = [e_0, e_1, \cdots, e_p]$ is an oriented $p-$simplex and $\hat{e_i}$ means that e_i is omitted. Thus, the boundary of a tetrahedron is the alternative sum of its four triangles, the boundary of a triangle is the alternative sum of its three edges and the boundary of a line segment is the difference of its two endpoints. A direct computation shows that

$$\partial_p \circ \partial_{p+1} = 0. \tag{3}$$

In other words

$$Im\partial_{p+1} \subset \ker \partial_p. \tag{4}$$

This yields the following exact sequence, called a *chain complex* of K :

$$0 = C_{n+1}(K) \xhookrightarrow{i} C_n(K) \xrightarrow{\partial_n} C_{n-1}(K) \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} C_{-1}(K) = 0 \tag{5}$$

where $\hookrightarrow$ denotes the inclusion map. The figure below illustrates the evolution of this chain complex.

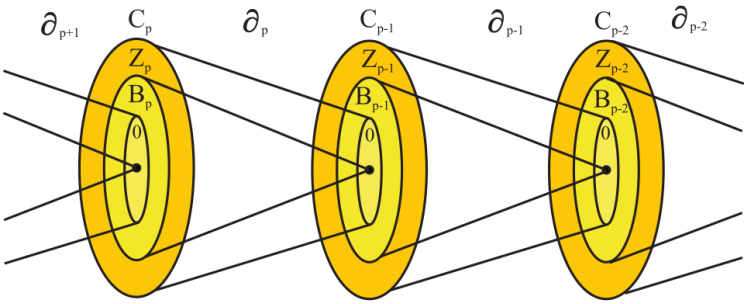


Figure 4: Chain, cycle and boundary groups and their mappings under boundary operators. (Horak, Maletić and Rajković 2009)

Elements of $Z_p := \ker \partial_p$ are called *p-cycles*, those of $B_p := Im\partial_{p+1}$ are called p-boundaries. In particular, any p-boundary is a cycle, but the inverse does not always hold. The obstruction for a cycle to be a boundary is encoded in the quotient

$$H_p(K) := \frac{Z_p}{B_p}. \tag{6}$$

called the p-th homology group of K. Its rank, defined as

$$\beta_p(K) := \dim_Z H_p(K), \tag{7}$$

is called the *p-th Betti number* of K and it encodes the number of $p-$dimensional holes in the simplicial complex K. In particular, β_0 denotes the number of connected components of K. For more details, we refer the reader to these standard references (Hatcher 2002) and (Spannier 1966).

2.1. Persistent homology

Persistent homology, one of the main tools in topological data analysis, proved its usefulness in many real world applications among which shape analysis, medical imaging and network sensing are only a few examples. In many of these applications, data is given as a point cloud. Persistent homology keeps track of homology classes as a nested sequence of simplicial complexes is built on top of the data. The "lifetime" of a homology class is an indication of the relevance or irrelevance of homological information.

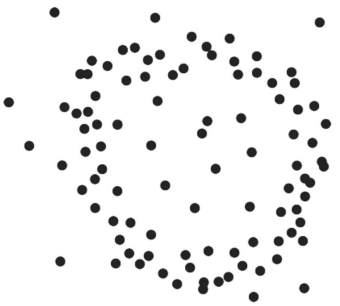


Figure 5: A noisy point cloud data

Let $\mathscr{P}$ be a cloud of points embedded in $\mathbb{R}^n$. One may associate a *filtration* to $\mathscr{P}$, that is a finite increasing sequence of sub-complexes

$$\mathscr{P} = K_0 \subset K_1 \subset \cdots \subset K_n. \tag{8}$$

For every $i \leq j$, the inclusion map $K_i \hookrightarrow K_j$ induces the homology homomorphism

$$f_p^{i,j} : H_p(K_i) \longrightarrow H_p(K_j) \tag{9}$$

at each dimension p. This yields the homology sequence

$$H_p(K_0) \longrightarrow H_p(K_1) \cdots \longrightarrow H_p(K_n). \tag{10}$$

As we go from K_{i-1} to K_i, we gain new homology classes and lose others as they become trivial or merge with each other. Persistent homology groups are defined as follows.

Definition 4 *The p-th persistent homology groups, denoted $H_p^{i,j}$, are defined to be the images $H_p^{i,j} := Imf_p^{i,j}$. Their ranks $\beta_p^{i,j} := rank(H_p^{i,j})$, are the corresponding p-th persistent Betti numbers.*

We note that

$$H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i)). \tag{11}$$

A class γ is born at time $t = i$ if $\gamma \notin H_p^{i-1,i}$. It dies at time $t = j$, when it becomes trivial or when it merges with an older class as we go from K_{j-1} to K_j, that is, $f_p^{i,j-1}(\gamma) \notin H_p^{i-1,j-1}$ but $f_p^{i,j}(\gamma) \in H_p^{i-1,j}$. The figure below illustrates this scenario.

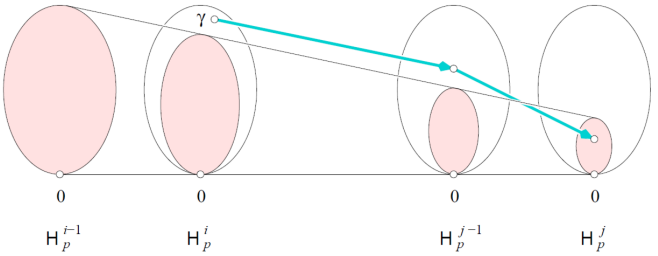


Figure 6: Example of a homology class with birth time $t = i$, and death time $t = j$. (Edelsbrunner and Harer 2010)

We can encode this evolution in a persistence barcode, which is a set of intervals whose first endpoint indicates the birth-time of the homology class, while the second one indicates its death-time. Short line segments correspond to noise, while persistent line segments imply relevant homological information.

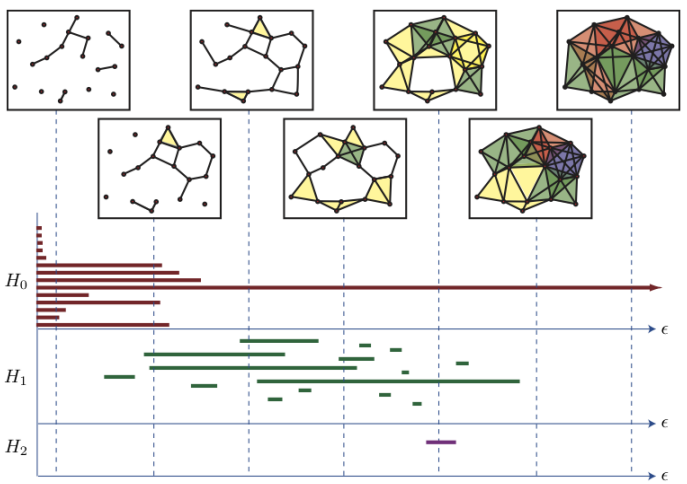


Figure 7: Example of a point cloud and its associated Vietoris-Rips complex and barcode (Ghrist 2008)

Barcodes can be computed efficiently by using a matrix reduction algorithm. Surprisingly, we can get all this information with a single reduction. We order the time appearance $t(\sigma_i)$ of a simplex σ_i as follows: $t(\sigma_i) < t(\sigma_j)$ whenever σ_i is a face of σ_j. Then we set the *boundary matrix*, ∂, which stores all that information, that is the binary matrix,

$$\partial[i,j] := \begin{cases} 1 & \text{if } \sigma_i \text{ is a face of } \sigma_j \text{ of co-dimension one ;} \\ 0 & \text{otherwise.} \end{cases}$$

Let $low(j)$ be the row index of the lowest non-null coefficient (when it exists) in the column j. A matrix is called reduced if $low(j) \neq low(k)$ whenever $j \neq k$. In other words, no two columns have lows in the same level. One way to get a reduced matrix R from the

boundary matrix ∂ is to add columns from left to right. (see Algorithm 1).

Algorithm 1 : Smith Reduction Algorithm

Input: Boundary matrix
Output: Reduced boundary matrix
for $j = 1$ to n **do**
  while $\exists j' < j$ with $low(j') = low(j)$ **do**
    add column j' to column j
  end while
end for

Theorem 1 (Pairing theorem, see (Edelsbrunner and Harer 2010))
Let R be the reduced matrix obtained from the boundary matrix. There is a persistence pairing (i, j) *of a homology class whenever* $i = low(j)$.

The filtrations built on top of data can also be described topologically using persistence diagrams. These are multisets of $\mathbb{R}^2$ that encode information about homology groups. A homology class that appears at time i and disappears at time j is represented by the point of coordinates (i, j). The multiplicity of that point represents the number of features with the same birth and death times. The persistence of each class is the real value $j - i$.

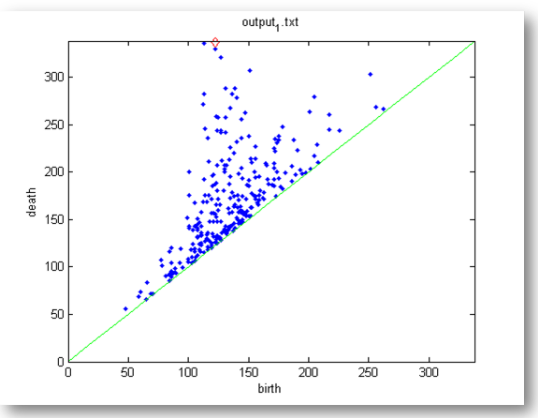


Figure 8: Example of a persistence diagram (Nanda 2017)

In order to compare topological signatures present in the resulting persistence diagrams, we compute their Bottleneck distance.

Definition 5 *Bottleneck distance*
Given two persistence diagrams D and E, their Bottleneck distance (w_∞) *is defined by:*

$$w_\infty(D,E) := \inf_{\eta} \sup_{x\in D} \|x-\eta(x)\|_\infty$$

where η ranges over bijections between D and E.

For further details on persistent homology, we refer the reader to these standard references (Edelsbrunner and Harer 2010) and (Ghrist 2008).

2.2. Zigzag Persistent homology

A more general approach to persistent homology is zigzag persistent homology, which we will use in this paper, in this section, we introduce some of its key principles. In this setting, both forward and backward maps are permitted between topological spaces. Let $\mathbb{X}_1 \leftrightarrow \mathbb{X}_2 \leftrightarrow \cdots \leftrightarrow \mathbb{X}_n$ be a sequence of topological spaces. The maps between these spaces induce maps between chain complexes $C(\mathbb{X}_1) \leftrightarrow C(\mathbb{X}_2) \leftrightarrow \cdots \leftrightarrow C(\mathbb{X}_n)$. The homology sequence $H_p(\mathbb{X}_1) \leftrightarrow H_p(\mathbb{X}_2) \leftrightarrow \cdots \leftrightarrow H_p(\mathbb{X}_n)$ obtained by applying the homology functor H_p forms a zigzag module.

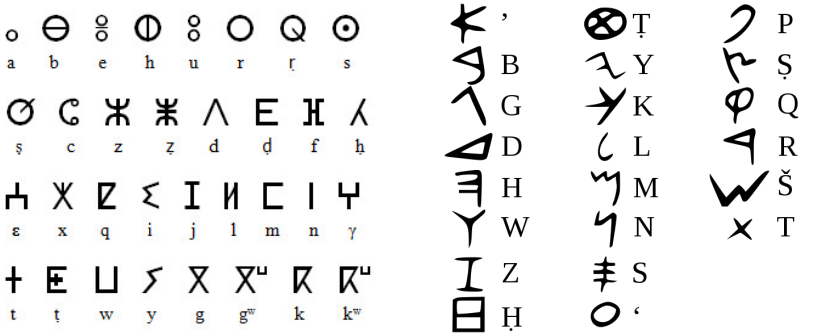
A finite-dimensional zigzag module can be decomposed as a direct sum of interval modules $\bigoplus \mathbb{I}_{[b,d]}$, where $\mathbb{I}_{[b,d]}$ is the homology class existing in the spaces from $H(\mathbb{X}_b)$ to $H(\mathbb{X}_d)$. The information needed to compute this decomposition is encoded in one filtration; the right filtration (Gunnar, De Silva and Morozov 2009).

Definition 6 ((Milosavljević, Morozov and Skraba (2011)) *The right-filtration of a space $H(\mathbb{X}_i)$ is the collection of its subspaces $R^n = (R_0, R_1, \ldots, R_n)$ such that $R_i \subseteq R_j$ whenever $i \leq j$. The filtration is defined as follows.*
If $n = 1$, then $R^0 = (0, H(\mathbb{X}_1))$.
If $H(\mathbb{X}_i) \xrightarrow{f} H(\mathbb{X}_{i+1})$, then $R^{n+1} = (f(R_0), f(R_1), \ldots, f(R_n), H(\mathbb{X}_{i+1}))$.
If $H(\mathbb{X}_i) \xleftarrow{g} H(\mathbb{X}_{i+1})$, then $R^{n+1} = (0, g^{-1}(R_0), g^{-1}(R_1), \ldots, g^{-1}(R_n))$.

For more details on zigzag persistent homology, we refer the interested reader to (Carlsson and De Silva 2010).

3. Method and results

Some scripts undergo a process of transformation over a long period of time, while others are a result of deliberate mixing of traits adopted from multiple other scripts. In this work, we assume that the Tifinagh and Phoenician scripts are related if, by introducing a sequence of minimal random transformations on the Phoenician letters, we obtain clusters of similar letters each containing letters from both scripts. We denote the set of Phoenician letters by $\mathbb{P}$ and that of Tifinagh letters by $\mathbb{T}$. To account for the dynamics of the script letters in $\mathbb{P}$, each of these letters is represented as a dynamic graph by allowing operations such as adding or removing vertices and edges. A dynamic graph is a graph $G = \{G_1, G_2, \ldots, G_n\}$ on which a sequence of updates is performed, G_i being the modified graph at time i. In this study,



((**a**)) Tifinagh letters ((**b**)) Phoenician letters

Figure 9: Tifinagh and Phoenician letters

we only allow graph transformations within 1 Graph Edit Distance (GED) from the original ones.

Zigzag persistent homology, which is a generalization of persistent homology, is more adapted for studying dynamical graphs. It allows for both inserting and deleting edges and vertices which correspond respectively to adding and removing simplices from the constructed simplicial complex.

A clustering is finally performed based on the pairwise bottleneck distances between topological signatures in order to detect similar letters in the two sets $\mathbb{P}$ and $\mathbb{T}$.

We summarize our approach in the following steps:

1. Represent each letter in $\mathbb{P}$ as a time-varying graph $G = \{G_0, G_1, \ldots, G_n\}$, G_i being the letter graph at step i.

2. Associate a metric space representation to each graph G and build a dynamical simplicial complex.

3. Compute the zigzag persistent homology of G.

4. Make a clustering of $\mathbb{P}$ and $\mathbb{T}$ graphs on the basis of the pairwise bottleneck distances between their topological signatures in order to detect similarity between letters.

3.1. Dynamic graphs

The dynamic graphs are constructed as follows:

- The letter is represented by a finite undirected graph. The vertex set V is the set of the intersection points present in that letter.

- A vertex function f_v that maps the state of the vertex v at time t to its state at time $t+1$.

- An edge function f_e that maps the state of the edge e at time t to its state at time $t+1$.

The undirected graph is represented by an adjacency matrix $\mathscr{M}$. The dynamics correspond to adding and removing nodes and edges and updating the matrix accordingly. For an edge addition (resp. deletion) event between vertices v_i and v_j, the function f_e assigns 1 (resp. 0) to $\mathscr{M}_{i,j}$ and $\mathscr{M}_{j,i}$. The function f_v on the other hand, adds a new row and column to the matrix $\mathscr{M}$ whenever a new edge is added and deletes the row and column corresponding to a vertex when it is removed.

3.2. Metric space representation

To each graph G_i, we associate a metric space representation. A metric on G_i is obtained by computing a matrix of shortest path distances between nodes using the Floyd Warshall algorithm (Floyd 1962) which we implemented using the Networkx library (Hagberg, Swart, and S Chult 2008). The Floyd Warshall algorithm is a dynamic programming algorithm, which works in the following fashion: let d_{ij}^k be the shortest path from i to j with intermediate vertices chosen among $\{1,2,\ldots,k\}$. Then, for $k > 1$, $d_{ij}^k = min(d_{ij}^{k-1}, d_{ik}^{k-1} + d_{kj}^{k-1})$.

We then build the Vietoris Rips complexes of the graphs on top of these metric spaces using Dionysus library (Morozov 2012). Given a distance matrix, we compute a sorted filtration filled with the 1-skeleton of the clique complex built on the points at distance at most six from each other, six being the maximum scale at which the Vietoris-Rips complex is computed.

3.3. Computing Zigzag Persistent Homology

In practice, given a time-varying graph $G = \{G_0, G_1, \ldots, G_n\}$, we start by constructing a simplicial complex of G_0, the graph instance at time $t = 0$. This simplicial complex is dynamically modified; as we add new vertices/edges or remove them, simplices are added or removed. At time $t = 0$, a simplicial complex K_0 is created. At time $t > 0$, K_t is the simplicial complex associated with the updated graph at time t. In the case of an addition event, a $k-$simplex is added to K_t if it was not present in K_{t-1}, while the simplices that were at time $t-1$ and did not appear at time t are removed from the complex in the case of a deletion event. We then compute zigzag persistence of this dynamic simplicial complex using *Dionysus* (Morozov 2012).

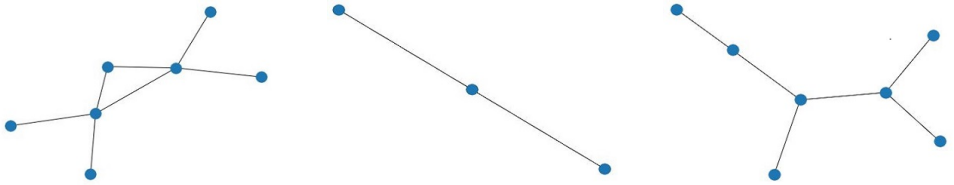


Figure 10: Sample Phoenician Letters Graphs

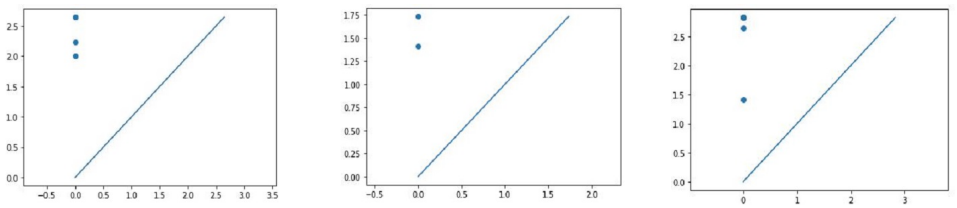


Figure 11: Their Persistence Diagrams

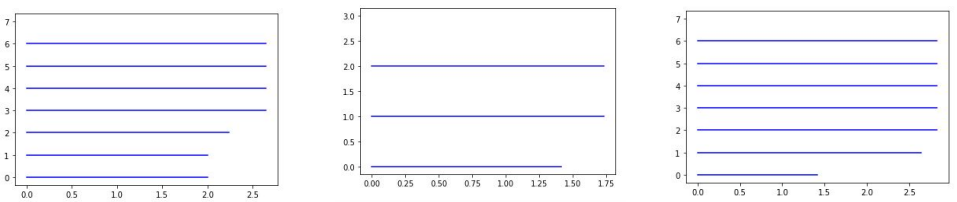


Figure 12: Their Barcodes

Figure 13: Sample Phoenician Letters

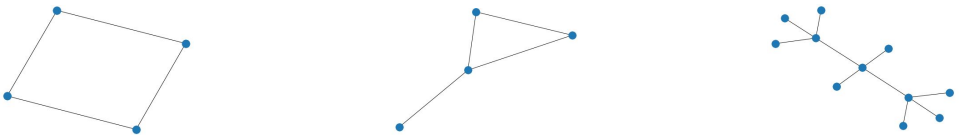


Figure 14: Sample Tifinagh Letters Graphs

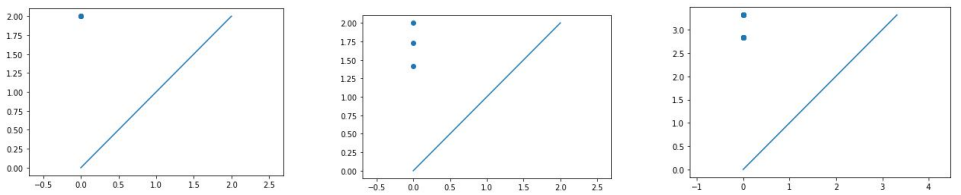


Figure 15: Their Persistence Diagrams

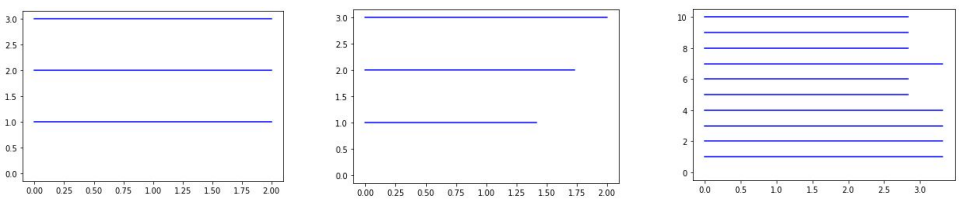


Figure 16: Their Barcodes

Figure 17: Sample Tifinagh Letters

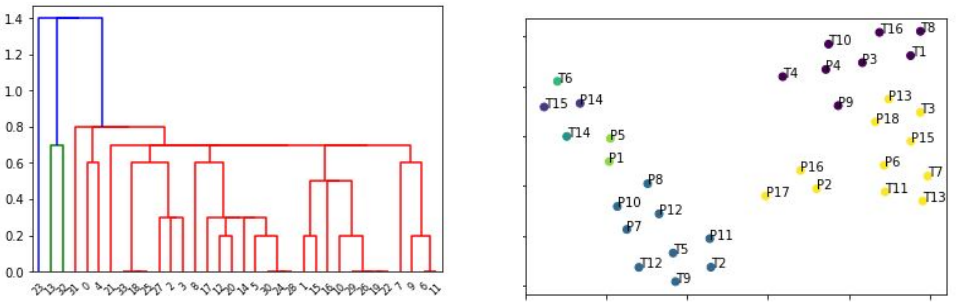
3.4. Clustering

In order to verify the aforementioned claim, i.e. Tifinagh being related to Phoenician, we measure similarity between the time-varying graphs representing the $\mathbb{P}$ letters and those representing the $\mathbb{T}$ letters. After computing the persistence diagrams associated with the simplicial complexes built on top of each graph, we compute the pairwise bottleneck distance between persistence diagrams. We obtain a distance matrix on the basis of which we perform hierarchical clustering, more specifically in this case an agglomerative clustering. Agglomerative clustering starts by considering each singleton as a cluster. The clusters are then inductively combined until some stop criterion is satisfied. In this work, the update at each step is performed using a complete linkage which measures inter-cluster dissimilarity based on the maximum distances between all data points.

3.5. Results

We notice that, except for a few distinct points, each cluster in the right figure contains both Phoenician and Tifinagh letters suggesting similarity between the two.

P_n denotes Phoenician letters while T_n denotes the Tifinagh letters. The agglomerative



((a)) Dendrogram **((b))** Clustering results

Figure 18: Dendrogram and Clustering Results

clustering with a complete linkage we used separates data into seven clusters, each containing both Tifinagh and Phoenician letters suggesting similarity between the two. A cluster also contains letters from the same script. This is due to the homogeneity present within each script; letters of the same script tend to have a common pattern that distinguishes them from other scripts. Hierarchical clustering produces a graphical representation between data points in the form of a hierarchical tree (Figure (A)) that we used for finding the optimal number of clusters.

The results we obtained only suggest a possible visual relationship between the graphemes of the script. This relationship can be due to Tifinagh being derived from Phoenician or Tifiniagh being created under the influence of Phoenician. The nature of that relationship might be a question for a future work.

4. Conclusion

In this work, we demonstrated how TDA and persistent homology in particular can be used to verify the relatedness between two writing systems. Even though we restricted our analysis to the study of similarity between the Phoenician and Tifiniagh scripts, the method we used can be extended to compare any two writing systems. A future work might explore the nature of this relatedness, i.e. whether one script is derived from the other or one was built under the other's influence.

References

Blanco, J., (2014) . Tifinagh & the the IRCAM, Explorations in Cursiveness and Bicamelarism in the Tifinagh script.*Master Of Arts In Typeface Design, University Of Reading*.

Briquel-Chatonnet, F., (1997). De l'araméen à l'arabe: quelques réflexions sur la genèse de l'écriture arabe. *Paris: Bibliothèque Nationale De France*.

McMahon, A., McMahon, R., (2011). Language Classification by Numbers. *Oxford: Oxford University Press*.

Sadouk, L. et al., (2020). Handwritten Phoenician Character Recognition and its Use to Improve Recognition of Handwritten Alphabets with Lack of Annotated Data. *International Journal Of Advanced Trends In Computer Science And Engineering*.

Fernández, M., Valiente, G, (2001). A graph distance metric combining maximum common subgraph and minimum common supergraph. *Pattern Recognition Letters*, 22, pp. 753–758.

Zhu, G., Lin, X., Zhu, K., Zhang, W., Xu Yu, J., (2012). TreeSpan: efficiently computing similarity all-matching. *Proceedings Of The 2012 ACM SIGMOD International Conference On Management Of DataMay*, pp. 529–540.

Gouda, K., Hassan, M., (2016). CSI_GED: An Efficient Approach for Graph Edit Similarity Computation. *Proc. Of ICDE'16*.

Hatcher, A., (2002). *Algebraic Topology*, Cambridge University Press.

Spanier, E., (1966). *Algebraic Topology*, McGraw-Hill Inc.

Edelsbrunner, H., Harer, J., (2010). *Computational Topology. An Introduction*, Amer. Math. Soc., Providence, Rhode Island.

Ghrist, R., (2008). Barcodes: the persistent topology of data.*Bulletin Of The American Mathematical Society*, 45, pp. 61–75.

Gunnar, C., De Silva, V., Morozov, D., (2009). Zigzag persistent homology and real-valued functions.*Proceedings Of The Twenty-fifth Annual Symposium On Computational Geometry (SCG '09)*, pp. 247–256.

Milosavljević, N., Morozov, D., Skraba, P., (2011). Zigzag persistent homology in matrix multiplication time. *Proceedings Of The Twenty-seventh Annual Symposium On Computational Geometry (SoCG '11)*. Association For Computing Machinery, New York, NY, USA, pp. 216–225.

Carlsson, G., De Silva, V., (2010). Zigzag persistence. *Found Comput Math 10*, pp. 367–405.

Floyd, R., (1962). Algorithm 97: Shortest Path. *Communications Of The ACM*, 5, p.345.

Hagberg, A., Swart, P., S Chult, D., (2008). Exploring network structure, dynamics, and function using NetworkX. *Proceedings Of The 7th Python In Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, And Jarrod Millman (Eds), Pasadena, CA USA*, 5, pp. 11–15.

Morozov, D., (2012). Dionysus. http://www.mrzv.org/software/dionysus/

Coulmas, F., (2008). Typology of Writing Systems. *Band 2*, pp. 1380–1387. Available at: `https://doi.org/10.1515/9783110147445.2.9.1380`

Gelb, I., (1963). *A Study of Writing*, Chicago University Press, 2nd edition.

Hill, A., (1967). The typology of writing systems. *Papers In Linguistics*, pp. 92–99.

Horak, D., Maletić, S., Rajković, M., (2009). Persistent homology of complex networks. *J. Stat. Mech. Theory And Experiment*, pp. 30–34.

Pulgram, E., (1976). *The typologies of writing-systems*, Mont Follick Series.

Pichler, W., (2007). The origin of the Libyco-Berber script. *Actes Du Colloque International, Le Libyco-berbére Ou Le Tifinagh: De L'authenticité À L'usage Pratique*, pp. 187–200.

Sampson, G., (1985).*Writing Systems: A Linguistic Introduction*, Hutchinson & Co. Ltd, London.

Slaouti Taklit, M. (2004). *L'alphabet Latin serait-il d'origine berbére?*, Harmattan, Paris.

Unger, J., DeFrancis, J. (1995). Logographic and Semasiographic Writing Systems: A Critique of Sampson's Classification. *Scripts And Literacy. Neuropsychology And Cognition*, 7, pp. 44–58.

Nanda, V., (2017). Perseus: The Persistent Homology Software. Available at: `http://people.maths.ox.ac.uk/nanda/perseus/`

Zhu, X., (2013). Persistent homology: An introduction and a new text representation for natural language processing. *Proceedings Of The Twenty-Third International Joint Conference On Artificial Intelligence*.

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 157–174, DOI 10.21307/stattrans-2021-032
Received – 24.02.2020; accepted – 07.10.2020

A new count data model applied in the analysis of vaccine adverse events and insurance claims

Showkat Ahmad Dar[1], **Anwar Hassan**[2], **Peer Bilal Ahmad**[3], **Sameer Ahmad Wani**[4]

ABSTRACT

The article presents a new probability distribution, created by compounding the Poisson distribution with the weighted exponential distribution. Important mathematical and statistical properties of the distribution have been derived and discussed. The paper describes the proposed model's parameter estimation, performed by means of the maximum likelihood method. Finally, real data sets are analyzed to verify the suitability of the proposed distribution in modeling count data sets representing vaccine adverse events and insurance claims.

Key words: poisson distribution, weighted exponential distribution, compound distribution, count data, maximum likelihood estimation.

1. Introduction

Compounding a discrete distribution with a continuous distribution is a valuable method for creating flexible distributions to assist modelling of count data. Count data distributions play a key role in several applications for applied fields and theoretical research like health, transport, insurance and engineering, etc. Barreto-Souza and Bakouch (2013) obtained a new class of compound distribution with decreasing failure rate by compounding zero-truncated Poisson Lindley distribution and exponential distribution. Hajebi et al. (2013) obtained a new lifetime model by compounding exponential distribution with negative binomial distribution. Mohmoudi and Jafari (2014) introduced a new lifetime compound probability distribution which generalizes the linear failure rate of distribution. Ghitany et al. (2011) obtained weighted Lindley

[1] Department of Statistics, University of Kashmir, Srinagar (J&K), India. Corresponding Author. E-mail: darshowkat2429@gmail.com. ORCID: https://orcid.org/0000-0002-3661-822X.
[2] Department of Statistics, University of Kashmir, Srinagar (J&K), India.
[3] Department of Mathematical Sciences, Islamic University of Science & Technology, Awantipora, Pulwama (J&K), India.
[4] Department of Statistics, University of Kashmir, Srinagar (J&K), India.

distribution and pointed that Lindley distribution is valuable in exhibiting biological data from mortality studies. Asgharzadeh et al. (2014) created a new class of distribution by mixing any continuous distribution and Poisson Lindley distribution through a compounding technique. Chesneau et al. (2020) introduced Cosine geometric distribution for count data modelling. Bourguigon et al. (2014) obtained the Birnbaum-Saunders power series distribution. The new lifetime distribution has a decreasing, increasing or constant hazard rate. Silva and Cordeiro (2015) created a new lifetime distribution by mixing Burr XII and power series distribution through a compounding technique. Pinho et al. (2015) obtained a new distribution by assuming that simple size distribution as Harris distribution. Bardbar and Nematollahi (2016) obtained a modified exponential distribution-geometric distribution with increasing or decreasing failure rate. Flores et al. (2013) obtained the complementary exponential power series distribution by considering the distribution of vectors through maximum components.

In this paper, we propose a new compounding distribution by compounding the Poisson distribution with the weighted exponential distribution, as there is a need to find a more flexible model for analysing statistical data. This model has over-dispersed nature so it will become most appropriate for analysing over-dispersed count data sets. This property makes this model unique as compared to other compounding models already in the statistical literature.

2. Definition of the proposed model (Poisson weighted exponential distribution)

If $Z|\nu \sim \mathrm{P}(\nu)$, where ν is itself a random variable following weighted exponential distribution with parameter (μ,σ), then determining the distribution that results from marginalizing over ν will be known as a compound of the Poisson distribution with that of weighted exponential distribution, which is denoted by $PWED(Z;\mu,\sigma)$. It may be noted that the proposed model will be a discrete one since the parent distribution is discrete.

Theorem 2.1: The probability mass function of a Poisson weighted exponential distribution, i.e. $PWED(Z;\mu,\sigma)$ is given by

$$P(Z=z)=\frac{\mu^2}{\mu+\sigma}\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}}\right];\ z=0,1,2,3,...,;\ \mu>0,\sigma\geq 0\ .$$

Proof: Using the definition (2), the pmf of a PWED $(Z;\mu,\sigma)$ can be obtained as

$$g(z\,|\,\nu)=\frac{e^{-\nu}\nu^{z}}{(z)!}\quad ;\ z=0,1,2,3,...,;\nu>0\,.$$

When its parameter ν follows weighted exponential distribution (WED) with pdf

$$h(\nu;\mu,\sigma)=\frac{\mu^2}{\mu+\sigma}(1+\sigma\nu)e^{-\mu\nu}\,;\nu>0,\mu>0,\sigma\geq 0\,.$$

We have

$$P(Z=z)=\int_0^\infty g(z\,|\,\nu).h(\nu;\mu,\sigma)d\nu$$

$$P(Z=z)=\frac{\mu^2}{\mu+\sigma}\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}}\right];z=0,1,2,3,...,;\mu>0,\sigma\geq 0 \tag{2.1}$$

which is the p.m.f. of PWED.

The corresponding c.d.f of PWED is obtained as:

$$F_Z(z)=\sum_{n=0}^{z}\left[\frac{\mu^2}{\mu+\sigma}\left[\frac{(1+\mu)+\sigma(n+1)}{(1+\mu)^{n+2}}\right]\right]$$

$$1-\frac{\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma x}{(1+\mu)^{z+2}(\mu+\sigma)}\;;z>0,\mu>0,\sigma\geq 0 \tag{2.2}$$

3. Special Cases

Case 1: If we put $\sigma=0$ the PWED reduces to the Poisson exponential distribution with pmf as

$$P_1(Z=z)=\left[\frac{\mu}{(1+\mu)^{z+2}}\right].$$

Case 2: If we put $\sigma=1$ the PWED reduces to the Poisson size biased exponential distribution with pmf as

$$P_2(Z=z)=\frac{\mu^2}{\mu+1}\left[\frac{(1+\mu)+(z+1)}{(1+\mu)^{z+2}}\right].$$

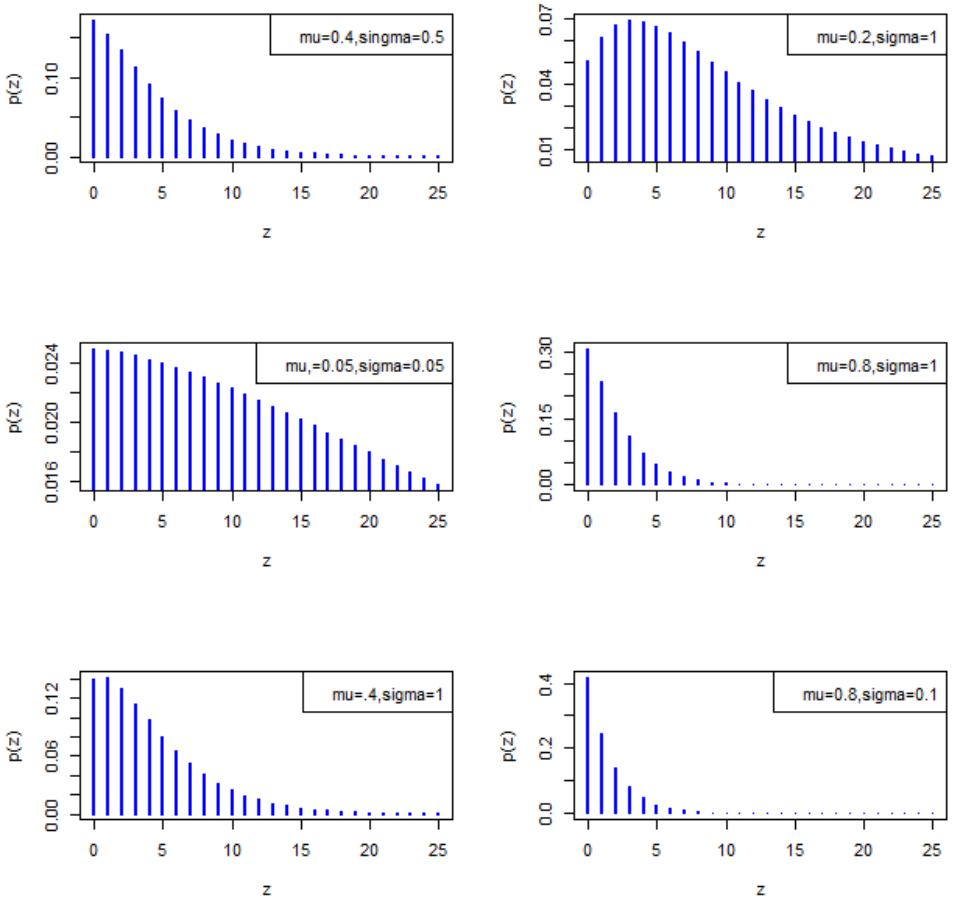


Figure 1. The above figures show the pmf plot for different values of μ and σ

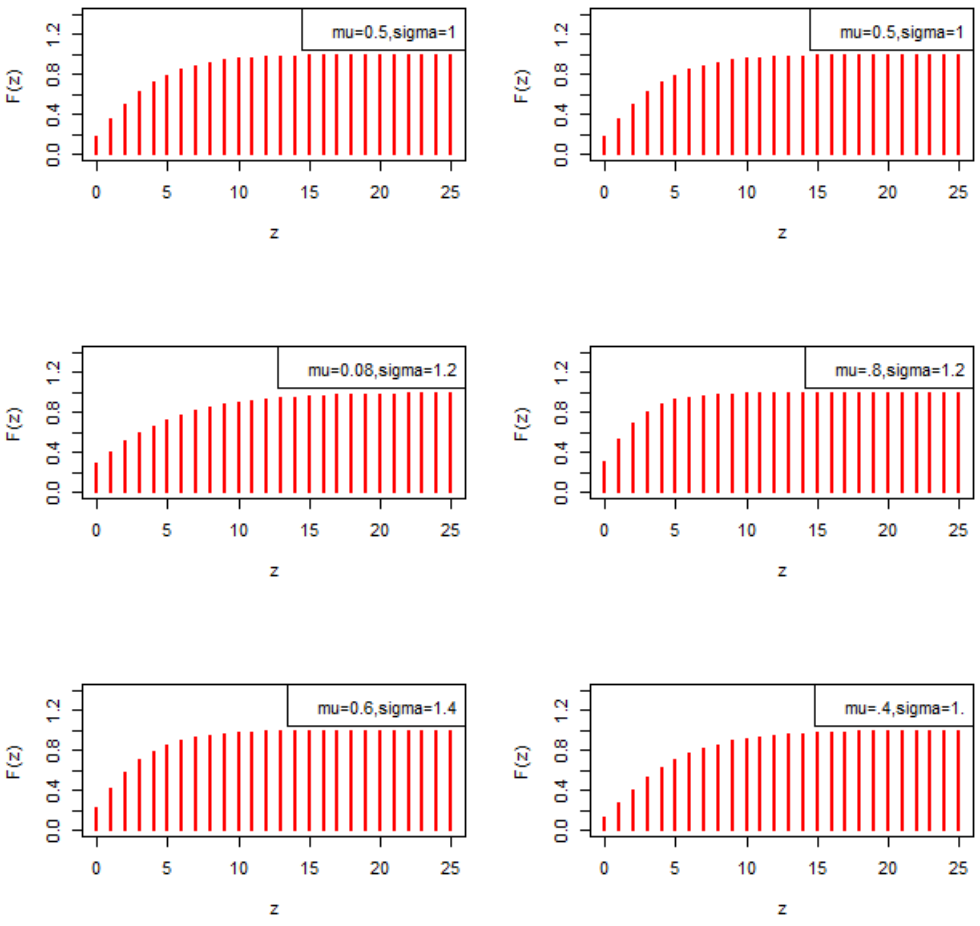


Figure 2. The above figures show the cdf plot for different values of μ and σ

4. Collective risk model

Theorem 4.1: Let Z follow PWED (μ,σ), be a primary distribution with exponential distribution (ζ) as a secondary distribution, then the aggregate loss U= $\sum_{i=0}^{M} Z_i$ has p.d.f given as

$$f_u(z) = \frac{\mu^2 e^{-\zeta\mu z}\zeta(\mu^2 + 2\mu(\sigma+1) + \sigma(\zeta z+2)+1)}{(1+\mu)^4(\mu+\sigma)},$$

whereas

$$f_u(0) = \frac{\mu^2((\mu+1)+\sigma)}{(\mu+\sigma)(1+\mu)^2}.$$

Proof: Let claim severity follow an exponential distribution $\zeta > 0$, we know that gamma (n, ν) distribution is n^{th} fold convolution of exponential distribution, which is given as

$$f^{*n}(z) = \frac{\zeta^n}{(n-1)!} z^{n-1} e^{-\zeta z}, \qquad n = 1,2,....$$

Therefore, the random variable U has p.d.f given as:

$$f_u(z) = \frac{\mu^2}{(1+\mu)(\mu+\sigma)} e^{-\zeta z} \sum_{n=1}^{\infty} \left\{ \frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^z} \right\} \frac{\zeta^n}{(n-1)!} z^{n-1} e^{-\zeta z}$$

$$f_u(z) = \frac{\mu^2 e^{-\zeta\mu z} \zeta(\mu^2 + 2\mu(\sigma+1) + \sigma(\zeta z+2)+1)}{(1+\mu)^4(\mu+\sigma)}.$$

The probability of no claim is given by

$$f_u(0) = \frac{\mu^2((\mu+1)+\sigma)}{(\mu+\sigma)(1+\mu)^2}.$$

Theorem 4.2: For collective risk model with PWED (μ, σ) as a primary distribution and Erlang $(2, \zeta)$ as secondary distribution, then the probability density function of aggregate loss random variable U= $\sum_{i=0}^{M} Z_i$ is given as

$$fu(z) = \frac{\mu^2 e^{-\zeta z}}{(\mu+\sigma)(1+\mu)^{\frac{7}{2}} 2z} \left(\begin{array}{l} z^2\zeta^2\sqrt{1+\mu}\,\sigma \cosh\frac{z\zeta}{\sqrt{1+\mu}} + z\zeta \sinh\frac{z\zeta}{\sqrt{1+\mu}} \\ \left(2\mu^2 + 3\mu\sigma + 2 + 4\mu\right) \end{array} \right),$$

whereas

$$f_u(0) = \frac{\mu^2((\mu+1)+\sigma)}{(\mu+\sigma)(1+\mu)^2}.$$

Proof: Let claim severity follow Erlang $(2, \zeta)$ and we know that gamma $(2n, \zeta)$ distribution is n^{th} fold convolution of Erlang $(2, \zeta)$ distribution with pdf given as

$$f^{*n}(z) = \frac{\zeta^{2n}}{(2n-1)!} z^{2n-1} e^{-\zeta z}, \qquad n = 1,2,....$$

So, the aggregate loss random variable U has pdf given as

$$fu(Z)=\sum_{n=1}^{\infty}\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}(\mu+\sigma)}\right]\frac{\zeta^{2n}}{(2n-1)!}z^{2n-1}e^{-\zeta z}$$

$$fu(z)=\frac{\zeta^3}{(\zeta\eta+2\beta)(1+\zeta)^3}e^{-\nu z}\sum_{n=1}^{\infty}\frac{((1+\mu)+\sigma(z+1))}{(1+\zeta)^n}\frac{\nu^{2n}}{(2n-1)!}z^{2n-1}$$

$$fu(z)=\frac{\mu^2 e^{-\zeta z}}{(\mu+\sigma)(1+\mu)^{\frac{7}{2}}2z}\left(\begin{array}{l}z^2\zeta^2\sqrt{1+\mu}\sigma\cosh\frac{z\zeta}{\sqrt{1+\mu}}+z\zeta\sinh\frac{z\zeta}{\sqrt{1+\mu}}\\ \left(2\mu^2+3\mu\sigma+2+4\mu\right)\end{array}\right).$$

The probability of no claim is given by

$$f_u(0)=\frac{\mu^2((\mu+1)+\sigma)}{(\mu+\sigma)(1+\mu)^2}.$$

5. Reliability Analysis

5.1. Reliability Function R(z): The reliability function is defined as the probability that a system survives beyond a certain time. The reliability function or the survival function of PWED is given as

$$R(z,\mu,\sigma)=\left(\frac{\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z}{(\mu+\sigma)^{z+2}(\mu+\sigma)}\right).$$

5.2. Hazard Function: The hazard function, also known as the hazard rate, is given as

$$\text{H.R=h}(z,\mu,\sigma)=\frac{f(z;\mu,\sigma)}{R(z;\mu,\sigma)}=\frac{\mu^2[(1+\mu)+\sigma(z+1)]}{\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z}.$$

5.3. Reverse Hazard Rate and Mills Ratio: The reverse hazard rate and the Mills ratio of PWED are respectively given as

$$\text{R.H.R=}h_r(z,\mu,\sigma)=\frac{\mu^2[(1+\mu)+\sigma(z+1)]}{(\mu+\sigma)^{z+2}(\mu+\sigma)-(\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z)},$$

$$\text{Mills ratio}=\frac{(\mu+\sigma)^{z+2}(\mu+\sigma)-(\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z)}{\mu^2[(1+\mu)+\sigma(z+1)]}.$$

6. Statistical properties

In this section, structural properties of the PWE model have been evaluated.

6.1. Moments

6.1.1. Factorial Moments

Using (2.1), the rth factorial moment about origin of the PWED (2.1) can be obtained. $\mu_{(r)}{}' = E[E(Z^{(r)} \mid \nu)]$, *where* $Z^{(r)} = Z(Z-1)(Z-2)...(Z-r+1)$

$$\mu_{(r)}{}' = \frac{\mu^2}{\mu+\sigma}\int_0^\infty \left[\nu^r\left(\sum_{z=r}^{\infty}\frac{e^{-\nu}\nu^{z-r}}{(z-r)!}\right)\right](1+\sigma\nu)e^{-\mu\nu}d\nu$$

Taking $u = z - r$, *we get*

$$\mu_{(r)}{}' = \frac{\mu^2}{\mu+\sigma}\int_0^\infty \left[\nu^r\left(\sum_{u=0}^{\infty}\frac{e^{-\nu}\lambda^u}{u!}\right)\right](1+\sigma\nu)e^{-\mu\nu}d\nu$$

$$\mu_{(r)}{}' = \frac{\mu^2 r!}{\mu+\sigma}\left[\frac{\mu+\sigma(r+1)}{\mu^{r+2}}\right] \qquad 6.1$$

Taking r=1,2,3,4 in (5.1), the first four factorial moments about origin of the PWED can be obtained as

$$\mu_{(1)}{}' = \frac{1}{\mu+\sigma}\left[\frac{\mu+2\sigma}{\mu}\right], \qquad \mu_{(3)}{}' = \frac{6}{\mu+\sigma}\left[\frac{\mu+4\sigma}{\mu^3}\right],$$

$$\mu_{(2)}{}' = \frac{2}{\mu+\sigma}\left[\frac{\mu+3\sigma}{\mu^2}\right], \qquad \mu_{(4)}{}' = \frac{24}{\mu+\sigma}\left[\frac{\mu+5\sigma}{\mu^4}\right].$$

6.1.2. Moments about origin (Raw moments)

Using the relationship between the factorial moments about origin and the moments about origin, the first four moments about origin of the PWED (2.1) can be obtained as:

$$\mu_1{}' = \frac{1}{\mu+\sigma}\left[\frac{\mu+2\sigma}{\mu}\right]$$

$$\mu_2{}' = \frac{\mu(\mu+2\sigma)+2(\mu+3\sigma)}{\mu^2(\mu+\sigma)}$$

$$\mu_3{}' = \frac{6(\mu+4\sigma)+6\mu(\mu+3\sigma)+\mu^2(\mu+2\sigma)}{\mu^3(\mu+\sigma)}$$

$$\mu_4{}' = \frac{24(\mu+5\sigma)+36\mu(\mu+4\sigma)+14\mu^2(\mu+3\sigma)+\mu^3(\mu+2\sigma)}{\mu^4(\mu+\sigma)}.$$

6.1.3. Moments about the Mean (Central moments)

Using the relationship $\mu_r = E(Y-\mu_1{}')^r = \sum_{H=0}^{r}\binom{r}{h}\mu_h{}'(-\mu_1{}')^{r-h}$ between the moments about the mean and the moments about origin, the moments about the mean of the

PWED (2.1) can be obtained as $\mu_2 = \frac{\mu(\mu+2\sigma)+2(\mu+3\sigma)}{\mu^2(\mu+\sigma)} - \left[\frac{1}{\mu+\sigma}\left[\frac{\mu+2\sigma}{\mu}\right]\right]^2$

$$\mu_3 = \frac{6(\mu+4\sigma)+6\mu(\mu+3\sigma)+\mu^2(\mu+2\sigma)}{\mu^3(\mu+\sigma)} - 3\frac{(\mu(\mu+2\sigma)+2(\mu+3\sigma))}{\mu^2(\mu+\sigma)}\frac{1}{\mu+\sigma}\left[\frac{\mu+2\sigma}{\mu}\right]$$

$$\mu_4 = \frac{24(\mu+5\sigma)+36\mu(\mu+4\sigma)+14\mu^2(\mu+3\sigma)+\mu^3(\mu+2\sigma)}{\mu^4(\mu+\sigma)} - 4\frac{6(\mu+4\sigma)+6\mu(\mu+3\sigma)+\mu^2(\mu+2\sigma)}{\mu^3(\mu+\sigma)}\frac{1}{\mu+\sigma}\left[\frac{\mu+2\sigma}{\mu}\right]$$
$$+6\frac{\mu(\mu+2\sigma)+2(\mu+3\sigma)}{\mu^2(\mu+\sigma)}\left[\frac{1}{\mu+\sigma}\left[\frac{\mu+2\sigma}{\mu}\right]\right]^2 - 3\left[\frac{1}{\mu+\sigma}\left[\frac{\mu+2\sigma}{\mu}\right]\right]^4$$

6.2. Coefficient of variation (c.v) , skewness $(\sqrt{\beta_1})$,, kurtosis (β_2) and Index of Dispersion (γ)

$$C.V = \frac{\sqrt{(\mu(\mu+2\sigma)+2(\mu+3\sigma))(\mu+\sigma)-(\mu+2\sigma)^2}}{(\mu+2\sigma)}$$

$$\sqrt{\beta_1} = \frac{(\mu+\sigma)(6(\mu+4\sigma)+6\mu(\mu+3\sigma)+\mu^2(\mu+2\sigma))-3(\mu(\mu+2\sigma)+2(\mu+3\sigma)(\mu+2\sigma))(\mu+2\sigma)}{((\mu+\sigma)(\mu(\mu+2\sigma)+2(\mu+3\sigma))-(\mu+2\sigma)^2)^{3/2}}$$

$$\beta_2 = \frac{((24(\mu+5\sigma)+36\mu(\mu+4\sigma)+14\mu^2(\mu+3\sigma)+\mu^3(\mu+2\sigma))-4(6(\mu+4\sigma)+6\mu(\mu+3\sigma)+\mu^2(\mu+2\sigma))(\mu+2\sigma)}{-6(\mu(\mu+2\sigma)+2(\mu+3\sigma))(\mu+2\sigma)^2-3(\mu+\sigma)(\mu+2\sigma))(\mu+\sigma)^2}$$
$$\frac{}{(\mu(\mu+2\sigma)+2(\mu+3\sigma))^2(\mu+\sigma)(\mu+2\sigma)^4}$$

$$\gamma = \frac{(\mu(\mu+2\sigma)+2(\mu+3\sigma))(\mu+2\sigma)-(\mu+2\sigma)}{\mu(\mu+\sigma)(\mu+2\sigma)}.$$

Table 1. Index of Dispersion, Mean and Variance of PWED (μ,σ) for different values of parameters

	μ	**0.2**	**0.5**	**0.8**	**1**	**1.2**	**1.5**	**1.8**
$\sigma = 1.2$	IOD	5.3	2.8	3.19	4.13	5.56	8.64	12.95
	VAR	50.6	10.09	6.83	6.88	7.53	9	10.98
	MEAN	9.55	3.6	2.14	1.66	1.35	1.05	0.84
$\sigma = 1.5$	IOD	5.35	2.7	2.89	3.65	4.85	7.52	11.34
	VAR	50.3	9.5	5.97	5.84	6.3	7.52	9.16
	MEAN	9.4	3.5	2.06	1.6	1.3	1	0.8
$\sigma = 1.8$	IOD	5.4	2.66	2.71	3.36	4.42	6.85	10.37
	VAR	50	9.1	5.43	5.19	5.53	6.6	8.06
	MEAN	9.3	3.41	2	1.54	1.25	0.96	0.77
$\sigma = 2$	IOD	5.3	2.76	3.07	3.93	5.27	8.19	12.31
	VAR	50.5	9.86	6.49	6.47	7.03	8.44	10.26
	MEAN	9.5	3.56	2.11	1.64	1.33	1.03	0.85
$\sigma = 2.2$	IOD	5.3	2.84	3.31	4.32	5.84	9.09	13.6
	VAR	50.7	10.3	7.17	7.3	8.02	9.66	11.7
	MEAN	9.6	3.62	2.16	1.68	1.37	1.06	0.86

6.3. Moment generating function and probability generating function of Poisson weighted Exponential Distribution

Theorem 6.3.1: If Z has the PWED (μ,σ), then the probability generating function $P_z(t)$ has the following form:

$$P_X(t) = \frac{\mu^2}{(\mu+\sigma)(1+\mu)}\left[\frac{(\mu+1-t)((1+\mu)+\sigma)+\sigma t}{(\mu+1-t)^2}\right]$$

Proof: We begin with the well-known definition of the probability generating function given by

$$P_Z(t) = \sum_{x=0}^{\infty} t^z\left[\frac{\mu^2}{\mu+\sigma}\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}}\right]\right]$$

$$P_Z(t) = \frac{\mu^2}{(\mu+\sigma)(1+\mu)}\left[\frac{(\mu+1-t)((1+\mu)+\sigma)+\sigma t}{(\mu+1-t)^2}\right]$$

Theorem 6.3.2: If X has the PWED (μ,σ), then the moment generating function $M_Z(t)$ has the following form:

$$M_Z(t) = \frac{\mu^2}{(\mu+\sigma)(1+\mu)}\left[\frac{(\mu+1-e^t)((1+\mu)+\sigma)+\sigma e^t}{(\mu+1-e^t)^2}\right]$$

Proof: We begin with the well-known definition of the moment generating function given by $M_Z(t) = \sum_{z=0}^{\infty} e^{tz}\left[\frac{\mu^2}{\mu+\sigma}\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}}\right]\right]$

$$M_Z(t) = \frac{\mu^2}{(\mu+\sigma)(1+\mu)}\left[\frac{(\mu+1-e^t)((1+\mu)+\sigma)+\sigma e^t}{(\mu+1-e^t)^2}\right]$$

Similarly Laplace and Fourier transforms as calculated as

$$L_Z(t) = \frac{\mu^2}{(\mu+\sigma)(1+\mu)}\left[\frac{(\mu+1+e^t)((1+\mu)+\sigma)+\sigma e^t}{(\mu+1+e^t)^2}\right]$$

$$F_Z(t) = \frac{\mu^2}{(\mu+\sigma)(1+\mu)}\left[\frac{(\mu+1+t)((1+\mu)+\sigma)+\sigma t}{(\mu+1+t)^2}\right]$$

6.4 Recurrence Relation between Probabilities

The PWED can be written as

$$P(Z=z) = \frac{\mu^2}{\mu+\sigma}\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}}\right]$$

$$P(Z=z+1) = \frac{\mu^2}{\mu+\sigma}\left[\frac{(1+\mu)+\sigma(z+2)}{(1+\mu)^{z+3}}\right].$$

Dividing P(Z=z+1) by P(Z=z), we find the recurrence relation between probabilities

$$P(Z=z+1)=\frac{(z+2)(\eta(1+\zeta)+\beta(z+3))}{(1+\zeta)(z+1)(\eta(1+\zeta)+\beta(z+2))}P(Z=z) \quad .$$

6.5. Quantile function

Theorem 6.5: The quantile function of the PWED (μ,σ) is

$$Q_Z(u)=-\frac{(\mu+\mu^2+\sigma+2\mu\sigma)}{\mu\sigma}-\frac{1}{\log(1+\mu)}W_{-1}\left[\frac{(\mu+\sigma)(u-1)\log(1+\mu)}{\mu\sigma(1+\mu)^{\frac{\mu+\mu^2+\sigma}{\mu\sigma}}}\right].$$

Proof: The cdf of the distribution is

$$F_Z(z)=1-\frac{\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z}{(1+\mu)^{z+2}(\mu+\sigma)}.$$

The u^{th} quantile function is obtained by solving $F_Z(z)=u$

$$\frac{-(\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z)}{(\mu\sigma)}=Z+\frac{(u+1)(1+\mu)^{z+2}(\mu+\sigma)}{\mu\sigma}$$

$$Q_Z(u)=-\frac{(\mu+\mu^2+\sigma+2\mu\sigma)}{\mu\sigma}-\frac{1}{\log(1+\mu)}W_{-1}\left[\frac{(\mu+\sigma)(u-1)\log(1+\mu)}{\mu\sigma(1+\mu)^{\frac{\mu+\mu^2+\sigma}{\mu\sigma}}}\right].$$

7. Order statistics

Let $Z_{(1)},Z_{(2)},Z_{(3)}\ldots,Z_{(n)}$ be the ordered statistics of the random sample $Z_1,Z_2,Z_3,\ldots Z_n$ drawn from the discrete distribution with cdf $F_Z(z)$ and pmf $P_Z(z)$, then the pmf of the rth order statistics $Z_{(r)}$ is given by:

$$f_{z(r)}(z,\mu,\sigma)=\frac{n!}{(r-1)!(n-r)!}P(z)[F(z)]^{r-1}[1-F(z)]^{n-r}.\ r=1,\ 2,\ 3,\ldots,n$$

Using the equations (2.1) and (2.2), the probability density function of the rth order statistics of the Poisson weighted exponential distribution is given by

$$f_{(r)}(z,\mu,\sigma)=\frac{n!}{(r-1)!(n-r)!}\frac{\mu^2}{\mu+\sigma}\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}}\right]\left[1-\frac{\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma x}{(\mu+\sigma)^{x+2}(\mu+\sigma)}\right]^{r-1}\left[\frac{\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z}{(\mu+\sigma)^{z+2}(\mu+\sigma)}\right]^{n-r}$$

Then, the pmf of the first order $Z_{(1)}$ Poisson weighted exponential distribution is given by

$$f_1(Z;\mu,\sigma)=n\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}}\right]\left[\frac{\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z}{(\mu+\sigma)^{z+2}(\mu+\sigma)}\right]^{n-1}.$$

And the pmf of the nth order $Z_{(n)}$ Poisson weighted exponential model is given as $f_{(n)}(z,\mu,\sigma)=n\left[\frac{(1+\mu)+\sigma(z+1)}{(1+\mu)^{z+2}}\right]\left[1-\frac{\mu+\mu^2+\sigma+2\mu\sigma+\mu\sigma z}{(\mu+\sigma)^{z+2}(\mu+\sigma)}\right]^{n-1}$.

8. Estimation of Parameters

In this section, we estimate the parameters of the Poisson weighted exponential distribution using methods of moments and the method of maximum likelihood estimation.

8.1. Method of Moments

In order to obtain sample moments, we replace population moments with sample moments:

$$a=\mu_1'=\frac{\mu+2\sigma}{\mu(\mu+\sigma)}$$
$$b=\mu_2'=\frac{\mu(\mu+2\sigma)+2(\mu+3\sigma)}{\mu^2(\mu+\sigma)}.$$

Solving the above equations of sample moments, we get

$$\hat{\sigma}=\frac{\hat{\mu}-a\hat{\mu}^2}{\hat{a}\hat{\mu}-2}$$
$$\hat{\mu}=\frac{2a+\sqrt{2a-2b+4a^2}}{b-a}.$$

Theorem 8.1: The MOM estimator $\hat{\mu}$ *of* μ is positively biased.

$$Proof: Let\ \hat{\mu}=h(\bar{Z})\ where\ \ h(t)=\frac{1-\sigma t+\sqrt{\sigma^2t^2+6\sigma t+1}}{2t} \qquad ,t>0$$

$$Since\ \ h''(t)=\frac{1}{t^3}+\frac{9\sigma t+15\sigma^2t^2+3\sigma^3t^3+}{t^3(1+\sigma^2t^2+6\sigma)^{\frac{3}{2}}}>0$$

Then, h(t) is strictly convex. Hence, by Jensen's inequality, we have
$E\{h(\bar{z})\}>h\{E(\bar{z})\}$

$$finally, since\ h\{E(\bar{z})\}=h(\mu)=h\left(\frac{2\sigma+\mu}{\mu(\mu+\sigma)}\right)=\mu,\ we\ obtain E(\hat{\mu})>\mu.$$

Theorem 8.2: The MOM estimator $\hat{\mu}$ *of* μ is consistent and asymptotically normal.

$$\sqrt{n}\,(\hat{\mu}-\mu)\to_d N(0,\upsilon^2(\mu))$$
$$\upsilon^2(\mu)=\frac{\mu^2(\sigma+\mu)^2(2\sigma^2+2\mu\sigma^2+3\sigma\mu^2+4\sigma\mu+\mu^3+\mu^2)}{(2\sigma^2+\mu^2+4\sigma\mu)^2}.$$

Where

prof: Consistency Since $\mu < \infty$, *then* $\bar{Z} \xrightarrow{p} \mu$. *Also,* sin *ceh(t) is continuous funcationat* $t=\mu$

$$then\ h(\bar{z}) \xrightarrow{p} h(\mu), i.e, \hat{\mu} \xrightarrow{p} \mu$$

Asymptotic normality: Since $\sigma^2 < \infty,$ then by the central limit theorem, we have

$$\sqrt{n}\ (\overline{Z} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Also, since $h(\mu)$ *is differentiable and* $h'(\mu) \neq 0,$ by the delta-method, we have

$$\sqrt{n}(h(\bar{z}) - h(\mu)) \rightarrow_d N(0, \upsilon^2(\eta))$$

$$\upsilon^2(\mu) = \frac{\mu^2(\sigma+\mu)^2(2\sigma^2 + 2\mu\sigma^2 + 3\sigma\mu^2 + 4\sigma\mu + \mu^3 + \mu^2)}{(2\sigma^2 + \mu^2 + 4\sigma\mu)^2}.$$

Where

The theorem follows.

As a result of this, the asymptotic $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{\mu} \pm z_{\frac{\alpha}{2}} \frac{\upsilon(\bar{\mu})}{\sqrt{n}},$$

Where $z_{\frac{\alpha}{2}}$ is the (1- $\frac{\alpha}{2}$) percentile of the standard normal distribution.

8.2. Method of Maximum Likelihood Estimation

This is one of the most useful methods for estimating the different parameters of the distribution. Let $Z_1, Z_2, Z_3, ..., Z_n$ be the random sample of size n draw from PWED, then the likelihood function of PWED is given as

$$L(z \mid \mu, \sigma) = \frac{\mu^{n2}}{(\mu+\sigma)^n} \prod_{i=1}^{n} \left(\left[\frac{(1+\mu) + \sigma(z+1)}{(1+\mu)^{z+2}} \right] \right)$$

$$\log L = 2n \log \mu - n \log(\mu+\sigma) + \sum_{i=1}^{n} \log((1+\mu) + \sigma(z+1)) - (\sum_{i=1}^{n} z_i + 2n) \log(1+\mu)$$

$$\frac{\delta}{\delta\mu} \log L = \frac{2n}{\mu} - \frac{n}{\mu} + \sum_{i=1}^{n} \frac{1}{((1+\mu) + \sigma(z+1)} - \frac{\sum_{i=1}^{n} z_i + 2n}{(1+\mu)} = 0$$

$$\frac{\delta}{\delta\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(z+1)}{((1+\mu) + \sigma(z+1)} = 0$$

The above equations can be solved numerically by using R software (3.5.2).

9. Applications of Poisson weighted exponential distribution

In this section, we fit our proposed distribution to a data set representing vaccine adverse event counts and the number of claims in automobile insurance so as to illustrate our claim that our proposed model fits well when compared to other competing models. The data sets are given in Table 2 and 5 respectively. In Table 6 the degree of freedom is zero for some distributions, and hence p-value is not given and thus in such tables comparisons can be done on the basis of the AIC and BIC values.

Table 2. Dataset representing vaccine adverse event counts (see C. E. Rose, S.W. Martain, K. A. Wannemueler, B. D. Plikaytis (2006))

Counts	0	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1437	1010	660	428	236	122	62	34	14	8	4	4	1

We compute the expected frequencies for fitting Poisson Weighted Exponential (PWED), Zero Inflated Poisson (ZIPD), Negative Binomial (NBD), Geometric (GD), Poisson Lindley (PLD), Poisson Akash (PAD), Poisson Distribution (PD) and Discrete Generalized Inverse Weibull Distribution (DGIWD) with the help of R studio statistical software, and Pearson's chi-square test is applied to check the goodness of fit of the models discussed. The calculated figures are given in Table 3 and 6. Based on the chi-square, we observe that the Poisson weighted exponential distribution provides a satisfactorily better fit for the data set representing vaccine adverse event counts in Table 3 and the number of claims in automobile insurance in Table 6 as compared to other distributions. Also the parameters are estimated by using the ML method. We have analysed the data using R software (3.5.2). Parameter estimates along and the model function of the fitted distributions are given in Table 3 and 6.

Table 3. Fitted proposed distribution and other competing models to a dataset representing vaccine adverse event counts

Z	Obs.freq	PD	ZIPD	NBD	GD	PLD	DGIW	PAD	PWED
0	1437	890.75	1436.4	1401.7	1603.5	1500.1	1354	1500.2	1417
1	1010	1342.35	810.6	1065.3	963.9	1003.5	1377.2	977.65	1048
2	660	1011.4	789.6	671.15	579.4	629.2	524.15	632.55	670
3	428	508	529.5	393.75	348.3	378.7	248.9	392	397.4
4	236	191.4	274.6	222.5	209.35	221.6	139	232.1	225.1
5	122	57.7	117.3	122.8	125.85	127	86.45	132.1	123.6
6	62	14.5	42.9	66.7	75.65	71	57.9	72.8	66.3
7	34	3.1	13.8	35.8	45.45	39.9	41	39	35
8	14	0.6	4	19.1	27.35	22	30.2	20.45	18.2

Table 3. Fitted proposed distribution and other competing models to a dataset representing vaccine adverse event counts (cont.)

Z	Obs.freq	PD	ZIPD	NBD	GD	PLD	DGIW	PAD	PWED
9	8	0.1	1	10.1	16.45	12	23	10.5	9.4
10	4	0.1	0.3	5.3	9.9	7	18	5.35	4.8
11	4	0.1	0.1	2.75	5.9	3.55	14	2.65	2.9
12	1	0.1	0.1	1.45	8.95	4	105	2.55	1.7
Total	4020								
d.f		5	6	8	9	9	7	9	8
Chi-square		1901	570	10.4	78.68	20.12	488	14.16	6.21
Parameter estimate (S.E)		$\hat{\lambda}=1.5$ (0.019)	$\hat{\mu}=2.04$ (0.031) $\hat{\sigma}=0.261$ (0.009)	$\hat{r}=1.52$ (0.07) $\hat{p}=0.5$ (0.01)	$\hat{p}=0.398$ (0.004)	$\hat{\theta}=0.99$ (0.016)	$\hat{a}=1.49$ (0.02) $\hat{b}=0.47$ (0.17) $\hat{t}=0.037$ (0.068)	$\hat{\lambda}=1.35$ (0.016)	$\hat{\mu}=1.14$ (0.04) $\hat{\sigma}=2.99$ (0.9)
p-value		0.00	0.00	0.23	0.00	0.017	0.00	0.11	0.62

Furthermore, from Table 4 and 7, it has been observed that the Poisson weighted exponential distribution have the lesser AIC and BIC values as compared to other competing models. Hence, we can conclude that the PWED leads to a best fit as compared to other competing models for analysing the data set given in Table 2 and 5.

Table 4. Model comparison criterion for fitted models to a data set representing vaccine adverse event counts

Criterion	PD	NBD	ZIPD	GD	PLD	DGIW	PAD	PWED
AIC	14464.2	13485.2	13741.5	13558	13494	14049.8	13487	13480
BIC	14470.5	13486.33	13754.1	13564.3	13500.3	14068.7	13493.3	13481

Table 5. Data set representing the number of claims in automobile insurance (see Klugman et al. (2012))

Claim counts	0	1	2	3	4
Observed frequency	1563	271	32	7	2

Table 6. Fitted proposed distribution and other competing models to a data set representing the number of claims in automobile insurance

Z	Obs.fre	PD	ZIPD	NBD	GD	PLD	DGIW	PAD	PWED
0	1563	1544.1	1562.9	1566.4	1570.1	1569.5	1562.7	1571.9	1564.4
1	271	299.8	265.2	261.5	255.25	256.34	274.6	252.7	268.4
2	32	29.1	42	40.15	41.5	41.34	27.15	41.85	35.7
3	7	1.9	4.45	6.6	6.75	6.6	6.3	7	5.6
4	2	0.1	0.35	1	1.3	1.0444	4.25	1.45	0.75
Total	1875								
d.f		1	-	1	2	2	-	2	1
Chi-square		57.04	-	3.61	3.29	3.87	-	3.72	1.51
Parameter estimate (S.E)		$\hat{\lambda} = 0.194$ (0.01)	$\hat{\alpha} = 0.31$ (.042) $\hat{\beta} = .38$ (.07)	$\hat{p} = 6.13$ (0.4) $\hat{r} = 1.19$ (0.07)	$\hat{p} = 0.83$ (0.007)	$\hat{\theta} = 05.89$ (0.3)	$\hat{a} = 3.16$(. $\hat{b} = .4$(.17 $\hat{t} = .03$(.1	$\hat{\theta} = 05.74$ (0.27)	$\hat{\alpha} = 7.87$ (2.3) $\hat{\beta} = 8.83$ (18.4)
p-value		0.00	-	0.05	0.19	0.14	-	0.15	0.22

Table 7. Model comparison criterion for fitted models to a data set representing the number of claims in automobile insurance counts

Criterion	PD	NBD	ZIPD	GD	PLD	DGIW	PAD	PWED
AIC	2005.5	1991.1	1995.5	1989.8	1989.7	1994.6	1990.25	1987
BIC	2011	1990.3	2006.2	1995.3	1995.25	2011.2	1995.8	1986.3

10. Conclusion

A new over-dispersed probability distribution is introduced using the compounding technique. Statistical properties of the proposed model are studied and applications in handling count data sets representing vaccine adverse counts and insurance claims are analysed.

References:

Berreto-Souza, W., Bakouch, H., (2013). A new lifetime model with decreasing failure rate. *A journal of theoretical and Applied Statistics*, 47(2), pp. 465–476 .

Hajebi,M., Rezaei, S. and Nadarajah, S., (2013). An exponential- negative binomial distribution. *REVSTAT-Statistical journal*, 11(2), pp. 191–210.

Asgharzadeh, A., Bakouch,H.S., Nadarajah, S. and Esmaeili, L., (2014). A new family of compound lifetime distributions. *Kybernetika*, 50, pp. 142–169.

Mohmoudi,E. Jafari, A. A., (2014). The compound class of linear failure rate power series distributions: model, properties and applications. *Communications in statistics-simulation and computation.*

Silva, R. B., Cordeiro, G. M., (2015). The Burr XII power series distribution: A new compounding family. *The Brazilian journal of probability and statistics*, 29, pp. 565–589.

Pinho, L. G. B., Cordeiro, G. M. and Nobre, J. S., (2015). On the Harris-G class of distributions: general results and application. *Brazilian journal of probability and statistics*, 29(4), pp. 813–832.

Bourguignon, M., Silva, R. B. and Cordeiro, G. M., (2014). Anew class of fatigue life distributions. *Journal of statistics Computation and Simulation*,84 , pp. 2619–2635.

Bordbar, F., Nematollahi, A. R., (2016). The modified exponential- geometric distribution. Communications in Statistics –*Theory and Methods*, 45(1), pp. 173–181.

Flores, D, J., Borges, P., Cancho, G. and Louzada, F., (2013). The complementary exponential power series distribution. *Brazilian journal of probability and statistics*, 27, pp. 565–584.

Ghitany M. E., Alqallaf F., Al- Mutairi D. K. and Husain H. A.,(2011). A two parameter weighted Lindley distribution and its applications to survival data, *Mathematics and Computers in Simulations*, 81, pp. 1190–1201.

Christophe Chesneau, Hassan S. Bakouch, Tassaddaq Hussain and Bilal A. Para, (2020). The cosine geometric distribution with count data modeling, Journal of Applied Statistics, DOI: 10.1080/02664763.2019.1711364.

R Core Team, (2019). R version 3.5.3: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Klugman S., Panjer H., Willmot G., (2012). Loss Models. From Data to Decisions. *John Wiley and Sons, New York.*

C. E. Rose, S. W.Martin, K. A. Wannemuehler, and B. D. Plikaytis, (2006). On the use of zero infilated and hurdle models for modeling vaccine adverse events count data, Journal of Biopharmaceutical Statistics, Vol. 16, pp. 463–481.

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 175–191, DOI 10.21307/stattrans-2021-033
Received – 24.11.2019; accepted – 07.06.2021

Modelling the occupational and educational choices of young people in Poland using Bayesian multinomial logit models

Wioletta Grzenda[1]

ABSTRACT

Binomial logit models are commonly used in the analysis of the situation of respondents on the labour market. Consequently, in most cases researchers consider two states: of being unemployed and employed or economically inactive and active. This paper focuses on the situation of young people aged 18 to 29 on the labour market in Poland. A major part of the people who comprise the studied group are still in education or combine education with work. Therefore, the participants of the research were divided into the following groups: the employed and not learning, those combining education with work, the unemployed, learners/students only, and those economically inactive and not at school. The model allowing an analysis which includes both the most common division into working and non-working persons as well as the division proposed in this study is a nested logit model. This model has a hierarchical structure and is a special case of a multinomial logit model. In this paper, all models were estimated within the Bayesian approach. The findings show that continuing education by young people may result from their problems with finding a job; moreover, combining work with education is not the preferred form of professional activity. In addition, the study examines the inequalities observed on the Polish labour market.

Key words: young people, labour market, education, multinomial logit model, Bayesian approach.

1. Introduction

In socio-economic research, models for the dichotomous dependent variables are very popular (Cramer, 2003; Allison, 2009). Unfortunately, with their use, only two states or two events for a given unit can be analysed. In the case of issues related to the labour market, division into economically active and economically inactive, as well as employed and unemployed persons is usually made. In the case where the examined feature has more than two levels, a better solution than combining selected categories is to use models for discrete outcome variables that can take more than two possible

[1] SGH Warsaw School of Economics, Collegium of Economic Analysis, Institute of Statistics and Demography, Poland. E-mail: wgrzend@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-2226-4563.

values. Among this group of models, two main classes are distinguished: models for ordinal response variables and models for dependent variables with unordered categories. The second group includes: the MultiNomial Logit Model (MNLM), the Conditional Logit Model (CLM), the Mixed Logit Model (MLM) and the Nested Logit Model (NLM) (Cameron and Trivedi, 2005).

The choice of a model depends primarily on whether the independent variables included in the model vary across alternatives or they are the same across alternatives (Cameron and Trivedi, 2005). The standard multinomial logit model can be used when the model takes into account only the features of the individuals studied without taking into account the features of the selected categories. If this assumption is not met, the conditional logit model is used unless both types of features are considered. In the latter case the mixed logit model is used. In addition, according to Stanisz (2016), in order for the standard multinomial model to be used, the categories of the responding variable should be independent and distinguishable for the decision maker. Both the multinomial logit model and the conditional logit model have some limitations regarding the assumption of independence from irrelevant (unrelated) alternatives (IIA). The model in which this assumption can be slightly weakened, and also can take into account the hierarchy of alternatives, is the nested logit model considered in this paper. This model is not widely used due to the problems related to the estimation of its parameters. To avoid these problems, the Bayesian approach and Markov Chain Monte Carlo methods (MCMC) were used in this work (Robert and Casella, 2004).

The purpose of this study is to analyse the occupational and educational choices of young people aged 18 to 29 in Poland. In most studies on this issue, the division of young people into those who have already completed education and those who continue their education, e.g. at a higher level (de Dios Jiménez and Salas-Velasco, 2000), economically active and economically inactive (MRPiPS, 2018) or unemployed and employed (Gallie and Paugam, 2000; Grzenda, 2012; Bieszk-Stolorz and Markowicz, 2013) are considered. The binary divisions presented above can be further detailed. For example, among the economically inactive there are both those who are unwilling to take up employment despite their abilities and young people who remain in the education system and have not started their careers yet. In addition, it is worth considering in the research that young people sometimes combine education with work. Therefore, in this study, the respondents were divided into employed, combining education with work, learners only, and unemployed or persons economically inactive but not being learners. The methodological approach proposed in this work makes it possible to consider in the analysis both a more general division into working and non-working persons, as well as a more detailed division taking into account education of youth. Information on educational and economic activity of young people in Poland was obtained from the Labour Force Survey (LFS).

The subject addressed in this study is very important because according to many reports (CSO, 2016a; CSO, 2016b; MRPiPS, 2018) the situation of young people on the labour market in Poland is the worst compared to other age groups. In addition, economists are concerned about the growing phenomenon of NEET (not in employment, education or training) (Chłoń-Domińczak and Strawiński, 2013), which affects young people who are neither in education nor working. The consequences of this phenomenon apply to the entire economy as well as to individuals who lose their competence over time. Youth unemployment has also a social dimension, lack of employment negatively affects family and fertility decisions, and, as a result, the demographic situation of the country. Therefore, the identification of factors determining the educational and professional decisions of young people may help identify solutions that may improve the situation of these people on the labour market in Poland.

2. Multinomial models

Models for unordered categorical dependent variable are also considered as discrete choice models and are most often used in marketing research (Anderson, De Palma and Thisse, 1992). In the case of the binomial logit model, it can be assumed that a given unit has two variants to choose from. Suppose now that the i-th unit $(i = 1, \dots, n)$ has to select not two but J unordered categories. These categories are mutually exclusive and constitute a whole set of possible selection options for the units under consideration. In the case where the independent variables do not differ for the alternatives considered, a standard multinomial logit model (MNLM) is considered. For this model, the probability of observing the choice by the i-th unit $(i = 1, \dots, n)$ of j-th category $(j = 1, \dots, J)$ is given by the formula:

$$p_{ij} = \frac{exp(\mathbf{x}'_i \boldsymbol{\beta}_j)}{\sum_{k=1}^{J} exp(\mathbf{x}'_i \boldsymbol{\beta}_k)}, i = 1, \dots, n, j = 1, \dots, J,$$

where $\mathbf{x}$ denotes the vector of independent variables and $\boldsymbol{\beta}$ is the vector of parameters. The sum of these probabilities for all categories $j = 1, \dots, J$ is 1.

If the independent variables differ for the alternatives considered, the standard multinomial model cannot be used; the conditional logit model (CLM) is considered then. In the case of this model, the probability of observing the selection of the j-th category $(j = 1, \dots, J)$ by the i-th unit $(i = 1, \dots, n)$ is given by the formula:

$$p_{ij} = \frac{exp(\mathbf{x}'_{ij} \boldsymbol{\beta})}{\sum_{k=1}^{J} exp(\mathbf{x}'_{ik} \boldsymbol{\beta})}, i = 1, \dots, n, j = 1, \dots, J.$$

The combination of both considered models is the mixed logit model (MLM) (Cameron and Trivedi, 2005).

The presented models can also be considered more generally in the context of the additive random utility models (ARUM) and discrete choice theory. In this approach, each unit assigns to each category j certain utility $U_j, j = 1, \dots, J$ and selects the one with the highest utility. Let

$$U_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}, i = 1, \dots, n, j = 1, \dots, J,$$

denote the utility function. By making different assumptions about the random component of utility, different multinomial logit models can be obtained.

In the standard multinomial logit model, the random components ε_j $(j = 1, \dots, J)$ are independent and identically Gumbel distributed (have the type I extreme-value distribution), with the density function given by the formula:

$$f(\varepsilon_j) = e^{-\varepsilon_j} exp(-e^{-\varepsilon_j}), j = 1, \dots, J.$$

According to assumptions made in (McFadden, 1974), to be able to use a standard logit multinomial model, the categories analysed must meet the assumption of independence from irrelevant alternatives (IIA). This assumption also applies to the conditional logit model. However, it is often not fulfilled. By eliminating or adding one alternative, the quotient of the probability of the categories considered so far often changes. Unfortunately, there are no tests that conclusively determine whether IIA assumption is met. Cheng and Long (2007) have shown that two existing tests by Hausman and McFadden (1984) and Small and Hsiao (1985) can be unreliable. Then the solution may be to use another model, namely the nested logit model (Train, 2009).

The nested logit model has a hierarchical structure. The set of all possible alternatives is divided into the so-called nests so that the assumption of independence from irrelevant alternatives (IIA) is met only in each nest, but it does not have to be met between the nests. Therefore, in the nested logit model, all random components ε_{ij} $(j = 1, \dots, J)$ do not have to be independent. In addition, instead of the Gumbel distribution, the generalized extreme-value distribution (GEV) is assumed for these components.

Let K denote the number of disjoint subsets (nests) $S_1, S_2, \dots, S_K$, into which the possible alternatives have been divided. Then, the cumulative distribution function for the random components vector $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iJ})$, is given by the formula:

$$F(\boldsymbol{\varepsilon}_i) = exp\left(-\sum_{k=1}^{K}\left(\sum_{j \in S_k} exp(-\varepsilon_{ij}/\lambda_k)\right)^{\lambda_k}\right).$$

Within each of the nests, random components ε_{ij} $(j = 1, \dots, J)$ are correlated. The λ_k parameter is a function of the correlation coefficient between possible alternatives in the k-th nest and is used to measure the correlation between the categories in the nest. The value of 1 for the λ_k parameter means no correlation in the

k-th nest, therefore if the value of this parameter for all nests is 1, then the nested logit model can be replaced with a standard multinomial logit model.

With the previously introduced notation, the choice probability for alternative $j \in S_k$ by i-th $(i = 1, \dots, n)$ unit for the nested logit model is given by the formula:

$$P(y_{ij} = 1) = \frac{exp(\mathbf{x}'_{ij}\boldsymbol{\beta}/\lambda_k)\left(\sum_{m \in S_k} exp(\mathbf{x}'_{im}\boldsymbol{\beta}/\lambda_k)\right)^{\lambda_k - 1}}{\sum_{l=1}^{K}\left(\sum_{m \in S_l} exp(\mathbf{x}'_{im}\boldsymbol{\beta}/\lambda_l)\right)^{\lambda_l}}.$$

Then, the likelihood function is in the form:

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\lambda}) = \prod_{i=1}^{N}\prod_{j=1}^{J}\left(P(y_{ij} = 1)\right)^{y_{ij}},$$

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$.

In this article, the Bayesian approach was used to estimate the parameters of the nested logit model (Lahiri and Gao, 2002; Rossi, Allenby and McCulloch, 2005). This approach requires a prior distribution for the vector of coefficient parameters $\boldsymbol{\beta}$ and the parameter vector $\boldsymbol{\lambda}$. For the parameter vector $\boldsymbol{\beta}$, depending on the prior information, the most common are flat priors or the normal prior distributions. For the components of the $\boldsymbol{\lambda}$ parameter vector and for $a > 0$, the following prior distribution was used in this paper:

$$p(\lambda) = \begin{cases} a\lambda^{a-1}\exp(-\lambda^a) & \text{for } \lambda > 0, \\ 0 & \text{for } \lambda \leq 0. \end{cases}$$

Examples of other prior distributions for the parameter vector $\boldsymbol{\lambda}$ can be found in Lahiri and Gao (2002). This could be, for example, a beta or gamma distribution. Using the notation applied for the nested logit model, the formula for the posterior distribution has the form:

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\lambda})p(\boldsymbol{\beta})p(\boldsymbol{\lambda}).$$

In this paper, Markov Chain Monte Carlo (MCMC) methods were used to determine the marginal posterior distributions, in particular the methods used were the Metropolis algorithm (Gelman, et al., 2000) and the Gamerman algorithm (Gamerman, 1997).

3. Reference data

To analyse the situation of young people on the labour market in Poland, data from the Labour Force Survey (LFS) were used. The LFS is a quarterly panel survey with a rotational sample selection scheme. In this study, the research sample comprised units that were surveyed for two consecutive quarters in 2015. These are people from the samples numbered 63-65 and 67-69. This selection of the sample enabled, *inter alia*,

verification of the answers given. In the first stage of the analysis, in accordance with the adopted research objective, people aged 18 to 29 were selected from the entire data set, thus separating a sample of 16,144 respondents. Then, the respondents were divided into five categories due to their situation on the labour market:

1. only learners/students,
2. employed but not being learners,
3. combining education with work,
4. unemployed persons but economically active,
5. economically inactive people but not being learners.

Learners were selected based on question No. 90 (*In the last 4 weeks, including as a last week the week of the survey, were you a student*?). Then, they were divided into people who had a job and those who did not. Having a job as described in this article means doing professional work in accordance with survey question 12 or having a job but temporarily not doing related work, as identified based on the answer to question 13. (Questions: *12. Did you perform work for at least 1 hour, which provided earnings or income in the week under study from Monday to Sunday, or assist in a family business for free?* 13. *Did you have a job in the week under study, but did not perform it temporarily?*). Then, from among persons who did not have a job and did not learn, economically active and economically inactive people were distinguished. Economically active persons mean those who were looking for a job and were ready to take up a job in accordance with survey questions 71 and 79. (*71. In the last 4 weeks including the week of the survey, did you look for a job? 79. Could you take a job in the 2 weeks following the week of the survey?*).

Table 1. A set of potential explanatory variables

Variable	Description	Categories	Percent
age_group	Age group at the time of the survey	1 = from 18 to 19 years old	17.68
		2 = from 20 to 24 years old	40.96
		3 = from 25 to 29 years old	41.35
sex	Sex	0 = woman	49.27
		1 = man	50.73
education	Level of education	1 = higher	22.76
		2 = post-secondary and secondary professional	22.01
		3 = secondary general	23.19
		4 = basic vocational	12.38
		5 = primary school	19.66
marital_status	Marital status	0 = unmarried, a widower, a widow, separated or divorced	78.81
		1 = married	21.19

Table 1. A set of potential explanatory variables (cont.)

Variable	Description	Categories	Percent
child	The presence of a child under 15 years in the household	0 = no	77.87
		1 = yes	22.13
place_residence	Class of place of residence during the survey	0 = village	47.63
		1 = town	52.37
region	Region of Poland	1 = Central Łódzkie, Mazowieckie)	15.03
		2 = Southwest (Dolnośląskie, Opolskie)	11.92
		3 = South (Małopolskie, Śląskie)	14.33
		4 = Northwest (Wielkopolskie, Zachodniopomorskie, Lubuskie)	15.04
		5 = North (Kujawsko-Pomorskie, Warmińsko-Mazurskie, Pomorskie)	18.14
		6 = East (Lubelskie,Podkarpackie, Świętokrzyskie, Podlaskie)	25.55

The employed only persons were the largest part of the entire group – 43.99%. Learners constituted 30.22%, among them were both economically inactive and unemployed people. Introducing a more detailed breakdown of learners would mean introducing more values of a dependent variable, and when interpreted against one reference level, it could give hardly clear results. In addition, substantive considerations also had an impact on this division. Namely, this group includes, for example, part-time students who did not enter full-time studies and often have difficulties in determining whether they are not working because they cannot find a job, or because a lot of their time is consumed by studying or it can be an obstacle that they have to attend weekend classes starting on Fridays. The next subgroup includes persons economically inactive but not being learners. The high percentage of economically inactive and not learning persons is worrying as the share of this group is 10.83%. The share of unemployed was 8.46%. Considering the general population in the period under study, it is worth emphasizing that among all the unemployed people aged 18 to 29 accounted for as much as 37.39% (CSO, 2016a). The smallest percentage share was obtained for working and studying people – 6.5%.

Based on the presented breakdown, the dependent variable was constructed. To do this, the last two groups, i.e. groups 4 and 5 were combined into one group: the unemployed and the inactive but not learning. In this way, a group of people unemployed and persons economically inactive but not being learners, which, consider phenomenon of NEET, was then selected as a reference group in the paper. One of the

research objectives was to analyse the impact of individual characteristics of the respondents on their situation on the labour market. Therefore, a set of potential explanatory variables included in this study was developed, which is presented in Table 1.

4. The model estimation

In the first stage of the analysis, the nested logit model was estimated in the Bayesian approach. Due to the primary division of the surveyed respondents into working and non-working persons, the two-nest model was chosen. The first nest contains both learning and unemployed, and persons economically inactive but not being learners, while the second one employed and people who combine work and education. Taking into account the large sample size, all considered models were estimated using normal non-informative prior distributions. For the parameter vector $\boldsymbol{\beta}$, the normal prior distributions with mean equal to 0 and variance equal to 100 were adopted in all models. The formula for the prior distribution for the lambda parameter has been presented in Section 2. In this paper, the Metropolis algorithm (Gelman, et al., 2000) or the Gamerman algorithm (Gamerman, 1997) have been used for sampling from multidimensional distributions, depending on the model under consideration.

The results for the nested logit model are presented in Table 2. The assessment of convergence of generated chains was made using the Geweke test. Based on the results obtained for both models at the significance level of $\alpha = 0.05$, the null hypothesis that the obtained chains for the considered parameters of these models are convergent cannot be rejected (Table 2). Two nests were included in the model and none of them was degenerated, therefore posterior values for two lambda parameters were determined. These parameters are used to measure the correlation between alternatives in each nest. The lambda values obtained are less than 1, therefore the nested logit model is a better model for analysing the situation of young people on the labour market in Poland in the examined period, compared to the standard multinomial logit model, because it takes into account the correlation in the considered nests.

Based on the results contained in Table 2, it can be concluded that if the option of non-working and not in education is not considered, then the second option, i.e. employed but not learning is the most important for the respondents, while the second most important one is only learning, in both cases compared to the option of non-working and not in education. On the other hand, the option of combining education with professional work definitely loses its significance, also compared to the reference option.

Table 2. Statistics of the posterior samples and Geweke convergence diagnostics for the nested logit model

Parameter	Posterior expected values	Posterior standard deviation	Highest probability density interval ($\alpha = 0.05$)		Geweke diagnostics	
					z	*p-value*
option 1	0.4025	0.1577	0.0935	0.7055	0.4957	0.6201
option 2	0.7383	0.3372	0.0743	1.3663	0.6567	0.5114
option 3	-1.0159	0.8474	-2.5516	0.7159	1.2513	0.2108
lambda 1	0.8966	0.3483	0.2164	1.5690	0.4646	0.6422
lambda 2	0.9172	0.3664	0.0294	1.5417	-1.0613	0.2886

In the next stage of the study, attempts were made to estimate the nested logit model with variables describing the characteristics of the respondents. Unfortunately, despite various attempts to improve the quality of generated chains, their convergence could not be achieved. Therefore, a standard multinomial model was considered, which is a generalization of the nested logit model (Allison, 2009). This approach was possible because the independent variables included in the model only describe the characteristics of the respondents and not the characteristics of the alternatives. The results of the estimation are presented in Table 3. This model was estimated under the same initial conditions as adopted in the first model. Prior to the interpretation of the results, the convergence of generated chains was also assessed using the Geweke test. Based on the results obtained, it was found that at the significance level $\alpha = 0.01$, the null hypothesis that the obtained chains for the considered parameters of these models are convergent cannot be rejected. Then, the received posterior expected values were interpreted.

In the case of the feature describing the respondent's sex, it was obtained that women, compared to men, were 2.18% less likely to remain in the education system as compared to the option of remaining unemployed or inactive, but not learning. In addition, they had a 58.25% less chance of doing work and a 27.42% less chance of staying in the education system and at the same time working than men, in both cases compared to persons non-working and not in education.

Young people aged 18 and 19 were more than 72 times more likely to remain in the education system, compared to people from the oldest age group 25–29, while people aged 20-24 were about 15 times more likely to remain in the education system compared to the same age group, in both cases compared to the option of persons non-working and not in education. The youngest and those aged 20–24 had 60.60% and 21.46%, respectively, less chance of having a job than people from the oldest age group. In addition, people aged 18 and 19 had more than 7 times more chances to have a job and remain in the education system compared to people from the oldest age group, while people aged 20 to 24 had these chances four times higher.

Of course, all interpretations presented in the paper remain valid under the assumption of *ceteris paribus*. In addition, in further interpretation of the results obtained, the reference level of the target variable is the same, i.e. we assume that for each of the considered options the reference level is a sum of unemployed and persons economically inactive but not being learners.

Single people were more than six times more likely to remain in the education system than married people, and more than twice as likely to combine work and study compared to married people. However, there were no major differences due to marital status in terms of performing only professional work.

Table 3. Statistics of the posterior samples and Geweke convergence diagnostics for the multinomial logit model with variable

Parameter		Posterior expected values	Posterior standard deviation	Highest probability density interval ($\alpha = 0.05$)		Geweke diagnostics	
						z	*p-value*
sex 0	1	-0.0220	0.0585	-0.1375	0.0918	0.1953	0.8452
sex 0	2	-0.8734	0.0482	-0.9714	-0.7837	-0.8965	0.3700
sex 0	3	-0.3205	0.0784	-0.4709	-0.1647	0.1490	0.8815
age_group 1	1	4.2766	0.1199	4.0418	4.5087	0.2409	0.8096
age_group 2	1	2.6963	0.0922	2.5088	2.8724	0.3924	0.6948
age_group 1	2	-0.9315	0.1214	-1.1637	-0.6913	-0.2102	0.8335
age_group 2	2	-0.2415	0.0524	-0.3445	-0.1386	-0.7510	0.4527
age_group 1	3	2.0762	0.1605	1.7528	2.3803	-0.7171	0.4733
age_group 2	3	1.3900	0.0956	1.2001	1.5712	0.2087	0.8347
marital_status 0	1	1.8844	0.1347	1.6100	2.1376	0.1940	0.8461
marital_status 0	2	-0.0004	0.0606	-0.1139	0.1218	0.5431	0.5871
marital_status 0	3	0.7593	0.1152	0.5345	0.9834	-0.8147	0.4152
education 1	1	1.1135	0.1090	0.8954	1.3210	-0.8506	0.3950
education 2	1	-0.5467	0.0960	-0.7313	-0.3558	0.4423	0.6583
education 3	1	0.8835	0.0891	0.7035	1.0521	-0.6394	0.5226
education 4	1	-2.3453	0.1294	-2.6017	-2.1006	-0.7995	0.4240
education 1	2	2.0446	0.0900	1.8702	2.2228	-0.6612	0.5085
education 2	2	1.4084	0.0839	1.2459	1.5712	-1.3094	0.1904
education 3	2	1.0270	0.0891	0.8618	1.2082	-0.0209	0.9834
education 4	2	0.9030	0.0856	0.7361	1.0753	-0.4503	0.6525
education 1	3	2.4821	0.1550	2.1858	2.7869	-0.8424	0.3996
education 2	3	1.0031	0.1491	0.7126	1.2938	-1.0947	0.2737
education 3	3	1.5160	0.1429	1.2411	1.7987	-0.3626	0.7169
education 4	3	-0.4994	0.1857	-0.8590	-0.1374	0.3333	0.7389
child 0	1	0.0519	0.0967	-0.1373	0.2364	-0.1032	0.9178
child 0	2	-0.6439	0.0614	-0.7656	-0.5236	0.7671	0.4430
child 0	3	-0.8808	0.0988	-1.0697	-0.6886	0.7285	0.4663
place_residence 0	1	-0.3460	0.0584	-0.4582	-0.2324	-0.3128	0.7544
place_residence 0	2	0.0440	0.0473	-0.0522	0.1332	-0.3450	0.7301

Table 3. Statistics of the posterior samples and Geweke convergence diagnostics for the multinomial logit model with variable (cont.)

Parameter		Posterior expected values	Posterior standard deviation	Highest probability density interval ($\alpha = 0.05$)		Geweke diagnostics	
						z	*p-value*
place_residence 0	3	-0.2625	0.0798	-0.4181	-0.1079	-2.4367	0.0148
region 1	1	0.1133	0.0919	-0.0664	0.2927	1.3343	0.1821
region 2	1	-0.0421	0.0971	-0.2329	0.1488	-0.1950	0.8454
region 3	1	0.2145	0.0938	0.0263	0.3925	0.1968	0.8440
region 4	1	-0.0856	0.0903	-0.2702	0.0819	1.1744	0.2402
region 5	1	-0.2119	0.0847	-0.3766	-0.0457	-0.6986	0.4848
region 1	2	0.5751	0.0751	0.4310	0.7237	-0.4479	0.6542
region 2	2	0.4017	0.0786	0.2486	0.5566	-0.8153	0.4149
region 3	2	0.4922	0.0765	0.3412	0.6410	0.6756	0.4993
region 4	2	0.4283	0.0724	0.2867	0.5681	0.8986	0.3688
region 5	2	0.2772	0.0687	0.1437	0.4110	0.7898	0.4297
region 1	3	0.5633	0.1233	0.3302	0.8084	0.6649	0.5061
region 2	3	0.4660	0.1310	0.2130	0.7237	-1.3811	0.1672
region 3	3	0.4651	0.1291	0.2152	0.7202	0.7235	0.4694
region 4	3	0.4529	0.1211	0.2223	0.6948	-0.4774	0.6330
region 5	3	0.3632	0.1141	0.1431	0.5893	0.7771	0.4371

People with higher education had three times more chances to remain only in the education system than people with basic education, for people with post-secondary and secondary vocational education these chances were 42.11% lower, for people with general secondary education the chances were over twice as large, and 90.42% less for people with basic vocational education. Chances for performing only professional work were more than seven times higher for people with higher education, four times higher for people with post-secondary and secondary vocational education, more than twice higher for people with general secondary education and basic vocational education, in each case compared to persons with basic education. The chances of combining education with work were more than eleven times higher for people with higher education, more than twice as high for people with post-secondary and secondary vocational education, more than four times higher for people with general secondary education and 39.31% lower for people with basic vocational education. In each case, compared to people with basic education.

For the variable describing the presence of a child under the age of 15 in a household in which the respondent or his/her spouse is its head, it was obtained that the lack of a child was only associated with a 5.32% increase in the chances of staying in the education system, compared to persons residing in households, in which a child or children were present. On the other hand, the chances of only working and combining

education with work were about 50% lower for these people, also compared to people living in households with children.

People living in the countryside had 29.25% less chance of staying only in the education system compared to urban residents. In addition, they were more than four times more likely to work only and had 23.09% less chance to combine education with work, in both cases compared to young people living in cities.

Comparing the eastern region to other regions of Poland, it was found that the inhabitants of each of them had at least 31% greater chances of only working and combining education with work compared to the inhabitants of the eastern region. In addition, residents of the central (Łódzkie, Mazowieckie) and southern (Małopolskie, Śląskie) regions were more likely to remain in the education system only, by 12% and 23%, respectively, compared to the eastern region. In other regions, these chances were lower compared to the eastern region, with the smallest chance of remaining only in the education system obtained for the northern region.

5. Conclusions

The occupational and educational choices of young people depend on many different factors related to both individual characteristics of these people including their work motivation (Davidescu, Roman, Strat and Mosora, 2019) as well as the socio-economic situation of a country. This work focuses on this first group of factors except for work motivation, due to the lack of relevant data for Poland in this regard. For modelling, the multinomial logit model, and its special case, i.e. the nested logit model (Allison, 2009) were chosen. Using the latter model, it was possible to take into account the hierarchical division of young people due to their status on the labour market as well as their education. However, as far as the analysis of the impact of the individual characteristics of the respondents on their occupational and educational choices is concerned, the standard multinomial model turned out to be the better model.

According to the human capital theory, the wage differences among occupations and the ability to learn during education are the main factors influencing occupational and educational choices of young people (Dale, 2009). What follows from our study is that young people in Poland prefer to focus mainly on working and to a lower extent on education. This may be associated with high opportunity costs of higher education, mainly foregone earnings. Combining education with work is not their preferred form of economic activity too. It can, therefore, be concluded that remaining in the education system is associated with the inability to find a suitable job, or with the prospect of getting a better job after obtaining higher education. Given the current massification of higher education (Jasiński, Bożykowski, Chłoń-Domińczak, Zając and Żółtak, 2017),

it is important to look for other solutions to improve the situation of young people on the labour market in Poland. Importantly, the professional situation of this age group is the worst compared to other age groups (CSO, 2016b).

In recent years, in Poland, the approach of young people to the balance between learning and work has been constantly changing. In this paper, we have shown to what extent this approach depends on their age too. It was found that people from the youngest age group from 18 to 19 years old had the best chance of remaining only in the education system, these people also had the least chance of being employed in comparison with people aged 25 to 29. Moreover, the people who most often combined work and study also belonged to the youngest age group, for the next age group, i.e. persons aged from 20 to 24, these chances were almost two times lower. The obtained result may be slightly disturbing in the face of "lifelong learning" widely promoted in the European Union. On the other hand, it is difficult to say unequivocally whether the experience of combining education and employment in youth may facilitate lifelong learning or it is a factor discouraging from pursuing further learning. However, many people from this age group decide to continue to combine both work and study as they are aware of the need of further education (Brooks, 2006).

According to G.S Becker (1991), the main determinant of the economic activity of an individual is education. Our empirical evidence suggests that people with higher education as well as post-secondary and secondary vocational education have greater chance of having a job compared to persons with basic education. On the other hand, people with general secondary education most often combined work and education, such a combination was least often among people with basic vocational education. People who had basic vocational education most often ended their educational activity at this stage. According to Jasiński, Bożykowski, Chłoń-Domińczak, Zając and Żółtak (2017), young people should choose carefully their educational pathway because the employment chances of university graduates in Poland depend on the study area, moreover they also change over time. In the context of combining work and learning they indicated that prior experience in the labour market has an impact on employment chances, but only in the first months after graduation.

The social inequalities in the labour market, including inequalities due to gender (Becker, 2010), have been a challenge for many labour markets in Europe. Our study indicates that during the period considered in Poland women had less chance of employment compared to men, and even less chance to combine education with work. According to Castellano and Rocca (2017,) education is the most important factor determining the gender gap in the labour market. In Poland, women are better educated than men. Therefore, we can agree with Castellano and Rocca that gender inequalities in the labour markets may depend on cultural factors, too. Moreover, as expected, single people were more likely to remain in the education system than

married people, and they were also more likely to combine work with education. In the case of respondents working only, there were no major differences due to marital status. In the context of having a family, it was found that having a child is no longer a major problem with continuing education and is also conducive to greater professional activity. According to other studies (Michaud and Tatsiramos, 2011), having a child mainly affects women's employment, but the effects of this influence vary from country to country. Based on one of the latest studies on Polish women (Grzenda, 2019), it was found that the differences in the professional careers of women without children and having children are becoming smaller.

Considering access to education of young people living in cities and in the countryside, the results of other studies have been confirmed (Kołaczek, 2005), according to which at the primary level of compulsory education there are no differences in access to education between cities and villages, these differences are only revealed at a secondary and higher level. In this study, it was found that rural residents had less chance of continuing education, as well as combining work with education, but these differences did not exceed 30% compared to young people living in cities. Given territorial division it was observed that in comparison with the inhabitants of the eastern region of Poland, the inhabitants of all other regions had a greater chance of employment as well as combining work with education. In addition, the inhabitants of the central and southern regions, in which larger scientific centres are concentrated, had a better chance of staying only in the education system compared to the eastern region.

The research methods proposed in this paper made it possible to determine the impact of individual characteristics of young people in Poland on their occupational and educational choices. In addition, our contribution to research in this area consisted of including in the model as many as four different states of their activity: employed but not learning; combining education with work; only learners/students; unemployed and persons economically inactive but not being learners. This provided new insight into how young people enter the labour market in Poland.

Acknowledgements

This paper was presented on the MSA 2019 conference which financed its publication. Organization of the international conference "Multivariate Statistical Analysis 2019" (MSA 2019) was supported from resources for popularization of scientific activities of the Minister of Science and Higher Education in the framework of agreement No. 712/P-DUN/202019.

This study has been prepared as part of the project granted by the National Science Centre, Poland, entitled "The modelling of parallel family and occupational careers with Bayesian methods" (2015/17/B/HS4/02064).

References

Allison, P. D., (2009). Logistic Regression Using the SAS®. Theory and Application. 8th ed. Cary, NC: SAS Institute Inc.

Anderson, S. P., De Palma, A. and Thisse, J. F., (1992). Discrete choice theory of product differentiation. Massachusetts: The MIT Press.

Becker, G. S., (1991). A Treatise on the Family. Cambridge: Harvard University Press.

Becker, G. S., (2010). The Economics of Discrimination. USA: University of Chicago Press.

Bieszk-Stolorz, B., Markowicz, I., (2013). Men's and Women's Economic Activity in Poland. Acta Universitatis Lodziensis. Folia Oeconomica, 285, pp. 221–227.

Brooks, R., (2006). Learning and work in the lives of young adults. International Journal of Lifelong Education, 25(3), pp. 271–289.

Cameron, A. C., Trivedi, P. K., (2005). Microeconometrics: methods and applications. Cambridge: Cambridge University Press.

Castellano, R., Rocca, A., (2017). The dynamic of the gender gap in the European labour market in the years of economic crisis. Quality & Quantity, 51(3), pp. 1337–1357.

Cheng, S., Long, J. S., (2007). Testing for IIA in the multinomial logit model. Sociological Methods & Research, 35(4), pp. 583–600.

Chłoń-Domińczak, A. and Strawiński, P., (2013). Wchodzenie osób młodych na rynek pracy w Polsce. In Proceedings of the 9th Congress of Polish Economists, pp. 28–29.

Cramer, J. S., 2003. Logit Models from Economics and Other Fields. Cambridge: Cambridge University Press.

Central Statistical Office of Poland, (2016a). Aktywność ekonomiczna ludności Polski w latach 2013 – 2015. Warszawa: CSO.

Central Statistical Office of Poland, (2016b). Monitoring Rynku Pracy, Kwartalna informacja o rynku pracy. Warszawa: CSO.

Dale, K., (2009). Household skills and low wages. Journal of Population Economics, 22(4), pp. 1025–1038.

Davidescu, A. A. M., Roman, M., Strat, V. A. and Mosora, M., (2019). Regional sustainability, individual expectations and work motivation: A multilevel analysis. Sustainability, 11(12), 3331.

de Dios Jiménez, J., Salas-Velasco, M., (2000). Modeling educational choices. A binomial logit model applied to the demand for higher education. Higher Education, 40(3), pp. 293–311.

Gallie, D., Paugam, S. eds., (2000). Welfare regimes and the experience of unemployment in Europe. Oxford: OUP Oxford.

Gamerman, D., (1997). Sampling from the posterior distribution in generalized linear mixed models. Statistics and Computing, 7(1), pp. 57–68.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B., (2000). Bayesian Data Analysis. London: Chapman & Hall/CRC.

Grzenda, W., (2012). Badanie determinant pozostawania bez pracy osób młodych z wykorzystaniem semiparametrycznego modelu Coxa. Przegląd Statystyczny, 59(1), pp. 123–139.

Grzenda, W., (2019). Modelowanie karier zawodowej i rodzinnej z wykorzystaniem podejścia bayesowskiego. Warszawa: Wydawnictwo Naukowe PWN.

Hausman, J., McFadden, D., (1984). Specification tests for the multinomial logit model. Econometrica, 52, pp. 1219–1240.

Hsiao, C., Small, K., (1985). Multinomial logit specification tests. International economic review, 26(3), pp. 619–627.

Jasiński, M., Bożykowski, M., Chłoń-Domińczak, A., Zając, T. and Żółtak, M., (2017). Who gets a job after graduation? Factors affecting the early career employment chances of higher education graduates in Poland. Edukacja Quarterly, 143(4).

Kołaczek, B., (2005). Podstawowe uwarunkowania społeczne dostępu młodzieży do kształcenia. Polityka Społeczna, 1.

Lahiri, K., Gao, J., (2002). Bayesian Analysis of Nested Logit Model by Markov Chain Monte Carlo. Journal of Econometrics, 11, pp. 103–133.

McFadden, D., (1978). Modelling the Choice of Residential Location. In: A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, eds. Spatial Interaction Theory and Planning Models. Amsterdam: North-Holland, pp. 75–96.

Michaud, P. C., Tatsiramos, K., (2011). Fertility and female employment dynamics in Europe: the effect of using alternative econometric modeling assumptions. Journal of Applied Econometrics, 26(4), pp. 641–668.

Ministerstwo Rodziny, Pracy i Polityki Społecznej, Departament Rynku Pracy, (2018). Sytuacja na rynku pracy osób młodych w 2017 roku, Warszawa: MRPiPS.

Robert, C. P., Casella, G., (2004). Monte Carlo Statistical Methods. 2nd ed. New York: Springer.

Rossi, P. E., Allenby, G. M. and McCulloch, R., (2005). Bayesian Statistics and Marketing. Chichester. UK: John Wiley & Sons.

Stanisz, A., (2016). Modele regresji logistycznej: zastosowania w medycynie, naukach przyrodniczych i społecznych, Kraków: Wydawnictwo StatSoft Polska.

Train, K. E., (2009). Discrete Choice Methods with Simulation. 2nd ed. Cambridge: Cambridge University Press.

STATISTICS IN TRANSITION new series, September 2021
Vol. 23, No. 3, pp. 193–205, DOI 10.21307/stattrans-2021-034
Received – 12.10.2019; accepted – 17.02.2021

On the improvement of paired ranked set sampling to estimate population mean

Syed Abdul Rehman[1], **Javid Shabbir**[2]

ABSTRACT

In ecological and environmental sampling the quantification of units is either difficult or overly demanding in terms of the time, money, workload, it requires. For this reason efficient and cost-effective sampling methods need to be devised for data collecting. The most commonly used method for this purpose is the Ranked Set Sampling (RSS). In this paper, a sampling scheme called Improved Paired Ranked Set Sampling (IPRSS) is proposed to estimate the population mean. The performance of the proposed IPRSS is evaluated under perfect and imperfect rankings. A simulation study based on selected hypothetical distributions and a real-life data set showed that IPRSS is more precise than RSS, Paired RSS (PRSS) or Extreme RSS (ERSS).

Key words: order statistics, ranked set sampling, relative efficiency, unbiased estimator, imperfect ranking.

1. Introduction

There are several methods of sampling that can be used to survey the natural resources of agriculture, biology, ecology, environmental management and forestry, etc. Whatever method of sampling is used, the main objective is to obtain precise estimates of population parameters at lowest cost of time, money and labour. One efficient sampling method is RSS, which provides more efficient estimates than simple random sampling (SRS). RSS is used when the exact quantification of selected units is expensive yet ranking a small set of selected units is inexpensive. For example, if interest lies in estimating the average weight of abalone, then it is easy to rank a small set of abalone with respect to their visual size. Similarly, the fuel consumption of vehicles can be ranked by a visual inspection of the vehicle size. McIntyre (1952) was the first to suggest the RSS method and estimated the average pasture and forage yields. The theory of RSS procedure was developed by Takahasi and Wakimoto (1968) assuming perfect ranking. Dell and Clutter (1972) showed that the mean estimator under imperfect RSS remains an unbiased estimator of population mean. Lynne Stokes (1977) showed the possibility of the ranking of the study variable based on an inexpensive concomitant variable. More applications of RSS can be seen from Johnson et al. (1993), Patil (1995), Mode et al. (1999), Al-Saleh and Al-Shrafat (2001), Yu and Tam (2002), Al-Saleh and Al-Hadrami (2003), Chen et al. (2003), Buchanan et al. (2005), Haq

[1]Balochistan University of Information Technology, Engineering and Management Sciences, Pakistan. E-mail: rehmanbukhari725@gmail.com. ORCID: https://orcid.org/0000-0003-0992-9310.
[2]Quaid-i-Azam University, Islamabad, Pakistan. E-mail: javidshabbir@gmail.com. ORCID: https://orcid.org/0000-0002-0035-7072.

et al. (2013) and references cited therein.
Muttlak (1996) introduced the paired RSS (PRSS) and Median RSS (MRSS) schemes for estimation of population mean. Muttlak (1998) extended the work of Patil (1995) and showed that the mean estimator under MRSS is more efficient than the mean estimators under SRS and RSS. Samawi et al. (1996) suggested Extreme RSS (ERSS) for estimation of population mean. Muttlak (2003) suggested quartile RSS (QRSS) to get more representative units as a sample. Biradar and Santosha (2015) proposed Independent Extreme RSS (IERSS) by modifying original RSS. It is well understood that population parameters like mean and variance might not be estimated more efficiently when the population under study is suspected of containing some extreme values or outliers. The selection of any of these extreme values in a sample will affect both the precision and accuracy of estimates. All the above discussed RSS schemes, especially usual RSS, MRSS and ERSS, might suffer the consequences of extreme values being selected in a sample. In order to overcome this deficiency, we propose a new RSS scheme by taking the advantage of ranking made in RSS. Our proposed RSS scheme called Improved Paired RSS (IPRSS) is more efficient as compared to other existing RSS schemes to estimate the population mean. Another advantage of using the IPRSS is that it requires less number of units to identify or observe as compared to RSS and MRSS, while requiring more units to identify as compared to PRSS.
Let Y and X respectively be the study variable and concomitant variable respectively. Let $Y_{i(i)j}$ be the i^{th} order statistic in the i^{th} set of the j^{th} cycle of the RSS scheme. Let $\bar{y}_{RSS}$ be the sample mean of study variable based on the RSS sample. Let $\Delta_i = \mu_{(i)} - \mu$ be the deviation of the ith order statistic from true mean. Similarly, other deviations can be defined in the same manner.

2. Ranked Set Sampling (RSS)

By means of the RSS procedure, initially m^2 units are identified from the population and randomly allocated to m sets each of size m. Now within each set, units are ranked visually with respect to the variable under study or by any other economical method. The lowest ranked unit from the first set is selected and the second lowest ranked unit is selected from the second set. The procedure is carried out until the highest ranked unit is selected from the last set. A sample of m units is collected which also completes one cycle of a RSS. Repeating this procedure times results in a sample of size $n = mr$.
Let $Y_{1(1)j}, Y_{1(2)j}, \cdots, Y_{1(m)j}; Y_{2(1)j}, Y_{2(2)j}, \cdots, Y_{2(m)j}; \cdots, Y_{m(1)j}, Y_{m(2)j}, \cdots, Y_{m(m)j}$ be m independent random sets (samples) in the jth cycle, each of size m, such that $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, r$. Hence, the collected units in a sample by means of the RSS procedure are: $Y_{1(1)j}, Y_{2(2)j}, \cdots, Y_{m(m)j}$. Takahasi and Wakimoto (1968) showed that under perfect ranking, the mean of RSS is $\bar{y}_{RSS} = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m} Y_{i(i)j}$, which is an unbiased estimator of population mean μ and is more precise than $\bar{y}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. The variance of $\bar{y}_{RSS}$ is given by

$$Var(\bar{y}_{RSS}) = Var(\bar{y}_{SRS}) - \frac{1}{rm^2}\sum_{i=1}^{m}\Delta_{(i)}^2 \tag{1}$$

where $\Delta_{(i)}$ is defined earlier.

2.1. Mathematics of RSS procedure

Let Y be the study variable with pdf $f(y)$ and CDF $F(y)$ with mean μ and variance σ^2. Let $Y_1, Y_2, \cdots, Y_m$ be a simple random sample of size m drawn from $f(y)$ for which the order statistics is $Y_{(1:m)}, Y_{(2:m)}, \cdots, Y_{(m:m)}$. The pdf and CDF of the ith order statistics $Y_{(i:m)}$ for $i = 1, 2, \cdots, m$, respectively, are given by

$$f_{(i:m)}(y) = \frac{m!}{(m-i)!(i-1)!}\{F(y)\}^{i-1}\{1-F(y)\}^{m-i}f(y) \qquad -\infty < y < \infty \tag{2}$$

and

$$F_{(i:m)}(y) = \sum_{i=1}^{m}\binom{m}{r}\{F(y)\}^{r}\{1-F(y)\}^{m-r} \tag{3}$$

The expression for mean and variance of $Y_{(i:m)}$, respectively, are given by

$$\mu_{(i:m)}(y) = \int y f_{(i:m)}(y)dy \qquad \text{and} \tag{4}$$

$$\sigma^2_{(i:m)}(y) = \int (y-\mu_{(i:m)}(y))^2 f_{(i:m)}(y)dy, \tag{5}$$

for detail see David and Nagaraja (2003).

3. RSS with errors in ranking (imperfect)

In practical situations, we come across problems in which visual ranking of the study variable is difficult or impossible. Lynne Stokes (1977) introduced a model to rank the study variable (Y) with respect to ranks of the auxiliary variable (X) correlated with (Y). This procedure was named imperfect RSS (IRSS). Following Lynne Stokes (1977), it is assumed that (Y,X) follows a bivariate normal distribution and the regression of Y on X is linear, i.e.

$$Y_{i[i]j} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}\left(X_{i(i)j} - \mu_x\right) + \varepsilon_{ij} \tag{6}$$

where ρ is the population correlation coefficient between Y and X, and ε_{ij} is the error term with zero mean and a constant variance, i.e.,

$$E(\varepsilon_{ij}) = 0 \quad and \quad Var(\varepsilon_{ij}) = \sigma^2_\varepsilon = \sigma^2_y(1-\rho^2) \tag{7}$$

Here, $X_{i(i)j}$ is the i^{th} order statistics selected from the i^{th} sample in the j^{th} cycle of the auxiliary variable whose $X_{i[i])j}$ corresponding is the i^{th} judgment order statistics of the study variable. The sample mean $\bar{y}_{IRSS}$ and its variance under IRSS respectively, are given by

$$\bar{y}_{IRSS} = \frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{m} Y_{i[i]j} \tag{8}$$

and

$$Var(\bar{y}_{IRSS}) = \frac{1}{rm^2}\left\{m\sigma_y^2(1-\rho^2)+\rho^2\frac{\sigma_y^2}{\sigma_x^2}\sum_{i=1}^{m}\sigma_{x(i)}^2\right\} \tag{9}$$

4. Paired Ranked Set Sampling (PRSS)

The PRSS procedure was suggested by Muttlak (1996) to estimate the mean of finite population. The procedure is as follows: For even sample size m, identify $m/2$ sets each of size m and rank the units in each set. The lowest and highest ranked units are collected from the first set, and the second lowest and second highest ranked units are collected from the second set. The process continues until $(m/2)^{th}$ and $((m+2)/2)^{th}$ ranked units are collected from the last set. For odd sample size m , identify $((m+1)/2)$ sets, each of size m , and select units as previously mentioned in the case of even size until $(m/2)^{th}$ set. Also $((m+1)/2)^{th}$ ranked unit is collected from the last set. This concludes in one cycle of a PRSS of size m. By repeating this process r times a sample of size m can be obtained. The sample mean under PRSS depending on even (E) and odd (O) sample sizes are calculated as, respectively

$$\bar{y}_{PRSS}^{E} = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m}\left(Y_{i(i)j}+Y_{i(m+1-i)j}\right) \tag{10}$$

and

$$\bar{y}_{PRSS}^{O} = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{(m+1)/2}Y_{i(i)j}+\sum_{i=1}^{(m-1)/2}Y_{i(m+1-i)j}\right) \tag{11}$$

The bias and variances of $\bar{y}_{PRSS}^{E}$ and $\bar{y}_{PRSS}^{O}$ respectively, are given by

$$Bias(\bar{y}_{PRSS}^{E}) = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m}\left(\Delta_{i(i)j}+\Delta_{i(m+1-i)j}\right) \tag{12}$$

$$\begin{aligned} Var(\bar{y}_{PRSS}^{E}) &= Var(\bar{y}_{SRS}) - \frac{1}{rm^2}\sum_{j=1}^{r}\sum_{i=1}^{m}\left(\Delta_{i(i)j}^2+\Delta_{i(m+1-i)j}^2\right) \\ &\quad + \frac{2}{(rm)^2}\sum_{j=1}^{r}\sum_{i=1}^{m}\Delta_{i(i)j}\Delta_{i(m+1-i)j} \end{aligned} \tag{13}$$

and

$$Bias(\bar{y}_{PRSS}^{O}) = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{(m+1)/2}\Delta_{i(i)j}+\sum_{i=1}^{(m-1)/2}\Delta_{i(m+1-i)j}\right) \tag{14}$$

$$\begin{aligned} Var(\bar{y}_{PRSS}^{O}) &= Var(\bar{y}_{SRS}) - \frac{1}{rm^2}\left(\sum_{i=1}^{(m+1)/2}\Delta_{(i)}^2+\sum_{i=1}^{(m-1)/2}\Delta_{(m+1-i)}^2\right) \\ &\quad + \frac{2}{r^2m^2}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2}\Delta_{(i)}\Delta_{(m+1-i)} \end{aligned} \tag{15}$$

In the case of symmetrical distribution both $\bar{y}^{E}_{PRSS}$ and $\bar{y}^{O}_{PRSS}$ are unbiased estimators of population mean.

5. Extreme Ranked Set Sampling (ERSS)

The ERSS procedure was introduced by Samawi et al. (1996) to estimate the population mean. It is an essential sampling scheme in the case where population is fat-tailed, e.g. students t-distribution, Uniform distribution, Cauchy distribution, etc., units deviates from their mean beyond 5-sigma limits. The ERSS technique is executed as:
From the population, identify m^2 units and allocate them randomly to m sets, each of size . Now rank the units within each set. For $m = even$, select the lowest ranked units from the first $m/2$ sets and highest ranked units from the last $m/2$ sets. For $m = odd$, select the lowest ranked unit from the first $(m-1)/2$ sets, and the highest ranked units from set $(m+1)/2$ to set $(m-1)$. The median, i.e. $((m+1)/2))^{th}$ unit is selected from the last set. This conclude one cycle of ERSS of size m. A sample of size $n = mr$ can be obtained by repeating the ERSS process r times.
In the j^{th} cycle , ERSS is termed as: $Y_{1(1)j}, Y_{2(1)j}, \cdots, Y_{m/2(1)j}, Y_{m+2/2(m)j}, \cdots, Y_{m(m)j}$ for even m. While, for odd m, $Y_{1(1)j}, Y_{2(1)j}, \cdots, Y_{m-1/2(1)j}, Y_{m+1/2(m)j}, \cdots, Y_{m(m)j}, Y_{m(m-1/2)j}$ units are selected. The estimators based on ERSS, are given by:

$$\bar{y}^{E}_{ERSS} = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{m/2} Y_{i(1)j} + \sum_{i=(m+2)/2}^{m} Y_{i(m)j}\right) \tag{16}$$

and

$$\bar{y}^{O}_{ERSS} = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{(m-1)/2} Y_{i(1)j} + \sum_{i=(m+1)/2}^{m-1} Y_{i(m)j} + Y_{m(m+1)/2}\right) \tag{17}$$

The bias and variance of each estimator is obtained as

$$Bias(\bar{y}^{E}_{ERSS}) = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{m/2} \Delta_{(1)} + \sum_{i=(m+2)/2}^{m} \Delta_{(m)}\right) \tag{18}$$

$$Var(\bar{y}^{E}_{ERSS}) = Var(\bar{y}_{SRS}) - \frac{1}{rm^2}\left(\sum_{i=1}^{m/2} \Delta^2_{(1)} + \sum_{i=(m+2)/2}^{m} \Delta^2_{(m)}\right) \tag{19}$$

and

$$Bias(\bar{y}^{O}_{ERSS}) = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{(m-1)/2} \Delta_{(1)} + \sum_{i=(m+1)/2}^{m-1} \Delta_{(m)} + \Delta_{((m+1)/2)}\right) \tag{20}$$

$$Var(\bar{y}^{O}_{ERSS}) = Var(\bar{y}_{SRS}) - \frac{1}{rm^2}\left(\sum_{i=1}^{(m-1)/2} \Delta^2_{(1)} + \sum_{i=(m+1)/2}^{m-1} \Delta^2_{(m)} + \Delta^2_{((m+1)/2)}\right) \tag{21}$$

6. Improved Paired Ranked Set Sampling (IPRSS)

The proposed RSS procedure called IPRSS is an improvement in PRSS and is an effort to overcome the drawback of ERSS. There is a great chance that ERSS will select the extreme values contained by population under study by considering only the first and last order statistic of sets, which not only affects the representativeness of a sample but the accuracy and precisions both are also highly affected. The proposed IPRSS scheme is also an effort to counter this drawback.

The IPRSS procedure is as follows:

For an even sample size m , identify $m/2$ sets each of size $m+2$ from a population and within each set rank the units. Now, choose the second lowest and second highest ranked units from the first set. Similarly select the third lowest and third highest ranked units from the second set. The procedure continues until $(m+4)^{th}$ and $(m+4/2)^{th}$ ranked units are selected from the last set.

For an illustration, a selection of units by IPRSS for $m=4$ is given by

$$\begin{array}{cccccc} Y_{1(1)j} & \underline{Y_{1(2)j}} & Y_{1(3)j} & Y_{1(4)j} & \underline{Y_{1(5)j}} & Y_{1(6)j} \\ Y_{2(1)j} & Y_{2(2)j} & \underline{Y_{2(3)j}} & \underline{Y_{2(4)j}} & Y_{2(5)j} & Y_{2(6)j} \end{array}$$

In case of odd sample size m, identify $m+1/2$ sets each of size $m+2$ and select units as aforesaid in case of even until $m-1/2$ set. Also $(m+3/2)^{th}$ ranked unit is choosen from the last set. For example $m=5$, then the IPRSS procedure can be illustrated as

$$\begin{array}{ccccccc} Y_{1(1)j} & \underline{Y_{1(2)j}} & Y_{1(3)j} & Y_{1(4)j} & Y_{1(5)j} & \underline{Y_{1(6)j}} & Y_{1(7)j} \\ Y_{2(1)j} & Y_{2(2)j} & \underline{Y_{2(3)j}} & Y_{2(4)j} & \underline{Y_{2(5)j}} & Y_{2(6)j} & Y_{2(7)j} \\ Y_{3(1)j} & Y_{3(2)j} & Y_{3(3)j} & \underline{Y_{3(4)j}} & Y_{3(5)j} & Y_{3(6)j} & Y_{3(7)j} \end{array}$$

This concludes one cycle of an IPRSS of size m. The process can be repeated r times to obtain IPRSS of size $n=mr$. The sample means under IPRSS depending on even (E) and odd (O) sample sizes are respectively given by

$$\bar{y}^{E}_{IPRSS} = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m/2}\left(Y_{i(i+1)j}+Y_{i(m+2-i)j}\right) \tag{22}$$

and

$$\bar{y}^{O}_{IPRSS} = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{(m+1)/2} Y_{i(i+1)j} + \sum_{i=1}^{(m-1)/2} Y_{i(m+2-i)j}\right) \tag{23}$$

In case of a symmetric distribution about zero, i.e. $Y_{(i+1)} \cong Y_{(m+2-i)}$ for $i=1,2,\cdots,m$. Arnold et al. (1992) showed that $\mu_{(i+1)} = -\mu_{(m+2-i)}$ and $\sigma^2_{(i+1)} \approxeq -\sigma^2_{(m+2-i)}$. This indicates that if m is odd, then $\mu_{(m+1/2)} = \mu = 0$. Hence, the suggested IPRSS provides unbiased estimators, however in the case of asymmetrical distribution the expressions for bias along

with the variance of $\bar{y}^E_{IPRSS}$ and $\bar{y}^O_{IPRSS}$ respectively, are given as

$$Bias(\bar{y}^E_{IPRSS}) = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m/2}\left(\Delta_{(i+1)} + \Delta_{(m+2-i)}\right) \tag{24}$$

$$\begin{aligned} Var(\bar{y}^E_{IPRSS}) &= Var(\bar{y}_{SRS}) - \frac{1}{rm^2}\sum_{i=1}^{m/2}\left(\Delta^2_{(i+1)} + \Delta^2_{(m+2-i)}\right) \\ &\quad + \frac{2}{(rm)^2}\sum_{j=1}^{r}\sum_{i=1}^{m/2}\Delta_{(i+1)}\Delta_{(m+1-i)} \end{aligned} \tag{25}$$

and

$$Bias(\bar{y}^O_{IPRSS}) = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{(m+1)/2}\Delta_{(i+1)} + \sum_{i=1}^{(m-1)/2}\Delta_{(m+2-i)}\right) \tag{26}$$

$$\begin{aligned} Var(\bar{y}^O_{IPRSS}) &= Var(\bar{y}_{SRS}) - \frac{1}{(rm)^2}\left(\sum_{i=1}^{(m+1)/2}\Delta^2_{(i+1)} + \sum_{i=1}^{(m-1)/2}\Delta^2_{(m+2-i)}\right) \\ &\quad + \frac{2}{(rm)^2}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2}\Delta_{(i+1)}\Delta_{(m+2-i)} \end{aligned} \tag{27}$$

7. IPRSS with ranking errors (imperfect ranking)

In the case of ranking errors, the quantified observations of the study variable from the i^{th} sample may not be the i^{th} order statistics instead of the i^{th} judgment order statistics. Dell and Clutter (1972) demonstrated that the sample mean of RSS with errors in ranking is an unbiased estimator of the population mean, and has smaller variance than the usual estimator based on SRS with the same sample size. However, the variance of the estimator with errors in ranking will be larger than the variance of the estimator with perfect ranking. Let $Y_{[i+1]}$ and $Y_{[m+2-i]}$ denote the $(i+1)^{th}$ and $(m+2-i)^{th}$ judgment order statistics of the sample for $i = 1,2,\cdots,m+2$. Then, the IPRSS estimator of population mean with errors in ranking is defined as:

$$\bar{y}^E_{IPRSS} = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m/2}\left(Y_{i[i+1]j} + Y_{i[m+2-i]j}\right) \tag{28}$$

and

$$\bar{y}^O_{IPRSS} = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{(m+1)/2}Y_{i[i+1]} + \sum_{i=1}^{(m-1)/2}Y_{i[m+2-i]j}\right) \tag{29}$$

the bias and variance of $\bar{y}^{E}_{IPRSS}$ and $\bar{y}^{O}_{IPRSS}$ are

$$Bias(\bar{y}^{E}_{IPRSS}) = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m/2}\left(\Delta_{[i+1]}+\Delta_{[m+2-i]}\right) \tag{30}$$

$$\begin{aligned} Var(\bar{y}^{E}_{IPRSS}) &= Var(\bar{y}_{SRS}) - \frac{1}{rm^2}\sum_{i=1}^{m/2}\left(\Delta^2_{[i+1]}+\Delta^2_{[m+2-i]}\right) \\ &\quad + \frac{2}{(rm)^2}\sum_{j=1}^{r}\sum_{i=1}^{m/2}\Delta_{[i+1]}\Delta_{[m+1-i]} \end{aligned} \tag{31}$$

and

$$Bias(\bar{y}^{O}_{IPRSS}) = \frac{1}{rm}\sum_{j=1}^{r}\left(\sum_{i=1}^{(m+1)/2}\Delta_{[i+1]}+\sum_{i=1}^{(m-1)/2}\Delta_{[m+2-i]}\right) \tag{32}$$

$$\begin{aligned} Var(\bar{y}^{O}_{IPRSS}) &= Var(\bar{y}_{SRS}) - \frac{1}{(rm)^2}\left(\sum_{i=1}^{(m+1)/2}\Delta^2_{[i+1]}+\sum_{i=1}^{(m-1)/2}\Delta^2_{[m+2-i]}\right) \\ &\quad + \frac{2}{(rm)^2}\sum_{j=1}^{r}\sum_{i=1}^{(m-1)/2}\Delta_{[i+1]}\Delta_{[m+2-i]} \end{aligned} \tag{33}$$

The relative efficiency (RE) of the proposed $\bar{y}_{IPRSS}$ with respect to $\bar{y}_{RSS}$, $\bar{y}_{PRSS}$ and $\bar{y}_{ERSS}$ is defined as:

$$RE(\bar{y}_{IPRSS},\bar{y}_{\bullet}) = \frac{MSE(\bar{y}_{\bullet})}{MSE(\bar{y}_{IPRSS})} \tag{34}$$

where

$$MSE(\bar{y}_{\bullet}) = Var(\bar{y}_{\bullet}) + (Bias(\bar{y}_{\bullet}))^2 \tag{35}$$

and

$$MSE(\bar{y}_{\bullet}) = Var(\bar{y}_{IPRSS}) + \left(Bias(\bar{y}_{IPRSS})\right)^2 \tag{36}$$

8. Empirical study

To evaluate the efficiency comparison of IPRSS against other competing procedures, a simulation study is conducted based on the following steps:
Generate a hypothetical (Normal, Logistic, Exponential and Poisson) distribution, say "X" of size 1000 and also generate "$e \sim N(1000,0,1)$". Compute the study variable "Y" by using the relation $Y = \rho X + e\sqrt{1-\rho^2}$, where ρ is the population coefficient of correlation between the study variable and the concomitant variable whose value is taken fixed as 0.50, 0.70 and 0.90. Draw a sample of size $n = mr$, using SRSWOR, and estimate mean for each sampling procedure, for all the pairs of $r = 1,3$ and $m = 4,6,8$. This procedure is repeated 10,000 times. The results are given in Tables 1 and 2.

Table 1: RE of IPRSS over other sampling procedures based on perfect ranking

		RSS		ERSS		PRSS	
Distribution	m/r	1	3	1	3	1	3
Normal(1000,100,5)	4	1.05	1.09	1.24	1.28	1.47	1.48
	6	1.04	1.03	1.45	1.45	4.98	11.96
	8	1.01	1.04	1.58	1.66	12.33	35.00
Logistic(1000,5,2)	4	1.29	1.27	1.68	1.73	1.73	1.64
	6	1.24	1.21	2.05	2.05	5.17	12.29
	8	1.18	1.18	2.36	2.43	12.41	33.98
Exponential(1000,1)	4	1.16	1.09	1.48	1.44	1.56	1.48
	6	1.07	1.04	1.80	2.05	4.79	11.5
	8	1.08	1.03	2.43	3.04	12.11	30.58
Poisson(1000,3)	4	1.39	1.36	2.03	1.98	1.91	1.76
	6	1.44	1.39	3.11	3.27	5.00	11.97
	8	1.40	1.29	4.24	4.62	12.21	30.4

Table 2: RE of IPRSS over other sampling procedures based on imperfect ranking

			RSS			ERSS			PRSS		
Distribution	r	m/ρ	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
		4	1.04	1.05	1.05	1.18	1.21	1.22	1.35	1.4	1.43
	1	6	1.05	1.05	1.05	1.36	1.39	1.40	3.66	4.17	4.44
Normal		8	1.04	1.03	1.02	1.42	1.47	1.50	7.98	9.59	10.49
(1000,100,5)		4	1.08	1.08	1.08	1.23	1.25	1.25	1.34	1.39	1.42
	3	6	1.03	1.03	1.03	1.34	1.38	1.40	8.13	9.61	10.41
		8	1.08	1.07	1.06	1.50	1.56	1.58	21.76	26.76	29.53
		4	1.20	1.24	1.26	1.47	1.57	1.62	1.44	1.54	1.61
	1	6	1.19	1.21	1.22	1.64	1.79	1.89	3.14	3.79	4.20
Logistic		8	1.13	1.16	1.18	1.76	1.96	2.08	6.11	7.94	9.17
(1000,5,2)		4	1.18	1.22	1.24	1.45	1.53	1.58	1.44	1.55	1.62
	3	6	1.14	1.17	1.19	1.63	1.79	1.88	6.59	8.41	9.58
		8	1.12	1.14	1.16	1.82	2.02	2.14	15.61	20.97	24.56
		4	1.02	1.03	1.04	1.11	1.17	1.25	1.06	1.09	1.13
	1	6	1.01	1.01	1.01	1.25	1.4	1.57	1.38	1.55	1.74
Exponential		8	1.01	1.00	1.00	1.46	1.78	2.13	1.8	2.2	2.66
(1000,1)		4	1.02	1.02	1.04	1.09	1.15	1.21	1.03	1.04	1.07
	3	6	1.01	1.01	1.02	1.52	1.78	2.05	1.88	2.36	2.84
		8	1.00	1.00	1.00	2.07	2.71	3.37	2.86	3.96	5.09
		4	1.16	1.22	1.27	1.43	1.6	1.74	1.31	1.44	1.55
	1	6	1.16	1.23	1.29	1.69	2.04	2.35	1.96	2.41	2.81
Poisson		8	1.09	1.14	1.19	1.89	2.4	2.87	3.12	4.25	5.3
(1000,3)		4	1.15	1.18	1.21	1.42	1.58	1.70	1.25	1.34	1.41
	3	6	1.11	1.15	1.18	1.79	2.18	2.51	3.43	4.68	5.74
		8	1.09	1.11	1.13	2.22	2.92	3.53	6.61	9.75	12.51

The simulation results show that the proposed IPRSS performs better than RSS, PRSS and ERSS. It is also observable from the simulation results that the efficiency of IPRSS significantly increases in comparison with other techniques when the number of cycles (r) is increased, i.e. in each cycle IPRSS avoids two suspected extreme values that are selected by other RSS schemes. Hence, with the increase in cycle (r), the number of suspected extreme values avoided by IPRSS also increases, which help us getting a sample free from outliers or extreme values. It is also notable that under imperfect ranking, all the RSS techniques including IPRSS perform poor relative to the case of perfect ranking. Since IPRSS collects a sample that avoids extreme values, so the results of IPRSS will be certainly better than other techniques. Since IPRSS is an improvement in the PRSS and it is designed to avoid the selection of extreme values in a sample so it is obvious that IPRSS will select more representative sample especially when population contains extreme values.

9. Application on real-life data set

For practical application, we consider the data provided by Singh and Mangat (2013). The net irrigated area is considered as a study variable while the number of tractors is considered as a concomitant variable. A simulation study is conducted by selecting 10,000 samples (the simulation procedure is discussed above). For simplicity we fix $r = 1$ and $m = 3,4,5,6,7$ using the procedure of RSS, ERSS, PRSS and IPRSS, by means of both perfect and imperfect rankings.
The $MSE(\bar{y}(\bullet))$ of each sampling procedure is calculated as,

$$MSE(\bar{y}(\bullet)) = \frac{1}{10,000} \sum_{i=1}^{10,000} (\bar{y}(\bullet) - \mu_y)^2 \tag{37}$$

The relative efficiencies are given in Table 3.

Table 3: RE of IPRSS over other sampling procedures

		RE		
Ranking	m	RSS	ERSS	PRSS
Perfect	3	1.746	1.731	1.910
	4	1.358	2.436	1.613
	5	1.334	2.338	1.450
	6	1.157	4.270	1.379
	7	1.127	4.195	1.195
Imperfect	3	1.617	1.613	1.715
	4	1.441	2.390	1.532
	5	1.345	2.190	1.390
	6	1.207	3.582	1.239
	7	1.124	3.358	1.190

The results in Table 3 show that IPRSS performs better than RSS, ERSS and PRSS on real-life data set. Furthermore, IPRSS performs better than ERSS distinguishably when the sample size is increased.

10. Conclusions

In this article we proposed an efficient sampling scheme for the estimation of population mean. We showed numerically that the mean estimators under IPRSS are more precise than the mean estimators under RSS, ERSS and PRSS, for both perfect and imperfect rankings based on various distributions and a real-life data set. Generally, the proposed mean estimates under IPRSS are more efficient than the estimates under RSS, ERSS, and PRSS when more cycles of small samples are selected. Hence, the use of IPRSS is recommended over the existing RSS schemes. The current work can be extended to develop ratio and regression type estimators of population mean based on IPRSS to get more precise estimates.

Using the idea suggested in this paper, other RSS techniques can be modified to eliminate the effects of extreme values in samples.

Acknowledgements

Both authors are thankful to the unknown referees for their valuable comments, which helped in improving the quality of the paper.

References

Al-Saleh, M. F. and Al-Hadrami, S. A. (2003). Parametric estimation for the location parameter for symmetric distributions using moving extremes ranked set sampling with application to trees data. *Environmetrics: The official journal of the International Environmetrics Society*, 14(7):651–664.

Al-Saleh, M. F. and Al-Shrafat, K. (2001). Estimation of average milk yield using ranked set sampling. *Environmetrics: The official journal of the International Environmetrics Society*, 12(4):395–399.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (1992). A first course in order statistics, vol. 54. *SIAM, Philadelphia*.

Biradar, B. and Santosha, C. (2015). Estimation of the population mean based on extremes ranked set sampling. *American Journal of Mathematics and Statistics*, 5(1):32–36.

Buchanan, R. A., Conquest, L. L., and Courbois, J.-Y. (2005). A cost analysis of ranked set sampling to estimate a population mean. *Environmetrics: The official journal of the International Environmetrics Society*, 16(3):235–256.

Chen, Z., Bai, Z., and Sinha, B. (2003). *Ranked set sampling: theory and applications*, volume 176. Springer Science & Business Media.

David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*. John Wiley and Sons New York, 3 edition.

Dell, T. and Clutter, J. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, pages 545–555.

Haq, A., Brown, J., Moltchanova, E., and Al-Omari, A. I. (2013). Partial ranked set sampling design. *Environmetrics*, 24(3):201–207.

Johnson, G., Patil, G., and Sinha, A. (1993). Ranked set sampling for vegetation research. *Abstracta Botanica*, pages 87–102.

Lynne Stokes, S. (1977). Ranked set sampling with concomitant variables. *Communications in Statistics-Theory and Methods*, 6(12):1207–1211.

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian journal of agricultural research*, 3(4):385–390.

Mode, N. A., Conquest, L. L., and Marker, D. A. (1999). Ranked set sampling for ecological research: accounting for the total costs of sampling. *Environmetrics: The official journal of the International Environmetrics Society*, 10(2):179–194.

Muttlak, H. (1998). Median ranked set sampling with concomitant variables and a comparison with ranked set sampling and regression estimators. *Environmetrics: The official journal of the International Environmetrics Society*, 9(3):255–267.

Muttlak, H. A. (1996). Pair rank set sampling. *Biometrical journal*, 38(7):879–885.

Muttlak, H. A. (2003). Investigating the use of quartile ranked set samples for estimating the population mean. *Applied Mathematics and Computation*, 146(2-3):437–443.

Patil, G. (1995). ranked set sampling. *Environmental and Ecological Statistics*, 2(4):271–285.

Samawi, H. M., Ahmed, M. S., and Abu-Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sampling. *Biometrical Journal*, 38(5):577–586.

Singh, R. and Mangat, N. S. (2013). *Elements of survey sampling*, volume 15. Springer Science & Business Media.

Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the institute of statistical mathematics*, 20(1):1–31.

Yu, P. L. and Tam, C. Y. (2002). Ranked set sampling in the presence of censored data. *Environmetrics: The official journal of the International Environmetrics Society*, 13(4):379–396.

STATISTICS IN TRANSITION new series, September 2021
Vol. 22, No. 3, pp. 207–211,

About the Authors

Abu Awwad Raed R. is an assistant professor with a nine-year experience of working at the Department of Mathematics at the University of Petra. He received his Ph.D. from the University of Jordan. His research interests include mathematical statistics, Bayesian and non-Bayesian inferences and ordered data analysis. He has published over 10 research papers in international journals.

Abufoudeh Ghassan has served as the Chairman of the Department of Mathematics at the University of Petra since 2017. He graduated from the Jordan University with a Ph.D. in mathematics in 2013. His research areas of interest include mathematical statistics and entropy measures. After graduation, he worked as an Assistant Professor at the Faculty of Arts and Sciences in the Department of Mathematics at the University of Petra. At present he is an Associate Professor in Mathematics.

Ahmad Peer Bilal is presently working as Assistant Professor (Statistics) in the Department of Mathematical Sciences, IUST, J&K, India. He received his Ph.D. degree in Statistics from University of Kashmir, Srinagar, India, in 2008. He has more than 15 years of teaching and corporate experience and has published a number of research papers in reputed international journals with Sci, Scopus, Copernicus and Zentralblatt Math etc. indexing. He has attended more than twenty five training programs and workshops. He has presented research papers in more than ten national and international conferences. His main research interests are probability distributions, their applications in data analysis and Bayesian analysis.

Ali Mir Masoom is George and Frances Ball Distinguished Professor Emeritus of Statistics at Ball State University in the USA. He obtained his B.Sc. (Honors) in 1956 and a M.Sc. in 1957, both in statistics, from University of Dhaka, and his second M.Sc and a Ph.D. from the University of Toronto, in 1967 and 1969, respectively. Dr. Ali has published about 250 articles in finite sampling, optimal spacing, skew-symmetric and generalized distributions, and several others. He has published one book, submitted one for publication, and is currently working on an edited volume. He received many awards, including gold medals from two statistical organizations, the Outstanding Researcher Award and the Outstanding Faculty Award from Ball State and the Sagamore of the Wabash (the highest award of the State of Indiana, USA). He is a Fellow of several statistical organizations.

Al-Nasser Amjad is a Full Professor of Statistics at the Department of Statistics, Science Faculty, Yarmouk University, and Vice President for Academic Affairs at Al Falah University, Dubai. His research interests cover: generalized maximum entropy, customer satisfaction index, measurement error models, data envelopment analysis, ranked set sampling and acceptance sampling plans. Professor Al-Nasser has published over 140 research papers in both local and international journals and conferences. He has also published six books and monographs. Professor Al-Nasser is an active member of many scientific professional bodies and he is the founding editor of the Electronic Journal of Applied Statistical Analysis (EJASA), indexed on WoS and Scopus: http://siba-ese.unile.it/index.php/ejasa.

Bdair Omar M. is currently a visiting associates professor in the Department of Mathematics and Statistics at the McMaster University, Canada. He received his Ph.D. from the University of Jordan. His research interests include statistical models involving ordered data analysis, applied statistics, biostatistical applications on pharmaceutical trials and cohort studies, and statistical programming using different statistical packages. He has authored or co-authored 28 publications.

Boratyńska Agata is an Associate Professor at Collegium of Economic Analysis, Warsaw School of Economics (SGH). Her main areas of interest include: Bayesian statistical models, robustness of statistical models with respect to a prior distribution, application of Bayesian models in insurance, prediction of reserves and the credibility theory.

Bouazzaoui Hajar is a Ph.D. student at the faculty of Sciences and Techniques Settat, the Hassan I university. Her work focuses mainly on topological data analysis and its applications both inside and outside mathematics. She investigates questions ranging from teaching computers how to detect patterns and solving linguistics problems to studying the dynamics of billiards.

Dar Showkat Ahmad is a research scholar in the Department of Statistics, University of Kashmir, Srinagar, India. His area of research is probability distributions and their applications in data analysis and demography. He has published several research articles in reputed journals of statistical and mathematical sciences.

Elomary Mohamed Abdou is a Full Professor of Mathematics. His research interests are algebraic theory of quadratic forms, cryptography over elliptic curves and TDA. Professor Elomary Hassan published several papers in international journals, e.g. the *Journal of Algebra* and *Math Z*. He served as the Pedagogical Director of SIST (Super Institute of Science and Technology) in Morocco in 2002-2003, and Vice President of the SMM (Moroccan Mathematical Society) from 2015 to 2019.

Enang Ekaette is a Professor of Statistics with a specialisation in sample surveys. Her research interest cuts across ratio estimation, calibration estimation and the small area and domain estimation. She has published over 40 articles in both local and international journals and has attended and presented papers in many conferences. She is a member of the Nigerian Statistical Association and several other professional bodies.

Goual Hafida is an Associate Professor at the Laboratory of Probabilities and Statistics, Department of Mathematics, Faculty of Sciences, University of Badji Mokhtar Annaba of Algeria. Her main areas of interest include: distribution theory, density estimation, R programming, reliability theory, survival analysis, statistical inference, applied statistics, statistical modelling, and hypothesis testing.

Grzenda Wioletta is an Associate Professor at the Institute of Statistics and Demography at the SGH Warsaw School of Economics. She earned her Ph.D. in mathematics from the SGH Warsaw School of Economics and received a habilitation in economics and finance from the same university, for her works on the Bayesian modelling of family and occupational careers. She published papers on the applications of the Bayesian and classical statistical methods in the analysis of the unemployment and fertility, and in the probability theory. She is an author and co-author of several books on Bayesian statistics, advanced statistical methods, and programming in data analytics.

Hassan Anwar, Professor of Statistics, Dean of the School of Physical and Mathematical Science and Director of DIQA (Directorate of Internal Quality Assurance), University of Kashmir, Srinagar, India. He earned his M.Phil and Ph.D. in statistics from AMU, Aligarh and Patna University, Patna. He has published over 75 research articles and several notable scholoars were taught by him. His main research interest lies in generalized probability distributions. He took part in the 2011 Census as a special charge officer of the University of Kashmir. He also co-conducted all India's surveys on higher education (AISHE) as an officer of Kashmir University and its affiliated colleges for the University Grants Commission, Delhi, India.

Ibrahim Mohamed is an Assistant Professor of Statistics, Department of Applied Mathematical and Actuarial Statistics, Faculty of Commerce, Damietta University, Damietta, Egypt. He earned an M.Sc. in applied statistics in 2009, from Mansoura University, Egyp, and a. Ph.D. in applied statistics in 2015 from the same university. He has authored or co-authored 30 publications. His research interests encompass distribution theory, generalized class of distributions, Bayesian analysis, discrete distributions, discrete class of distributions, estimation theory, bivariate distributions and bivariate class of distributions.

Iseh Matthew holds a Ph.D. in Statistics with a specialization in sampling theory and survey methods. He serves as Head of the Department of Statistics, Akwa Ibom State University. His area of interest include small area estimation, calibration of weights and nonresponse adjustments, Markov decision process and survival modelling, and data analysis. Dr. Iseh has published 7 research articles in the local and international journals and conferences over the last two years. He also serves as a Survey enumerator in the Central Bank of Nigeria and is an active member of Nigerian Statistical Association and Professional Statisticians Society of Nigeria.

Kumar Vinod is a Full Professor of Economics at the Panjab University Regional Centre, Sri Muktsar Sahib, Punjab, India. His research interests cover urban finance, banking, foreign direct investment, and international economics. Professor Kumar has published over 30 research papers in the local and international journals and conferences as well as two books. He is also an active member of many key administrative bodies at his university.

Mamouni My Ismail is a Full Professor and Head of the Department of Mathematics, CRMEF Rabat, Morocco. His main areas of interest include algebraic topology (both pure and applied) and mathematical education. Currently he is the editor of two local journals. Professor Mamouni has published over 20 research papers in international journals, and a book. He is an active member of many scientific professional bodies.

Rehman Syed Abdul is a lecturer of statistics at BUITEMS, Quetta, Pakistan. He is currently a doctoral student of statistics at QAU, Islamabad, under the supervision of Professor Javid Shabbir. Mr. Rehman's research interests encompass survey sampling, statistical inference, quality control and big data analysis. He has won the HEC overseas scholarship, and his current Ph.D. studies are funded by the HEC scholarships. He is also a social activist, a writer and a blogger. He works for and writes about the rights of minorities, equality and justice in society.

Rossa Agnieszka is an Associate Professor at the Department of Demography, Faculty of Economics and Sociology, University of Lodz. Her main areas of interest include demographic forecasting, event history analysis and multivariate statistical analysis. Currently she is a member of the Committee on Demographic Studies of the Polish Academy of Sciences.

Shabbir Javid is working as a Tenured Professor of Statistics at the Quaid-i-Azam University of Islamabad, Pakistan. His fields of interests are: survey sampling, randomised response, and ranked set sampling. He has published more than 200 research papers in the local and international journals, and nurtured 110 MPhil and 14 PhD students. He is the Associate Editor of the *Journal of Statistical Theory and Practice*, Springer-Verlag.

Sharma Vipin is an Associate Professor at the University School of Business-APEX-MBA, Faculty of Economics, Chandigarh University, Punjab, India. He also serves as a department coordinator of the NBA, NAAC, and IQAC. His main areas of interest include intra-regional trade, econometrics, international trade, financial inclusion, and the economics of education. Dr. Vipin Sharma has published 17 (7 research papers in local, 3 in International, 2 in the conference and 5 five chapters in edited books) research papers in the local and international journals and conferences and five book chapters. He has also presented 13 research papers at National/International Conferences. He has delivered several lectures as a resource person/keynote speaker at FDPs, workshops, debate competitions and conferences held at both the national and international levels

Szymański Andrzej, Ph.D. – retired academic teacher at the University of Lodz. His main areas of interest are multivariate statistics and fuzzy modelling.

ul Haq Muhammad Ahsan received his MPhil degree in statistics from the College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan. At present, he is a Ph.D. research scholar at the University of the Punjab, Pakistan, under the supervision of Dr. Ahmed Zogo Memon and Dr. Sohail Chand. His research interests include mathematical and applied statistics, distribution theory, reliability analysis and mixture distributions.

Yousof Haitham M. is an Assistant Professor of Statistics at the Department of Statistics, Mathematics and Insurance, Faculty of Commerce, Benha University, Egypt. He earned his B.Sc. in statistics, mathematics and insurance in 2004 from the Faculty of Commerce, Benha University, Egypt, and his M.Sc. in applied statistics in 2011 and his Ph.D. in 2015, also in applied statistics, from the same university. He authored or co-authored 161 publications. His research interest cover the probability theory, continuous distributions, discrete distributions, continuous families of distributions, discrete families of distributions, generalized class of distributions, bivariate families of distributions, Bayesian analysis, semi-parametric regression, parametric regression, nonparametric regression, and new goodness-of-fit tests.

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).